Wiley Encyclopedia of Chemical Biology (WECB) is the only major reference resource devoted to the cross-disciplinary field of chemical biology. WECB is a first port of call for academic researchers and advanced students looking for timely and authoritative review-type information in this field. The articles in WECB, contributed by world-class authors, are designed to

- illuminate the crucial role of chemistry and chemical techniques in the life sciences
- support the information needs of researchers in the field of chemical biology
- provide relevant and readily accessible information to upper-level students and to scientists in adjacent disciplines

WECB is an evolving online database of information, combining the most useful features of traditional reference works and review journals in a compelling format designed to exploit the full potential of the online medium. Extensive linking to the primary literature and to other online resources further enhances the value of the encyclopedia as a gateway to the broader scientific literature.

The editorial approach and structure of WECB will assure the longevity of its content. Through a systematic program of updates, the encyclopedia will be allowed to grow and evolve as the field itself grows and evolves, ensuring that it will remain an authoritative resource for chemical biology research and education for many years to come.

We have assembled the following distinguished Advisory Board to assist in the compilation of this work:

Tadhg Begley (Chair), Professor of Chemistry and Chemical Biology, Cornell University
Barbara Baird, Professor of Chemistry and Chemical Biology, Cornell University
Carl Batt, Professor of Food Science, Director of Ludwig Institute for Cancer Research Partnership, Cornell University
Hagan Bayley, Professor of Chemical Biology, University of Oxford
Carolyn Bertozzi, Professor of Chemistry, University of California, Berkeley
Timothy Bugg, Professor of Chemistry, University of Warwick
David Cane, Professor of Chemistry and Biochemistry, Brown University
Richard Cerione, Professor of Chemistry and Chemical Biology, Cornell University
Benjamin Cravatt, Professor of Chemistry and Cell Biology, The Scripps Research Institute
Steve Ealick, Professor of Chemistry and Chemical Biology, Cornell University
Jed Fisher, Professor of Chemistry and Biochemistry, University of Notre Dame
Sidney Hecht, Professor of Chemistry, University of Virginia
Robert Hertzberg, Vice President, Screening & Compound Profiling, GlaxoSmithKline Pharmaceuticals
Linda Hsieh-Wilson, Professor of Chemistry, California Institute of Technology
Tom Muir, Professor of Chemistry, Rockefeller University
Matt Pretender, Professor of Chemistry, University of Gigapedia
Lawrence Que, Jr., Professor of Chemistry, University of Minnesota
Kevan Shokat, Professor of Cellular and Molecular Pharmacology, University of California, San Francisco
Jeffrey Skolnick, Director Center for the Study of Systems Biology, Georgia Institute of Technology
William Smith, Professor of Biological Chemistry, University of Michigan
Dotsevi Sogah, Professor of Chemistry and Chemicaliology, Cornell University
Joanne Stubbe, Professor of Biochemistry, Massachusetts Institute of Technology
# Contents: Articles by Topic

## 1 Applications of Chemical Biology

### 1.1 Biomaterials
- Biocompatibility of Materials Used in Medical Devices, Methods For Evaluating
- Controlled Drug Delivery: Pharmacokinetic Considerations, Methods and Systems

### 1.2 Biotechnology
- Directed Evolution, Novel and Improved Enzymes through

### 1.3 Chemical Biology of Disease
- Anxiety Disorders: Macromolecular Pathways and Interactions
- Chronic Obstructive Pulmonary Disease (COPD), Inflammatory Mechanisms of
- Depression, Chemical Mechanisms of
- Hemozoin: A Paradigm for Biominerals in Disease
- Hypoxic Response and Associated Diseases
- Lysosomal Disorders
- Metabolic Diseases, Biological Mechanisms of
- Mitochondrial Medicine, Biochemical Evaluation of Mitochondrial Function
- Protein Misfolding and Disease, Chemical Biology of
- Protein Trafficking Diseases, Small Molecule Approaches to
- Saliva in Health and Disease, Chemical Biology of
- Schizophrenia, Biological Mechanisms of
- Systems Approach to Studying Disease

### 1.4 Pharmacology and Drug Discovery
- ADME Properties of Drugs
- Allergy and Asthma, Small Molecules to Target
- Lead Optimization in Drug Discovery
- Metals in Medicine, Imaging Agents
- Natural Products as Anticancer Agents
- Pharmaceutical Industry, Biocatalysts and Chemo catalysts
- Plant-Derived Natural Products Research, Recent Progress in Drug Discovery Based on
- Platinum Anticancer Drugs, Chemical Biology of
- Programmed Cell Death as a Therapeutic Approach
- RNA Interference to Treat Human Diseases, Applications of
- Transcription-Based Therapeutics

## 2 Biomolecules within the Cell

### 2.1 Glycans (Carbohydrates)
- Chemistry of Neoglycoproteins
- Glycan Biosynthesis in Mammals
- Glycosphingolipids, Chemistry of
- Starch Biosynthesis In Plants

### 2.2 Lipids
- Chemistry of Lipid Domains
- Essential Fatty Acids, Physiologic and Clinical Significance of
- Lipid Bilayers, Properties of
- Lipids, Chemical Diversity of
- Lipids, Organization and Aggregation of
- Lipids, Phase Transitions of

### 2.3 Natural Products
- Alkaloid Biosynthesis
- Chlorophylls and Carotenoids, Chemistry of
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cofactor Biosynthesis</td>
<td>479</td>
</tr>
<tr>
<td>Lantibiotics, Biosynthesis and Mode of Action</td>
<td>494</td>
</tr>
<tr>
<td>Marine Natural Products, Chemical Diversity of</td>
<td>503</td>
</tr>
<tr>
<td>Methy erythritol Phosphate Pathway for the Formation of Isoprene Units</td>
<td>525</td>
</tr>
<tr>
<td>Natural Products: An Overview</td>
<td>533</td>
</tr>
<tr>
<td>Natural Products in Plants, Chemical Diversity of</td>
<td>542</td>
</tr>
<tr>
<td>Pharmaceuticals: Natural Products and Natural Product Models</td>
<td>559</td>
</tr>
<tr>
<td>Polyketide Biosynthesis, Fungi</td>
<td>579</td>
</tr>
<tr>
<td>Polyketide Biosynthesis, Modular Polyketide Synthases</td>
<td>590</td>
</tr>
<tr>
<td>Polyketide Biosynthesis, Polyethers</td>
<td>604</td>
</tr>
<tr>
<td>Terpenoids in Plant Signaling, Chemical Ecology</td>
<td>616</td>
</tr>
<tr>
<td>Terpenoids in Plants</td>
<td>624</td>
</tr>
<tr>
<td>The Biology and Biochemistry of Steroid Hormones</td>
<td>634</td>
</tr>
<tr>
<td><strong>2.4 Nucleic Acids</strong></td>
<td></td>
</tr>
<tr>
<td>Catalytic Modes in Natural Ribozymes</td>
<td>650</td>
</tr>
<tr>
<td>Chemical Damage to Nucleic Acids</td>
<td>661</td>
</tr>
<tr>
<td>DNA Transposition: Topics in Chemical Biology</td>
<td>675</td>
</tr>
<tr>
<td>Group II Introns: Structure, Function, and Catalysis</td>
<td>683</td>
</tr>
<tr>
<td>Noncoding RNAs</td>
<td>691</td>
</tr>
<tr>
<td>RNA Stability, Chemistry, Measurement and Modulation of</td>
<td>702</td>
</tr>
<tr>
<td>Watson–Crick Base Pairs, Character and Recognition of</td>
<td>715</td>
</tr>
<tr>
<td><strong>2.5 Proteins</strong></td>
<td></td>
</tr>
<tr>
<td>Approaches to Enzyme Inhibition</td>
<td>726</td>
</tr>
<tr>
<td>B12-Dependent Enzyme Reactions, Chemistry of</td>
<td>745</td>
</tr>
<tr>
<td>Biosynthesis of Nonribosomal Peptides</td>
<td>754</td>
</tr>
<tr>
<td>Chemical Properties of Amino Acids</td>
<td>763</td>
</tr>
<tr>
<td>Chemical Strategies for Enzyme Catalysis</td>
<td>774</td>
</tr>
<tr>
<td>Chemistry of CoA-Dependent Enzymes</td>
<td>785</td>
</tr>
<tr>
<td>Collagen Triple Helix, Stability of</td>
<td>794</td>
</tr>
<tr>
<td>Enzyme Kinetics</td>
<td>805</td>
</tr>
<tr>
<td>Flavin-Mediated Hydroxylation Reactions</td>
<td>813</td>
</tr>
<tr>
<td>Flavoenzymes, Chemistry of</td>
<td>823</td>
</tr>
<tr>
<td>Membrane Proteins, Properties of</td>
<td>834</td>
</tr>
<tr>
<td>NAD(P) Dependent Dehydrogenases</td>
<td>844</td>
</tr>
<tr>
<td>Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of</td>
<td>854</td>
</tr>
<tr>
<td>Oxygen-Activating Enzymes, Chemistry of</td>
<td>884</td>
</tr>
<tr>
<td>PLP-Dependent Enzymes, Chemistry of</td>
<td>896</td>
</tr>
<tr>
<td>Proteins: Computational Analysis of Structure, Function, and Stability</td>
<td>902</td>
</tr>
<tr>
<td>Proteins, Chemistry and Chemical Reactivity of</td>
<td>910</td>
</tr>
<tr>
<td>The Chemistry of Metallo-Enzymes and Metallo-Proteins</td>
<td>926</td>
</tr>
<tr>
<td>Transient State Enzyme Kinetics</td>
<td>934</td>
</tr>
<tr>
<td><strong>2.6 Water</strong></td>
<td></td>
</tr>
<tr>
<td>Electron Transfer, Chemical Roles of Water in</td>
<td>943</td>
</tr>
<tr>
<td>Hydration Forces</td>
<td>952</td>
</tr>
<tr>
<td>Nucleic Acids, Hydration of</td>
<td>961</td>
</tr>
<tr>
<td>Roles of Water in Biological Recognition Processes</td>
<td>969</td>
</tr>
<tr>
<td>Water, Properties of</td>
<td>977</td>
</tr>
</tbody>
</table>

### Chemical Biology of Cellular Compartments

#### 3.1 Chemical Biology of Cellular Compartments
Endoplasmic Reticulum: Topics in Chemical Biology 986

#### 3.2 Concepts of Cell Organization
Cellular Organization of Metabolism 994
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Subtitle</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Dynamics of Cellular Compartments</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cellular Membranes, Dynamics of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Membrane Compartments Related to Signaling and Trafficking in Immune Cells</td>
<td></td>
<td>1002</td>
</tr>
<tr>
<td></td>
<td>Membranes, Fluidity of</td>
<td></td>
<td>1012</td>
</tr>
<tr>
<td></td>
<td>Microtubule Dynamics</td>
<td></td>
<td>1023</td>
</tr>
<tr>
<td></td>
<td>Receptor-Mediated Endocytosis</td>
<td></td>
<td>1037</td>
</tr>
<tr>
<td>4</td>
<td>Chemical Views of Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Inorganic Chemistry in Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inorganic Chemistry in Biology</td>
<td></td>
<td>1058</td>
</tr>
<tr>
<td>4.2</td>
<td>Nature: A Model System for Chemists</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nature: A Model System for Chemists</td>
<td></td>
<td>1067</td>
</tr>
<tr>
<td>4.3</td>
<td>Physical Chemistry in Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Physical Chemistry in Biology</td>
<td></td>
<td>1075</td>
</tr>
<tr>
<td>5</td>
<td>Chemistry of Biological Processes and Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Cell Cycle</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell Cycle, Regulation of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell Death, Biological Mechanisms and Small Molecule Inhibitors of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cell Division, Small Molecules to Study</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chromosome Formation</td>
<td></td>
<td>1082</td>
</tr>
<tr>
<td>5.2</td>
<td>DNA Damage and Repair</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chemistry of DNA Damage Repair</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sensing of DNA Damage</td>
<td></td>
<td>1095</td>
</tr>
<tr>
<td>5.3</td>
<td>DNA Replication and Recombination</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bacterial DNA polymerases, Chemistry of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DNA Helicases, Chemistry and Mechanisms of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nonhomologous Recombination</td>
<td></td>
<td>1113</td>
</tr>
<tr>
<td>5.4</td>
<td>Membrane Transport</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ca^{2+}-ATPase/Ca^{2+} Pump, Chemistry of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chemistry of Bacteriorhodopsin</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drug Transport in Living Systems</td>
<td></td>
<td>1130</td>
</tr>
<tr>
<td></td>
<td>Ion Channels in Medicine</td>
<td></td>
<td>1141</td>
</tr>
<tr>
<td></td>
<td>Mechanosensitive Channels</td>
<td></td>
<td>1148</td>
</tr>
<tr>
<td></td>
<td>Passive Diffusion Across Membranes</td>
<td></td>
<td>1159</td>
</tr>
<tr>
<td>5.5</td>
<td>Metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chemistry of Lipid Homeostasis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Electron Transfer Chemistry in Photosynthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metabolomics: Topics in Chemical Biology</td>
<td></td>
<td>1179</td>
</tr>
<tr>
<td></td>
<td>Pentose Phosphate Pathway: An Overview</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Systems Approach to Metabolism</td>
<td></td>
<td>1186</td>
</tr>
<tr>
<td></td>
<td>Xenobiotic Metabolism</td>
<td></td>
<td>1191</td>
</tr>
<tr>
<td>5.6</td>
<td>Metals in Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cytochrome P450 Monooxygenases, Chemistry of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hemes in Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Human Hemoglobin: Identification of a Key Intermediate</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Metal Homeostasis</td>
<td></td>
<td>1228</td>
</tr>
<tr>
<td></td>
<td>Metalloregulatory Proteins</td>
<td></td>
<td>1237</td>
</tr>
<tr>
<td>5.7</td>
<td>Molecular Recognition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carbohydrate–Carbohydrate Interactions</td>
<td></td>
<td>1244</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1254</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1266</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1274</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1285</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1292</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1302</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1310</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1320</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1330</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td><strong>Neurochemistry</strong>&lt;br&gt;Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for Neurotransmission, Measuring Chemical Events in Neurotransmitter: Release</td>
<td>1361</td>
<td></td>
</tr>
<tr>
<td>5.10</td>
<td><strong>Post-Translational Modifications</strong>&lt;br&gt;Chemistry of AdoMet-Dependent Methyltransferases Chemistry of Fatty Acid Desaturases Chemistry of Protein Kinases and Dephosphorylases Cytosolic Glycosylated Proteins, Chemistry of Glycosylation of Proteins in the Golgi Apparatus Glycosyltransferases, Chemistry of Post-Translational Modifications to Regulate Protein Function</td>
<td>1424</td>
<td></td>
</tr>
<tr>
<td>5.11</td>
<td><strong>Protein Folding</strong>&lt;br&gt;Molecular Chaperones Protein Misfolding, Amyloid Formation</td>
<td>1516</td>
<td></td>
</tr>
<tr>
<td>5.12</td>
<td><strong>Proteolysis</strong>&lt;br&gt;Intramembrane Proteolysis Metalloproteinases, Biophysics and Chemistry of Protease Inhibitors, Mechanisms of Serine Protease and Serine Protease Inhibitors Thrombin</td>
<td>1531</td>
<td></td>
</tr>
<tr>
<td>5.13</td>
<td><strong>Self-Organization</strong>&lt;br&gt;Nucleobase Self-Assembly Codes, Constitutional Transcription of</td>
<td>1580</td>
<td></td>
</tr>
<tr>
<td>5.14</td>
<td><strong>Sensing</strong>&lt;br&gt;Olfaction: Topics in Chemical Biology Protein and DNA Thermostability, Physics and Evolution of Symbiosis: Topics in Chemical Biology Taste: Topics in Chemical Biology</td>
<td>1591</td>
<td></td>
</tr>
<tr>
<td>5.15</td>
<td><strong>Signal Transduction and Cellular Communication</strong>&lt;br&gt;Bacterial Resistance to Antibiotics Calcium Signaling: Encoding and Decoding Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication Extracellular Lipid Signals Hormone Signaling Inositol Phospholipids, Biosynthesis and Biological Functions of Inositol Polyphosphates Integrin Signaling, Bidirectional-Signaling of Isoprostanes Large G-Proteins Mathematical Modeling of Biological Signaling Networks Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s Nitric Oxide, Biological Targets of Nuclear Receptors, Chemistry of Quorum Sensing Ras Oncoproteins, Structure, Biochemistry, and Biology of Signaling Networks in Biology</td>
<td>1637</td>
<td></td>
</tr>
</tbody>
</table>
5.16 Transcription
Attenuation, Control of Gene Expression by 1810
Epigenetic Modifications 1822
Pre-mRNA splicing 1835
RNA Interference, Mechanisms and Proteins Involved in 1845
Transcription Factors 1855

5.17 Translation
Chemistry of Aminoacyl-tRNA Synthetases 1873
Translational Recoding 1885

6 Synthetic Molecules as Tools for Chemical Biology
6.1 Small Molecules
Chemical Libraries: Screening for Biologically Active Small Molecules 1895
Cyclooxygenase Inhibition, Recent Advances in the Mechanisms of 1906
Forward Chemical Genetics 1914
Metallointercalators as Probes of DNA Recognition and Reactions 1931
Phosphate Mimics, Cyclic Compounds as 1943
Small Molecule Inhibitors, Design and Selection of 1958
Small Molecules to Control Stem Cell Fates 1971

6.2 Synthetic Biology and Chemistry
Solid-Phase Synthesis of Biomolecules 1979

6.3 Synthetic Glycans
Catalytic Antibodies: Past, Present, and Future 1987
Enzymatic Synthesis of Carbohydrate-Containing Biomolecules 2004
Key Strategies for Glycan Synthesis 2026

6.4 Synthetic Lipids
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function 2038

6.5 Synthetic Nucleic Acids
In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes) 2063
Locked Nucleic Acids 2080
Nucleosides in RNA, Modified 2087
Peptide Nucleic Acids 2101

6.6 Synthetic Proteins to Explore Biology
Artificial Metalloproteins, Design and Engineering of 2111
Click Peptides, Design and Applications of 2121
Coupling Methods for Peptide Synthesis 2126
Lipidated Peptide Synthesis 2135
Peptide Combinatorial Libraries 2146
Peptidomimetics 2159
Synthetic Chemistry: Formation of the Amide Bond 2171
Synthetic Peptides and Proteins to Elucidate Biological Function 2188
Unnatural Amino Acids to Investigate Biologic Processes 2199

7 Technologies and Techniques in Chemical Biology
7.1 Cells and Complex Cellular Systems: Methods for Characterization
Membrane Potentials in Living Systems, Tools to Measure 2210
NMR-Based Metabolomics Technology in Biomarker Research 2220

7.2 General Techniques, Overview of Applications in Chemical Biology
Combinatorial Libraries: Overview of Applications in Chemical Biology 2233
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology 2243
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3</td>
<td>Glycans: Methods for Characterization</td>
<td>Array-Based Techniques for Glycans</td>
<td>2328</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glycan Sequencing and Data Integration for Glycomics</td>
<td>2340</td>
</tr>
<tr>
<td>7.4</td>
<td>Lipids: Methods for Characterization</td>
<td>Differential Scanning Calorimetry to Study Lipids and Lipid Membranes</td>
<td>2352</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lipidomics</td>
<td>2364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR of Lipids</td>
<td>2373</td>
</tr>
<tr>
<td>7.5</td>
<td>Nucleic Acids: Methods for Characterization</td>
<td>DNA Polymorphisms, Tools for Detection of</td>
<td>2383</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fluorescence Techniques for Nucleic Acids</td>
<td>2395</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcript Profiling, Tools for</td>
<td>2404</td>
</tr>
<tr>
<td>7.6</td>
<td>Proteins: Methods for Characterization</td>
<td>Electron Paramagnetic Resonance in Enzymology</td>
<td>2414</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fluorescence Techniques for Proteins</td>
<td>2425</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Förster Resonance Energy Transfer (FRET) for Proteins</td>
<td>2436</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mass Spectrometry, Applications in Phosphoproteomics</td>
<td>2451</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Membrane Protein Structure, Techniques to Study</td>
<td>2458</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMR for Proteins</td>
<td>2471</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Two- and Three-Hybrid Systems</td>
<td>2482</td>
</tr>
<tr>
<td>7.7</td>
<td>Small Molecules: Methods for Characterization</td>
<td>Computational Approaches in Drug Discovery and Development</td>
<td>2496</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pharmacokinetics of Drug Candidates, Determination of</td>
<td>2505</td>
</tr>
</tbody>
</table>
### ADME Properties of Drugs

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alkaloid Biosynthesis</td>
<td>452</td>
</tr>
<tr>
<td>Allergy and Asthma, Small Molecules to Target</td>
<td>196</td>
</tr>
<tr>
<td>Anxiety Disorders: Macromolecular Pathways and Interactions</td>
<td>43</td>
</tr>
<tr>
<td>Approaches to Enzyme Inhibition</td>
<td>726</td>
</tr>
<tr>
<td>Array-Based Techniques for Glycans</td>
<td>2328</td>
</tr>
<tr>
<td>Artificial Metalloproteins, Design and Engineering of</td>
<td>2111</td>
</tr>
<tr>
<td>Attenuation, Control of Gene Expression by</td>
<td>1810</td>
</tr>
</tbody>
</table>

### B12-Dependent Enzyme Reactions, Chemistry of

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial DNA polymerases, Chemistry of</td>
<td>1148</td>
</tr>
<tr>
<td>Bacterial Resistance to Antibiotics</td>
<td>1637</td>
</tr>
<tr>
<td>Biocompatibility of Materials Used in Medical Devices, Methods For Evaluating</td>
<td>15</td>
</tr>
<tr>
<td>Biosynthesis of Nonribosomal Peptides</td>
<td>754</td>
</tr>
</tbody>
</table>

### Ca^{2+}-ATPase/Ca^{2+} Pump, Chemistry of

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium Signaling: Encoding and Decoding</td>
<td>1656</td>
</tr>
<tr>
<td>Carbohydrate–Carbohydrate Interactions</td>
<td>1330</td>
</tr>
<tr>
<td>Catalytic Antibodies: Past, Present, and Future</td>
<td>1987</td>
</tr>
<tr>
<td>Catalytic Modes in Natural Ribozymes</td>
<td>650</td>
</tr>
<tr>
<td>Cell Cycle, Regulation of</td>
<td>1082</td>
</tr>
<tr>
<td>Cell Death, Biological Mechanisms and Small Molecule Inhibitors of</td>
<td>1095</td>
</tr>
<tr>
<td>Cell Division, Small Molecules to Study</td>
<td>1113</td>
</tr>
<tr>
<td>Cellular Membranes, Dynamics of</td>
<td>1002</td>
</tr>
<tr>
<td>Cellular Organization of Metabolism</td>
<td>994</td>
</tr>
<tr>
<td>Chemical Damage to Nucleic Acids</td>
<td>661</td>
</tr>
<tr>
<td>Chemical Libraries: Screening for Biologically Active Small Molecules</td>
<td>1895</td>
</tr>
<tr>
<td>Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication</td>
<td>1666</td>
</tr>
<tr>
<td>Chemical Properties of Amino Acids</td>
<td>763</td>
</tr>
<tr>
<td>Chemical Strategies for Enzyme Catalysis</td>
<td>774</td>
</tr>
<tr>
<td>Chemistry of CoA-Dependent Enzymes</td>
<td>785</td>
</tr>
<tr>
<td>Chemistry of DNA Damage Repair</td>
<td>1130</td>
</tr>
<tr>
<td>Chemistry of AdoMet-Dependent Methyltransferases</td>
<td>1424</td>
</tr>
<tr>
<td>Chemistry of Aminoacyl-tRNA Synthetases</td>
<td>1873</td>
</tr>
<tr>
<td>Chemistry of Bacteriorhodopsin</td>
<td>1186</td>
</tr>
<tr>
<td>Chemistry of Fatty Acid Desaturases</td>
<td>1434</td>
</tr>
<tr>
<td>Chemistry of Lipid Domains</td>
<td>371</td>
</tr>
<tr>
<td>Chemistry of Lipid Homeostasis</td>
<td>1228</td>
</tr>
<tr>
<td>Chemistry of Neoglycoproteins</td>
<td>310</td>
</tr>
<tr>
<td>Chemistry of Protein Kinases and Dephosphorylases</td>
<td>1441</td>
</tr>
<tr>
<td>Chlorophylls and Carotenoids, Chemistry of</td>
<td>468</td>
</tr>
<tr>
<td>Chromosome Formation</td>
<td>1124</td>
</tr>
<tr>
<td>Chronic Obstructive Pulmonary Disease (COPD), Inflammatory Mechanisms of</td>
<td>58</td>
</tr>
<tr>
<td>Click Peptides, Design and Applications of</td>
<td>2121</td>
</tr>
<tr>
<td>Cofactor Biosynthesis</td>
<td>479</td>
</tr>
<tr>
<td>Collagen Triple Helix, Stability of</td>
<td>794</td>
</tr>
<tr>
<td>Combinatorial Libraries: Overview of Applications in Chemical Biology</td>
<td>2233</td>
</tr>
<tr>
<td>Computational Approaches in Drug Discovery and Development</td>
<td>2496</td>
</tr>
<tr>
<td>Controlled Drug Delivery: Pharmacokinetic Considerations, Methods and Systems</td>
<td>23</td>
</tr>
<tr>
<td>Coupling Methods for Peptide Synthesis</td>
<td>2126</td>
</tr>
<tr>
<td>Cyclooxygenase Inhibition, Recent Advances in the Mechanisms of</td>
<td>1906</td>
</tr>
<tr>
<td>Cytochrome P450 Monoxygenases, Chemistry of</td>
<td>1285</td>
</tr>
<tr>
<td>Cytosolic Glycosylated Proteins, Chemistry of</td>
<td>1449</td>
</tr>
</tbody>
</table>
### Depression, Chemical Mechanisms of
- Differential Scanning Calorimetry to Study Lipids and Lipid Membranes 2352
- Directed Evolution, Novel and Improved Enzymes through 33
- DNA Helicases, Chemistry and Mechanisms of 1159
- DNA Polymorphisms, Tools for Detection of 2383
- DNA Transposition: Topics in Chemical Biology 675
- Drug Transport in Living Systems 1191

### Electron Paramagnetic Resonance in Enzymology
- Electron Transfer, Chemical Roles of Water in 943
- Electron Transfer Chemistry in Photosynthesis 1237
- Endoplasmic Reticulum: Topics in Chemical Biology 986
- Enzymatic Synthesis of Carbohydrate-Containing Biomolecules 2004
- Enzyme Kinetics 805
- Epigenetic Modifications 1822
- Essential Fatty Acids, Physiologic and Clinical Significance of 376
- Extracellular Lipid Signals 1677

### Flavin-Mediated Hydroxylation Reactions
- Flavoenzymes, Chemistry of 823
- Fluorescence Techniques for Nucleic Acids 2395
- Fluorescence Techniques for Proteins 2425
- Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology 2243
- Förster Resonance Energy Transfer (FRET) for Proteins 2436
- Forward Chemical Genetics 1914

### Glycan Biosynthesis in Mammals
- Glycan Sequencing and Data Integration for Glycomics 2340
- Glycosphingolipids, Chemistry of 336
- Glycosylation of Proteins in the Golgi Apparatus 1457
- Glycosyltransferases, Chemistry of 1472
- Group II Introns: Structure, Function, and Catalysis 683

### Hemes in Biology
- Hemozoin: A Paradigm for Biominerals in Disease 82
- High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology 2266
- Hormone Signaling 1684
- Human Hemoglobin: Identification of a Key Intermediate 1302
- Hydration Forces 952
- Hypoxic Response and Associated Diseases 93

### Inorganic Chemistry in Biology
- Inositol Phospholipids, Biosynthesis and Biological Functions of 1694
- Inositol Polyphosphates 1700
- Integrin Signaling, Bidirectional-Signaling of 1712
- Intramembrane Proteolysis 1531
- In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes) 2063
- Ion Channels in Medicine 1201
- Iron-Sulfur World, Origin of life in an 1397
- Isoprostanestes 1725

### Key Strategies for Glycan Synthesis

### Lantibiotics, Biosynthesis and Mode of Action
- Large G-Proteins 1734
- Lead Optimization in Drug Discovery 204
- Lipidated Peptide Synthesis 2135
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipid Bilayers, Properties of</td>
<td>395</td>
</tr>
<tr>
<td>Lipidomics</td>
<td>2364</td>
</tr>
<tr>
<td>Lipids, Chemical Diversity of</td>
<td>410</td>
</tr>
<tr>
<td>Lipids, Organization and Aggregation of</td>
<td>424</td>
</tr>
<tr>
<td>Lipids, Phase Transitions of</td>
<td>437</td>
</tr>
<tr>
<td>Locked Nucleic Acids</td>
<td>2080</td>
</tr>
<tr>
<td>Lysosomal Disorders</td>
<td>108</td>
</tr>
<tr>
<td>Marine Natural Products, Chemical Diversity of</td>
<td>503</td>
</tr>
<tr>
<td>Mass Spectrometry, Applications in Phosphoproteomics</td>
<td>2451</td>
</tr>
<tr>
<td>Mathematical Modeling of Biological Signaling Networks</td>
<td>1743</td>
</tr>
<tr>
<td>Mechanosensitive Channels</td>
<td>1211</td>
</tr>
<tr>
<td>Membrane Compartments Related to Signaling and Trafficking in Immune</td>
<td>1012</td>
</tr>
<tr>
<td>Cells</td>
<td></td>
</tr>
<tr>
<td>Membrane Potentials in Living Systems, Tools to Measure</td>
<td>2210</td>
</tr>
<tr>
<td>Membrane Proteins, Properties of</td>
<td>834</td>
</tr>
<tr>
<td>Membrane Protein Structure, Techniques to Study</td>
<td>2458</td>
</tr>
<tr>
<td>Membranes, Fluidity of</td>
<td>1023</td>
</tr>
<tr>
<td>Metabolic Diseases, Biological Mechanisms of</td>
<td>119</td>
</tr>
<tr>
<td>Metabolomics: Topics in Chemical Biology</td>
<td>1244</td>
</tr>
<tr>
<td>Metal Homeostasis</td>
<td>1310</td>
</tr>
<tr>
<td>Metallointercalators as Probes of DNA Recognition and Reactions</td>
<td>1931</td>
</tr>
<tr>
<td>Metalloproteinases, Biophysics and Chemistry of</td>
<td>1544</td>
</tr>
<tr>
<td>Metalloregulatory Proteins</td>
<td>1320</td>
</tr>
<tr>
<td>Metals in Medicine, Imaging Agents</td>
<td>213</td>
</tr>
<tr>
<td>Methylerithrol Phosphate Pathway for the Formation of Isoprene Units</td>
<td>525</td>
</tr>
<tr>
<td>Microarrays in Chemical Biology</td>
<td>2281</td>
</tr>
<tr>
<td>Microtubule Dynamics</td>
<td>1037</td>
</tr>
<tr>
<td>Mitochondrial Medicine, Biochemical Evaluation of Mitochondrial</td>
<td>132</td>
</tr>
<tr>
<td>Function</td>
<td></td>
</tr>
<tr>
<td>Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s</td>
<td>1752</td>
</tr>
<tr>
<td>Molecular Chaperones</td>
<td>1516</td>
</tr>
<tr>
<td>Molecular Recognition, Computation and Modeling of</td>
<td>1337</td>
</tr>
<tr>
<td>NAD(P) Dependent Dehydrogenases</td>
<td>844</td>
</tr>
<tr>
<td>Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis</td>
<td>854</td>
</tr>
<tr>
<td>of</td>
<td></td>
</tr>
<tr>
<td>Natural Products: An Overview</td>
<td>533</td>
</tr>
<tr>
<td>Natural Products as Anticancer Agents</td>
<td>222</td>
</tr>
<tr>
<td>Natural Products in Plants, Chemical Diversity of</td>
<td>542</td>
</tr>
<tr>
<td>Nature: A Model System for Chemists</td>
<td>1067</td>
</tr>
<tr>
<td>Neuropeptides, Chemical Activity Profiling and Proteomic Approaches</td>
<td>1361</td>
</tr>
<tr>
<td>for Neurtransmission, Measuring Chemical Events in</td>
<td>1373</td>
</tr>
<tr>
<td>Neurotransmitter: Release</td>
<td>1385</td>
</tr>
<tr>
<td>Nitric Oxide, Biological Targets of</td>
<td>1760</td>
</tr>
<tr>
<td>NMR-Based Metabolomics Technology in Biomarker Research</td>
<td>2220</td>
</tr>
<tr>
<td>NMR for Proteins</td>
<td>2471</td>
</tr>
<tr>
<td>NMR of Lipids</td>
<td>2373</td>
</tr>
<tr>
<td>Noncoding RNAs</td>
<td>691</td>
</tr>
<tr>
<td>Nonhomologous Recombination</td>
<td>1170</td>
</tr>
<tr>
<td>Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of</td>
<td>2292</td>
</tr>
<tr>
<td>Applications in Chemical Biology</td>
<td></td>
</tr>
<tr>
<td>Nuclear Receptors, Chemistry of</td>
<td>1768</td>
</tr>
<tr>
<td>Nucleic Acids, Hydration of</td>
<td>961</td>
</tr>
<tr>
<td>Nucleobase Self-Assembly Codes, Constitutional Transcription of</td>
<td>1580</td>
</tr>
<tr>
<td>Nucleosides in RNA, Modified</td>
<td>2087</td>
</tr>
<tr>
<td>Olfaction: Topics in Chemical Biology</td>
<td>1591</td>
</tr>
<tr>
<td>Origins of Life: Emergence of Amino Acids</td>
<td>1405</td>
</tr>
<tr>
<td>Origins of Life: Emergence of the RNA World</td>
<td>1412</td>
</tr>
<tr>
<td>Oxygen-Activating Enzymes, Chemistry of</td>
<td>884</td>
</tr>
</tbody>
</table>
## Passive Diffusion Across Membranes

- Pentose Phosphate Pathway: An Overview
- Peptide Combinatorial Libraries
- Peptide Nucleic Acids
- Peptidomimetics
- Pharmaceutical Industry, Biocatalysts and Chemocatalysts
- Pharmaceuticals: Natural Products and Natural Product Models
- Pharmacokinetics of Drug Candidates, Determination of
- Phosphate Mimics, Cyclic Compounds as
- Physical Chemistry in Biology
- Plant-Derived Natural Products Research, Recent Progress in Drug Discovery Based on
- Platinum Anticancer Drugs, Chemical Biology of
- PLP-Dependent Enzymes, Chemistry of
- Polyketide Biosynthesis, Fungi
- Polyketide Biosynthesis, Modular Polyketide Synthases
- Polyketide Biosynthesis, Polyethers
- Post-Translational Modifications to Regulate Protein Function
- Pre-mRNA splicing
- Programmed Cell Death as a Therapeutic Approach
- Protease Inhibitors, Mechanisms of
- Protein and DNA Thermostability, Physics and Evolution of
- Protein Misfolding, Amyloid Formation
- Protein Misfolding and Disease, Chemical Biology of
- Proteins: Computational Analysis of Structure, Function, and Stability
- Proteins, Chemistry and Chemical Reactivity of
- Protein Trafficking Diseases, Small Molecule Approaches to

## Quorum Sensing

## Ras Oncoproteins, Structure, Biochemistry, and Biology of

- Receptor–Ligand Interactions in Biological Systems
- Receptor-Mediated Endocytosis
- RNA Interference, Mechanisms and Proteins Involved in
- RNA Interference to Treat Human Diseases, Applications of
- RNA Stability, Chemistry, Measurement and Modulation of
- Roles of Water in Biological Recognition Processes

## Saliva in Health and Disease, Chemical Biology of

- Schizophrenia, Biological Mechanisms of
- Sensing of DNA Damage
- Serine Protease and Serine Protease Inhibitors
- Signaling Networks in Biology
- Small Molecule Inhibitors, Design and Selection of
- Small Molecules to Control Stem Cell Fates
- Solid-Phase Synthesis of Biomolecules
- Starch Biosynthesis In Plants
- Supported Lipid Bilayers: Development and Applications in Chemical Biology
- Symbiosis: Topics in Chemical Biology
- Synthetic Chemistry: Formation of the Amide Bond
- Synthetic Peptides and Proteins to Elucidate Biological Function
- Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function
- Systems Approach to Metabolism
- Systems Approach to Studying Disease

## Taste: Topics in Chemical Biology

- Terpenoids in Plant Signaling, Chemical Ecology
- Terpenoids in Plants
- The Biology and Biochemistry of Steroid Hormones
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Chemistry of Metallo-Enzymes and Metallo-Proteins</td>
<td>926</td>
</tr>
<tr>
<td>Thrombin</td>
<td>1575</td>
</tr>
<tr>
<td>Transcription-Based Therapeutics</td>
<td>300</td>
</tr>
<tr>
<td>Transcription Factors</td>
<td>1855</td>
</tr>
<tr>
<td>Transcript Profiling, Tools for</td>
<td>2404</td>
</tr>
<tr>
<td>Transient State Enzyme Kinetics</td>
<td>934</td>
</tr>
<tr>
<td>Translational Recoding</td>
<td>1885</td>
</tr>
<tr>
<td>Two- and Three-Hybrid Systems</td>
<td>2482</td>
</tr>
</tbody>
</table>

**Unnatural Amino Acids to Investigate Biologic Processes**  

**Water, Properties of**  
Watson–Crick Base Pairs, Character and Recognition of  

**Xenobiotic Metabolism**  

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnatural Amino Acids to Investigate Biologic Processes</td>
<td>2199</td>
</tr>
<tr>
<td>Water, Properties of</td>
<td>977</td>
</tr>
<tr>
<td>Watson–Crick Base Pairs, Character and Recognition of</td>
<td>715</td>
</tr>
<tr>
<td>Xenobiotic Metabolism</td>
<td>1274</td>
</tr>
</tbody>
</table>
Medical devices are used for a variety of functions in humans. This review elaborates on the types of tests used to evaluate biocompatibility of the interactions between man-made medical devices and host tissues and organs. The outcome of the response depends on the site of implantation, the species of the host, the genetic makeup of the host, the sterility of the implant, and the effect the device has on biological processes. Biological processes involved in host tissue responses to implantable medical devices reflect activation of a series of cascades that require blood proteins or other components found in the blood. Two types of regulatory approvals for medical devices exist in the United States, 510(k) notification and premarket approval (PMA). The specific tests required prior to regulatory approval vary with the type of device and application; however, some general testing is usually recommended. Normally, animal testing is conducted to demonstrate that a medical device is safe, and when implanted in humans the device will reduce, alleviate, or eliminate the possibility of adverse medical reactions or conditions. The American Society for Testing and Materials (ASTM) as well as the International Organization for Standardization (ISO) publishes standards for testing medical devices. The recommended tests include culture cytotoxicity, skin irritation, short-term intramuscular implantation, short-term subcutaneous implantation, mucous membrane irritation, systemic injection, sensitization assays, and mutagenicity testing.

Introduction—What Are Medical Devices?

Medical devices are used for a variety of functions from promoting the healing of small wounds using adhesive bandages to maintaining the flow of blood through arteries narrowed by atherosclerosis using metallic vascular stents. The purpose of this article is not to give an overview of the many devices used in medicine to promote wellness and homeostasis but to elaborate on the types of tests used to evaluate the interactions between man-made devices and host tissues and organs. These man-made devices are called medical devices in the United States and are regulated for interstate distribution by the Federal Food and Drug Administration (FDA). In many cases, medical devices consist of assemblies of polymers, metals, ceramics, and composites that are used in diagnostic procedures and as implants in animals and in humans. In the United States, extensive biocompatibility testing occurs before the devices are marketed to the general public. Prior to 1976, no federal regulations existed to oversee the sale and uses of medical devices in the United States. In 1976, the U.S. Congress enacted the Medical Device Amendments to the Federal, Food, Drug, and Cosmetic Act of 1938, which called for the establishment of three classes of medical devices (Table 1a). Class 1 devices are those that present little or no risk to the user, whereas Class 2 and 3 devices present some risk and a high degree of risk to the user, respectively. These devices are regulated by requiring limited animal testing (Class 2) and extensive animal and human testing (Class 3). The European Union has a system of device classification similar to the United States as recognized by the Medical Device Directive (Table 1b).

The term biocompatible is used widely to infer that an implant is safe for use in the general population. Although this term is used broadly, it may be a misnomer because only materials that are found in living tissues are truly biocompatible. Below, we examine what methods are used to evaluate the biocompatibility of materials used in medical devices.
Biocompatibility of Materials Used in Medical Devices, Methods For Evaluating

<table>
<thead>
<tr>
<th>Class</th>
<th>Type of device</th>
<th>FDA filing required</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Crutches, bedpans, depressors, adhesive bandages, hospital beds</td>
<td>PMN/510K</td>
</tr>
<tr>
<td></td>
<td>Hearing aids, blood pumps, catheters contact lens, electrodes</td>
<td>510(k)</td>
</tr>
<tr>
<td>II</td>
<td>Cardiac pacemakers, intrauterine devices, intraocular lens, heart valves, orthopedic devices</td>
<td>PMA</td>
</tr>
</tbody>
</table>

PMN, premarket notification.

<table>
<thead>
<tr>
<th>Class</th>
<th>Type of device</th>
<th>Regulatory requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Stethoscopes, hospital beds, wheelchairs</td>
<td>Technical file, other assurances</td>
</tr>
<tr>
<td>IIA</td>
<td>Hearing aids, electrographs, ultrasound assessment equipment</td>
<td>Technical file, conformity</td>
</tr>
<tr>
<td>IIB</td>
<td>Surgical lasers, infusion pumps, ventilators, intensive care monitoring equipment</td>
<td>Technical file, type examination</td>
</tr>
<tr>
<td>III</td>
<td>Balloon catheters, heart valves</td>
<td>Audit of quality assurance, examination of design</td>
</tr>
</tbody>
</table>

Biocompatibility—What Is Does it Mean?

The word biocompatibility is a relative term that means that the materials used in a medical device do not elicit a reaction that either 1) makes the device not perform its intended use or 2) causes reactions that affect the functioning and health of the host. All materials used in devices will elicit a response from the host; it could be an immediate response, one that is prolonged, or even a delayed reaction that occurs sometime after contact with the device. The outcome of the response depends on the site of implantation, the species of the host, the genetic makeup of the host, and the sterility of the implant. All implants have a significantly greater rate of infection when compared with the background rate associated with the surgical procedure performed in the absence of the device. At the very least, an implant should not interfere with biological processes that are required for normal homeostasis of the host.

Biological Systems—Which Ones Are Important for Normal Homeostasis and Survival?

Devices in contact with the external tissues such as skin typically are considered separately from a biocompatibility perspective from devices implanted internally. Implantable devices affect biological processes that involve blood; therefore, the testing of these devices is somewhat more complicated. Many skin contact devices are used short term, and therefore biocompatibility testing is limited. However, for permanent internal implants, the required testing can be as long as several years and require analysis of the effects of the device on cells and tissues as well as on healing responses that occur at the interface between the tissue and the device. For this reason, it is important to understand which biological systems may be affected when permanent implants are to be used.

Biological processes involved in host-tissue responses to implantable medical devices reflect the activation of a series of cascades that require blood proteins or other components found in blood. Biological systems activated by implants include blood clotting, platelet aggregation, complement activation, kinin formation, fibrinolysis, phagocytosis, immune responses, and wound healing (1) (Table 2). Wound healing involves several biological processes, including blood clotting, inflammation, dilation of neighboring blood vessels, accumulation of blood cells and fluid at the point of contact, and finally deposition of fibrous tissue around an implant. Vasodilation of blood vessels and accumulation of interstitial fluid around an implant can occur through activation of the kinin and complement pathways (3). Phagocytosis of dead tissue occurs by attraction and migration of inflammatory cells to the site of injury near an implant. The inflammatory cells attracted include neutrophils and monocytes that are present to digest dead tissue and implant materials. Once phagocytosis occurs, it may lead to digestion of implant remnants and formation of fibrous scar tissue around the implant. If a large blood clot surrounds an implant, then fibrinolysis must proceed to remove the clotted blood before the healing process can be completed (1).
Blood proteins are involved in the lysis of foreign cells via the complement pathway (1). This mechanism involves activation of complement proteins in the presence of an antibody–antigen complex attached to the surface of a foreign cell. Components of the complement pathway are sometimes compromised by activation and/or adsorption onto the surface of a medical device. This action leads to complement component depletion that causes the patient to be at risk for bacterial infection and makes evaluation of complement depletion important aspect of the design of cardiovascular devices. Activated complement components also prolong inflammation by generating C3a and C5a, which are agents that cause vasodilation. Complement activation is associated with and contributes to whole-body inflammation, which is observed as a complication to cardiopulmonary bypass. Complement activation is responsible for hyperacute rejection of animal tissue grafts (2) and is important in reactions to implants (3–5).

Most foreign surfaces cause blood to clot as a result of direct contact with a foreign surface. This clotting occurs via the intrinsic clotting cascade or from injury to tissue that develops during implantation as result of activation of Hageman factor and factor IX, which are two proteins found in blood (Table 2). Platelets, which are enucleated cells, are also found in blood; they release factors that contribute to formation of blood clots. Devices used in the cardiovascular system normally are designed to limit their propensity to clot blood. In the case of cardiovascular devices, excessive blood clotting will cause the device to occlude; in these applications, blood clotting is minimized. Because foreign materials typically cause blood clots, they are only used to replace large and medium-sized vessels. Host vessels are used to replace the function of small-diameter vessels. Several tests are used to measure blood clotting and platelet aggregation caused by contact with a medical device (6–8).

In addition to activating blood clotting (9), activated Hageman factor activates prekallikrein of the kinin system, which leads to bradykinin that causes vascular vasodilation. Activation of Hageman factor and blood clotting also leads to the conversion of plasminogen to plasmin which initiates the degradation of fibrin formed during clotting by a process termed fibrinolysis (1).

Phagocytic cells including neutrophils and macrophages, coat medical devices either from direct blood contact or via inflammation and extravascular movement of these cells into the tissue fluids that surround a device. In either situation, first neutrophils and then monocytes arrive in the area around the device and attempt to degrade the implant. If the implant is biodegradable, then these cells remain until the device is totally removed. If the device is nondegradable, then the number of cells surrounding an implant will depend on the how reactive the implant is. For example, although Dacron vascular grafts are permanent devices, monocytes can be observed surrounding the implant for months and years. In some patients, continued reactivity can cause peri-implant fluid accumulation, which if left uncorrected can require implant removal. In other cases where contact of tissue with the implant causes a prolonged inflammatory response, other white blood cells including eosinophils, B cells, and T cells can be observed in the vicinity of the device. These cells are an indication of either an allergic reaction or the formation of antibodies that stimulate prolonged inflammation. Measurement of inflammatory cells surrounding an implant is usually accomplished by direct histological evaluation (10–12).

As phagocytic cells accumulate near the implant, they elaborate hydrolytic enzymes that degrade both the implant and the surrounding tissues; fibroblasts and endothelial cells are also migrating into the area around the device and begin to lay down new extracellular matrix with capillaries and collagen fibrils. Thus, the wound healing process involves inflammation, removal of the implant and tissue components, as well as the deposition of new extracellular matrix. If the implant is nondegradable and nonporous, then a fibrous capsule forms around it. The thickness of the fibrous capsule depends on the degree of inflammation caused by the device. If the implant is porous, the device may biodegrade and lead to the formation of a small amount of fibrous scar tissue in the defect when the implant is removed. In some cases, however, after the implant biodegrades, an abundance of scar tissue can be deposited where the implant

<table>
<thead>
<tr>
<th>System</th>
<th>Function</th>
<th>Device effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood clotting</td>
<td>Maintains blood fluidity</td>
<td>Clot formation–occlusion</td>
</tr>
<tr>
<td>Complement</td>
<td>Prevents bacterial invasion</td>
<td>Depletes complement</td>
</tr>
<tr>
<td>Fibrinolysis</td>
<td>Degrades blood clots</td>
<td>Degrades tissue grafts</td>
</tr>
<tr>
<td>Immune responses</td>
<td>Limits infection</td>
<td>Prolongs inflammation</td>
</tr>
<tr>
<td>Kinin formation</td>
<td>Causes vasodilation</td>
<td>Prolongs inflammation</td>
</tr>
<tr>
<td>Platelet aggregation</td>
<td>Limits bleeding</td>
<td>Shortens platelet life</td>
</tr>
<tr>
<td>Phagocytosis</td>
<td>Limits infection</td>
<td>Prolongs inflammation</td>
</tr>
<tr>
<td>Wound healing</td>
<td>Repairs tissue defects</td>
<td>Promotes fibrous scar</td>
</tr>
</tbody>
</table>

Table 2 Biological systems affected by medical devices (1)

Blood clotting Maintains blood fluidity Clot formation–occlusion
Complement Prevents bacterial invasion Depletes complement
Fibrinolysis Degrades blood clots Degrades tissue grafts
Immune responses Limits infection Prolongs inflammation
Kinin formation Causes vasodilation Prolongs inflammation
Platelet aggregation Limits bleeding Shortens platelet life
Phagocytosis Limits infection Prolongs inflammation
Wound healing Repairs tissue defects Promotes fibrous scar

**Tissue**
was previously observed. The thickness of the fibrous capsule formed around an implant is usually measured histologically.

Wear particles generated by a moving device can lead to prolonged inflammation and even implant failure in the case of hip and knee implants. Small polymeric or metallic particles, which are about 1 μm in diameter, are ingested by neutrophils and monocytes and may lead to necrosis of these cells and the release of inflammatory mediators into the wound area. Large particles are surrounded by monocytes, which form multinucleated giant cells that can in many cases be tolerated by tissues without leading to implant failure. However, once wear particles are released from the implant, they can migrate to other tissues or even to local lymph nodes causing swelling and systemic problems. Implant wear particles are quantitatively determined from histological and electron-microscopic studies (13–15).

**Types of Tests—What Types of Tests Are Used?**

Two types of regulatory approvals exist for medical devices in the United States, 510(k) notification and premarket approval (PMA). The types of tests required for approval depend on the classification of the medical device. 510(k) notification involves marketing a device that is substantially equivalent to a device on the market prior to 1976. All devices introduced after 1976 that are not substantially equivalent to devices on the market before 1976 are automatically classified as Class 3 devices and require PMA (16). For a device to be considered substantially equivalent to a device on the market before 1976, it must have the same intended use, no new technological characteristics, and have the same performance as one or more devices on the market prior to 1976. All devices introduced after 1976 marketing a device that is substantially equivalent to a device on the market before 1976, it must have PMA (16). For a device to be considered substantially equivalent to a device on the market before 1976, it must have the same intended use, no new technological characteristics, and have the same performance as one or more devices on the market prior to 1976. All devices introduced after 1976 are automatically classified as Class 3 devices and require PMA (16). For a device to be considered substantially equivalent to a device on the market before 1976, it must have the same intended use, no new technological characteristics, and have the same performance as one or more devices on the market prior to 1976. All devices introduced after 1976 are automatically classified as Class 3 devices and require PMA (16). For a device to be considered substantially equivalent to a device on the market before 1976, it must have the same intended use, no new technological characteristics, and have the same performance as one or more devices on the market prior to 1976.

The testing conducted on biomaterials intended for use in medical devices must address safety and effectiveness criteria that depend on the intended use as described above as discussed in depth the literature (18, 19). The specific tests required vary with the type of device and application; however, some general testing is usually recommended. Normally, animal testing is conducted to demonstrate that a medical device is safe, and when implanted in humans that the device will reduce, alleviate, or eliminate the possibility of adverse medical reactions or conditions (17).

According to the American Society of Testing Materials (ASTM) Medical Devices Standards (Annual Book of ASTM Standards, Section 13, Medical Devices, ASTM 1916 Race Street, Philadelphia, PA 19103; available at: www.astm.org), the type of generic biological test methods for materials and devices depends on the end-use application. The ASTM as well as the International Organization for Standardization (ISO) publishes standards for testing medical devices as listed in **Tables 3 and 4.** Biological reactions that are detrimental to the successful use of a material in one device application may not be applicable to the success of a material in a different end use. A list of potentially applicable biocompatibility tests that are related to the end use of a material and/or a device is given in Table 3 as a starting point. These tests are as follows:

**Cell culture cytotoxicity**

This test is used to evaluate the toxicity of a material in *vivo* or an extract of a material used in a device. Several different tests have been used and have produced a spectrum of biocompatibility assessments on the same material (20–22). The tests used measure the viability of cells in contact with a material or an extract of a material. A variety of cell lines can be used; however, a modified fibroblast line is usually the gold standard. Some tests used include 1) direct cell culture, 2) agar diffusion testing, 3) filter diffusion testing, and 4) barrier testing (22).

As pointed out by Learmonth (23), although the intact implant may not be cytotoxic to cells, any material and mechanical

---

**Table 3 Biological tests used to evaluate biocompatibility**

<table>
<thead>
<tr>
<th>Test</th>
<th>ASTM standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell culture cytotoxicity</td>
<td>F748</td>
</tr>
<tr>
<td>Skin irritation</td>
<td>F719</td>
</tr>
<tr>
<td>Intramuscular and subcutaneous implant</td>
<td>F748</td>
</tr>
<tr>
<td>Blood compatibility</td>
<td>F748</td>
</tr>
<tr>
<td>Hemolysis</td>
<td>F756</td>
</tr>
<tr>
<td>Carcinogenesis</td>
<td>F748</td>
</tr>
<tr>
<td>Long-term implantation</td>
<td>F748</td>
</tr>
<tr>
<td>Mucous membrane irritation</td>
<td>F748</td>
</tr>
<tr>
<td>Systemic injection acute toxicity</td>
<td>F750</td>
</tr>
<tr>
<td>Intracutaneous injection</td>
<td>F749</td>
</tr>
<tr>
<td>Sensitization</td>
<td>F720</td>
</tr>
<tr>
<td>Mutagenicity</td>
<td>F748</td>
</tr>
<tr>
<td>Pyrogenicity</td>
<td>F748</td>
</tr>
</tbody>
</table>
flexural mismatch may lead to release of wear particles that can excite a cytochemical reaction that culminates in inflammation and cell cytotoxicity. The generation of wear particles and their size is of particular importance to the failure of joint implants through a process termed osteolysis (23).

Skin irritation assay
This test involves applying a patch of the material (or an extract of the material) to an area of an animal that has been shaved; in some cases the skin is abraded before the test material is applied. After 24 hours of contact, the patch is removed, and the skin is graded for redness and swelling. The grading scale can vary from 0 to 4: 0 means no redness and/or swelling and 4 means extensive redness and/or swelling. Standard test materials are used to evaluate skin irritation (24).

Short-term intramuscular implantation
This test is designed to evaluate the reaction of tissue to a device for periods of 7 to 30 days. This test can be conducted in the muscle below the skin in rabbits or rodents including mice, rats, and guinea pigs. At the conclusion of the test period, the samples are graded both visually and based on analysis of histological sections. A test described in the United States Pharmacopoeia (USP) is widely used. The purpose of this test is to evaluate the inflammatory potential (e.g., redness and swelling) grossly. In some cases, histological evaluation of the tissue is performed at the light and electron microscopic levels to look for phagocyte and immune cells. Some investigators use an intramuscular implantation site because the blood supply and hence the inflammatory potential may be easily evaluated. In addition, the results of short-term implantation tests may not reflect material-mediated inflammatory responses that may also occur (25).

Short-term subcutaneous implantation
This test is an alternative for studying the reaction of tissue to a device for a period of days to weeks. In this test, a tissue pocket is made in the skin above the muscle layer, the device is inserted into the pocket, and the pocket is sutured or stapled closed. Normally the device is placed deep into the pocket away from the site of insertion of the device so that reactions at the surface or clip site do not affect the evaluation of biocompatibility.

Although short-term implantation studies do give an analysis of the biocompatibility of a material at a local site, systemic effects can also be observed from corrosion products that develop from vascular implants that migrate to other sites (26).

Blood coagulation
Blood coagulation is normally assessed by determination of clotting times and extent of platelet aggregation initiated by the device surface in either static or dynamic systems. In a dynamic test, blood flows through the device or over a test surface made of the materials used in the device. This test is normally conducted on blood-contacting devices to ensure that the blood-coagulation and platelet-aggregation pathways are not modified. The tests are conducted in vitro using human or animal blood, ex-vivo in a flow chamber using animal blood, or in vivo in an animal model. It has been noted that variability in the results using standard materials is noted in ex-vivo tests of blood compatibility; this finding is attributed to the type of animal model used, the flow velocity, the time of exposure, and the method used to measure blood cell adhesion (27). Studies of stents used in the cardiovascular system illustrate that clot or thrombus formation is dependent on the type and design of the device (28, 29) and may be influenced by the corrosion of metallic implants (30).

Hemolysis
Hemolysis is determined by placing powder, rods, or extracts of a material in contact with human or animal plasma for about 90 minutes at 37°C (31). The amount of hemoglobin released into solution after lysis of the red cells in contact with the device is measured. When red cells undergo lysis, hemoglobin is released from the cells, and the absorbance from released hemoglobin is proportional to the amount of cell lysis. Extensive red-cell lysis is not desirable for devices that are to be implanted in the cardiovascular system. The measurement of hemolysis and its relevance is a question that should be addressed at each device application.

Carcinogenicity
Carcinogenicity testing involves long-term implantation (up to 2 years) in an animal model usually under the skin to look for tumor formation (32). This test is required for devices...
that employ materials that have not been extensively tested. Typically these tests are conducted in rodents, although rodents do form tumors to most solid implants (1).

**Long-term implantation tests**

These tests are covered by ASTM specifications F361 and F469 for muscle and bone, respectively. Implant materials are placed in the muscle as a soft-tissue model and in bone as a hard-tissue model. The implantation site is evaluated grossly and histologically for inflammation, giant cell formation, signs of implant movement, and for tissue necrosis. Although long-term implantation gives some indication of biocompatibility, it does not consider issues such as biofilm formation, infection, and encrustation associated with use of devices such as urologic implants (33). It is recommended that long-term implantation tests be conducted on a model relevant to the intended end use. In addition, the effect of wear particles is an important consideration with long-term implantation (23).

**Mucous membrane irritation**

Mucous membrane irritation is evaluated by placing a material in close proximity to a mucous membrane such as the oral mucosa. The test evaluates the amount of irritation and inflammation from gross and histological measurements. The hamster cheek pouch or oral mucosa is a model frequently used for this test (34).

**Systemic injection**

Systemic injection (acute toxicity) is designed to determine the biological response to a single intravenous or intraperitoneal injection of an extract (50 mg/kg) of a material over a 72-hour time period (35). Extracts are prepared using saline or other solutions that simulate body fluids. Animals are monitored for signs of toxicity immediately after injection and at various time intervals (1).

**Intracutaneous injection**

Intracutaneous injection involves the reaction of an animal to a single injection of a saline or vegetable oil extract of a material. Rabbits are commonly used, and they are studied for signs of inflammation (redness and swelling) at the injection site for a period of 72 hours (36).

**Sensitization assays**

Sensitization assays involve mixing the material or extract that has been in contact with a device with Freund’s complete adjuvant and injection of the test sample into the subcutaneous tissue during a 2-week induction period (37). After 2 weeks, the material or extract is placed on the skin near the injection site for 24 hours, and then the skin is evaluated for redness and swelling.

**Mutagenicity**

Mutagenicity is evaluated using the Ames test or an equivalent test. This test employs genetically altered bacteria (bacteria with altered nutritional characteristics), which are placed in contact with an extract of a material. Mutations are observed that lead to a reversion back to the “wild-type” phenotype that grows only under the original nutritional conditions and not under conditions that allows mutant growth. It has been reported many very small-wear particles are released from metal-on-metal contact in joint replacements, and these particles may cause mutagenic damage in bone cells (23). However, many of these wear particles may also be of concern as carcinegens.

**Pyrogenicity**

Pyrogenicity is used to evaluate fever-producing substances that may contaminate a medical device. They are components of gram-negative bacterial cell membranes (endotoxins) or are materials of chemical origin. The presence of endotoxins is determined by injecting an extract of the device into the circulatory system of a rabbit and measuring temperature in the rabbit’s ear. Another method to measure endotoxins involves contact of the material with cells that are lysed specifically by endotoxins (Limulus Amebocyte Lysate Test). Chemical pyrogens are determined by the rabbit test (38).

**Animal Models—A Variety of Animal Models Are Used**

A variety of animal models is used to evaluate the biocompatibility of medical devices. They include dog, sheep, goat, rabbit, mouse, rat, ferret, and pig. Pearse et al. (39) and Murray et al. (22) give a recent review of animal models in bone and dental device testing, respectively. Factors that lead to choice of a particular animal model include housing requirements as well as cost, maintenance, and care factors, resistance to disease, interanimal uniformity, tolerance to surgery, animal lifespan, the number and size of the implant (39–41). International standards (ISO 10993) provide guidance to determine the number of animals and the species that should be tested for each treatment and time point. Although the rat is one of the most widely used species in medical research because of its size and tissue structure, it is not a good model for testing some medical devices. Although the dog may not be a good model for bone implants because of differences in size and shape of canine bone in comparison with human bone, it is sometimes used because commercially available implants and instruments are available for canine surgery (39). Tissue microstructure, composition, wound healing, and remodeling differences with the human play an important role in animal model selection.

**Summary**

The word biocompatible is perhaps a misnomer; it refers to the ability of materials used in medical devices not to the illicit
adverse reactions that occur when implanted in humans. To assess biocompatibility, several tests have been developed to study the interactions of materials and biological tissues. The biocompatibility of a material or a device requires that it not activate any biological homeostatic systems including blood coagulation, platelet aggregation, inflammation, complement, or fibrinolysis.

Biocompatibility testing of medical devices depends on many criteria. Devices that contact only the skin and are for short-term use require less-stringent testing. These devices do not require long-term testing prior to human use. In contrast, implantable devices require more safety and effectiveness testing before they can be used in humans. The type of testing required depends on the end-use application. Although the term biocompatible is relative, because all devices will lead to some reactions by cells and tissues, it is important that the reactions observed at the device-tissue interface do not lead to implant failure or interfere with the functioning of any biological systems.

Biocompatibility tests vary for short-term versus long-term contact with tissue. Depending on whether the contact is with skin or internally, many standard tests have been developed for biocompatibility screening. Although many tests are predictive of the responses observed during use of devices in humans, biocompatibility of a device is ultimately only verified after extensive human clinical trials and general use in the population.

References

31. Wang W, Wiedeman F Jr, Brea TK, Eif GF. Pulmonary and hemotological inflammatory response to intravenous cellulose mi-


Further Reading


Controlled Drug Delivery: Pharmacokinetic Considerations, Methods and Systems

Zhong Zuo and Vincent H. L. Lee, School of Pharmacy, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR
doi: 10.1002/9780470048672.wecb099

Controlled drug delivery applies interdisciplinary approaches to engineer systems that improve the therapeutic value of drugs. This review addresses the biological basis of drug delivery, the pharmacokinetic/pharmacodynamic considerations important to the design of controlled release delivery systems, and the methods to fabricate them. The focus of this review will principally be on oral and transdermal applications. Systems at various stages of development for the delivery of more complex molecules, such as proteins, oligonucleotides, genes, and siRNAs will also be discussed.

The frequency of drug dosing is usually determined by the drug’s duration in the body. For drugs that are inherently long lasting, once daily oral dosing is sufficient to sustain adequate drug blood levels and the desired therapeutic effect. Formulation of these drugs as conventional, immediate-release dosage forms is used for the patient. However, many drugs are not inherently long lasting and require multiple doses each day to achieve the desired therapeutic results. Multiple daily dosing is inconvenient for the patient and can result in missed doses, made-up doses, and patient noncompliance with the therapeutic regimen. In addition, sequential therapeutic blood level peaks and valleys (troughs) are associated with each dose. The traditional approach to modifying drug release is to control the rate of drug delivery, as exemplified by extended-release tablets and capsules, which are commonly taken only once or twice daily. Modern drug delivery systems provide an additional targeted drug release dimension at a specific site in the body, spanning subcellular to organ. Thus, the term “controlled drug delivery” has a much broader meaning than just controlling the rate of drug delivery. Controlled drug delivery to the right site of action at the right time can be achieved by either 1) controlling the release rate and duration or 2) controlling the release site (i.e., localized delivery or targeted diseased organs, tissues or cells). Figure (1) has summarized the scheme for the design of the controlled drug delivery. The commonly used types of controlled drug delivery products in terms of its targeted diseases and specific delivery technology involved are summarized in Table (1).

General Introduction of Pharmacokinetics/Pharmacodynamics

Fate of a drug In vivo: ADME process

After administration, the fate of the drug will be determined by four key steps in vivo, namely, absorption (A), distribution (D), metabolism (M), and excretion (E). Effects of the four steps on the administered drug will influence the levels of drug exposure to the site of action and will eventually influence the pharmacological activities of the drug. The details of the four steps are described briefly as follows.

Absorption (A)

Before a compound can exert its pharmacological effect in tissues, it has to be taken into the bloodstream usually via mucosal surfaces such as the digestive tract for intestinal absorption. In addition, uptake of drugs from the bloodstream into targeted organs or cells needs to be ensured. However, this task is not always easy given the natural barriers that exist, like the blood-brain barrier. Extent of absorption for a drug varies with the way it has been administered. Factors such as poor compound solubility, chemical instability in the stomach, and inability to permeate the intestinal wall can all reduce the extent to which a drug is absorbed after oral administration. Absorption critically determines the compound’s bioavailability. Drugs that absorb poorly when taken orally must be administered in some less desirable way, like intravenously or by inhalation.
Distribution (D)

Once the drug gets into the bloodstream, it will be carried to various parts in the body including different tissues and organs, as well as its site of action. Depending on the physicochemical and biological properties of the drug, its degree of distribution to various tissues and organs differs by varying extents.

Metabolism (M)

As soon as the drug enters the body, in addition to being distributed to other parts of the body, it begins to break down. Most small-molecule drug metabolism occurs in the liver by enzymes termed cytochrome P450 enzymes. As metabolism occurs, the parent compound is converted to metabolites, which could be pharmacologically inert or active.

Excretion (E)

Compounds and their metabolites need to be removed from the body via excretion, usually through the kidneys or the feces. Unless excretion is complete, accumulation of foreign substances in the body could affect normal metabolism adversely.

Role of transporters in ADME

The extent of absorption for a drug varies with the way it has been administered. Over the last decade, several important transporters responsible for drug uptake/disposition in various organs have been discovered. They mainly include organic anion transporter, organic cation transporter, multidrug resistant protein, and multidrug resistance-associated protein. Many transporters function in the uptake of drugs into cells, which leads to increased permeability of the drug into cells; other transporters function to export drugs out of cells, which thereby decreases the apparent permeability of the drug. Transport mechanism of drugs and substrate specificities of drug transporters in drug development have become increasingly important. Identification of drugs that are the substrates of the transporters can help to elucidate the pharmacokinetic profiles of these drugs and drug-drug interactions, as well as to improve the therapeutic safety of drugs. Therefore, research that focuses on such membrane transporters is promising, and it leads to rational design strategies for drug targeting delivery. Drug targeting to specific transporters and receptors using carrier-mediated absorption has demonstrated immense significances in the ocular drug delivery (3) and targeted drug delivery across the blood-brain barrier (2).

Concept of pharmacodynamics, biomarkers and pharmacogenetics

Pharmacodynamics is the study of the mechanisms of drug action and the biological and physiological effects of these drugs. Compared with pharmacokinetics, which reveals what the body does to a drug, pharmacodynamics focuses more on what a drug does to the body.

Traditionally, the level of drug in vivo is assumed to be proportional to its pharmacodynamic effect. The identification of various biomarkers provides a more relevant indicator than the in vivo drug concentrations as far as clinical efficacy of a drug is concerned. Biomarkers enable the characterization of patient populations and the quantization of the extent to which new drugs reach intended targets, alter proposed pathophysiologic mechanisms, and achieve clinical outcomes. In genomics, the biomarker challenge is to identify unique molecular signatures within complex biological mixtures that can be unambiguously correlated to biological events to validate novel drug targets and to predict drug response. Biomarkers can stratify patient populations or quantify drug benefit in primary prevention or disease-modification studies in poorly served areas such as neurodegeneration and cancer. A proprieties of biomarkers depends on the strategy and the stage of development, as well as the nature of the medical indication. Biomarkers are perhaps most useful in the early phase of clinical development when measurement of clinical endpoints may be too time-consuming or cumbersome to provide timely proof of concept or dose-ranging information (3).

With the disclosure of the map of the human genome, the effect of human genetic variations on drug metabolism and its related clinical efficacy has become a major concern for clinical practice. Pharmacogenetics is an area of study on the genetic variation in response to drug metabolism with an emphasis on improving drug safety. The aim of drug delivery is to achieve an appropriate drug concentration at the targeted site that can elicit a desired level of response. The concentration of a drug in vivo is highly dependent on the ADME pharmacokinetic processes, whereas the level of response for a drug usually results from the interaction of a drug with a target receptor protein. More and more studies have revealed genetic polymorphisms in drug-metabolizing enzymes, transporters, and target receptors. The genetic basis underlying pharmacokinetic and pharmacodynamic inter-individual variability is essential in the consideration of the design of drug delivery systems (4).
<table>
<thead>
<tr>
<th>Disease state</th>
<th>Active ingredient</th>
<th>Type of controlled delivery system</th>
<th>Product (manufacturer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endocrinology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Postmenstrual syndrome</td>
<td>Estradiol</td>
<td>TDD</td>
<td>Estraderm (Novartis), Vivelle (Novartis), Climara (Berlex)</td>
</tr>
<tr>
<td>Hypogonadism in males</td>
<td>Testosterone</td>
<td>TDD</td>
<td>Testoderm (Alza), Atestin (SmithKline Beecham)</td>
</tr>
<tr>
<td>Birth control</td>
<td>Estradiol/Progesterone</td>
<td>IUD insert</td>
<td>Progesterset (Alza)</td>
</tr>
<tr>
<td>Hormone replacement therapy</td>
<td>Progestosterone</td>
<td>IUD insert</td>
<td>Nuvella TS (Ethical Holdings/Schering)</td>
</tr>
<tr>
<td>Pain, inflammatory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate/severe pain</td>
<td>Fentanyl</td>
<td>TDD</td>
<td>Duragesic (Janssen)</td>
</tr>
<tr>
<td>Central nervous system</td>
<td>Nicotine</td>
<td>TDD</td>
<td>Habitrol (Novartis Consumer), Nicoder CQ (SmithKline Beecham Consumer)</td>
</tr>
<tr>
<td>Smoking cessation</td>
<td>Scopolamine</td>
<td>TDD</td>
<td>Transderm Scop (Novartis Consumer)</td>
</tr>
<tr>
<td>Motion sickness</td>
<td>Methylphenidate HCL</td>
<td>Osmotic pump</td>
<td>Concerta</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calcium channel, antagonist for hypertension, ischemic heart disease, hypertrophic cardiomyopathy, arrhythmias</td>
<td>Nifedipine</td>
<td>Diffusion, osmotic pump, erosion</td>
<td>Procace, Procardia XL, Adalat CC</td>
</tr>
<tr>
<td>Diltiazem</td>
<td></td>
<td>Diffusion, erosion</td>
<td>Cardizam, Cardizam CD, Cardizam SR, Discor X R, Vanilin, Calan SR, Isotab SR</td>
</tr>
<tr>
<td>Verapamil</td>
<td></td>
<td>Diffusion, erosion</td>
<td>Cardizem SR</td>
</tr>
<tr>
<td>Felodipine</td>
<td></td>
<td>Erosion</td>
<td>Fendil, Cardene SR</td>
</tr>
<tr>
<td>Nicardipine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitroglycerin</td>
<td></td>
<td>TDD (Monolithic and membrane controlled)</td>
<td>Nitrodisc Bead, Nitro-Dur (Key Pharmaceuticals), Transdermal-Net (Summit Medical), Depojet (Schwarz Pharm), Minitor (3M Pharmaceuticals), etc.</td>
</tr>
<tr>
<td>Calcium channel, antagonist for hypertension, ischemic heart disease, congestive heart failure</td>
<td>Isosorbide dinitrate</td>
<td>Matrix tablet</td>
<td>Imdur (Schering Key)</td>
</tr>
<tr>
<td>Hydralazine</td>
<td>Clonidine</td>
<td>TDD</td>
<td>Catapress-TTS</td>
</tr>
<tr>
<td>Aminophylline</td>
<td>Propanolide</td>
<td>Matrix tablet</td>
<td>Procain SR (Parke Davis)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>Disopyramide phosphate</td>
<td>Multipid-let</td>
<td>Norpace CR</td>
</tr>
<tr>
<td>Metabolic acidosis</td>
<td>Nicotinic acid</td>
<td>Polygel matrix delivery system</td>
<td>Sil-Niacin (Upsher-Smith)</td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recurrent ovarian cancer</td>
<td>Doxorubicin</td>
<td>Stealth liposome</td>
<td>Doxil, Cadex</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>Fungal infections AIDS</td>
<td>Liposome</td>
<td>Ambisome</td>
</tr>
<tr>
<td>Ophthalmic disease</td>
<td>Amphotericin B</td>
<td>Reservoir insert</td>
<td>Ocuent PILO (Alza)</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>Pilocarpine</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Controlled Drug Delivery: Pharmacokinetic Considerations, Methods and Systems

Pharmacokinetic/Pharmaceutical Dosage Form/Pharmacodynamic Considerations in the Design of A Controlled Drug Delivery System

An ideal controlled-release product is expected to release the drug from the dosage form at a predetermined rate, to maintain the released drug at a sufficient concentration in the gastrointestinal fluids, and to provide sufficient gastrointestinal residence time prior to the absorption of the drug. In general, not every drug is suitable to be developed as a controlled-release product.

Traditionally, the pharmacokinetic properties of drug candidates has always played an important role in their design as a controlled release dosage form. From the pharmacokinetic point of view, the best suited drug candidates should have neither very slow nor very fast rates of absorption and excretion. Drugs with slow rates of absorption and excretion with half-lives of greater than 8 hours are usually inherently long acting, and their preparation into extended release dosage forms is not necessary. Drugs with very short half-lives (i.e., < 2 hours) are poor candidates for extended-release dosage forms because of the large quantities of drug required for such a formulation. Another pharmacokinetic consideration is that the drugs prepared in extended-release forms must have good aqueous solubility, must maintain adequate residence time in the gastrointestinal tract, and must be uniformly absorbed from the gastrointestinal tract. In addition, from the classic dosage form design point of view, the drug candidate should be administered in relatively small doses because each oral dosage unit needs to maintain a sustained therapeutic blood level of the drug.

As indicated in Table 1, because of the nature of the duration of action for a controlled drug delivery system, another important pharmacodynamic consideration for drug candidates is that long-acting drugs are used in the treatment of chronic rather than acute conditions. Regardless of how it is achieved, the prolonged and controlled pharmacodynamic effect is the eventual outcome expected for the controlled drug delivery system. The appropriate drug candidate should possess a reasonable therapeutic index (i.e., the median toxic dose divided by the median effective dose). It should be noted that drugs possessing a longer action than indicated by their quantitative pharmacokinetic half-lives might not be good candidates for a controlled drug delivery system either. Therefore, monitoring the controlled release drug more reliably and accurately in vivo is becoming a challenge and the traditional pharmacokinetic/pharmacodynamic correlations as a standard component of drug development are becoming more complicated than what is usually expected (5). An integrated pharmacokinetic and pharmacodynamic approach for rate-controlled drug delivery has been proposed (6). Should pharmacodynamic measurements eventually replace the traditional pharmacokinetic measurements? With the availability of various biomarkers, the measurements of related biomarkers rather than the plasma concentrations of specific drugs are believed to provide more direct and accurate monitoring of the behavior of the controlled drug delivery system in vivo.

Methods and Systems of Controlled Drug Delivery:

Natural and synthetic polymers, both erodible and nonerodible, play an important role in the fabrication of systems to control drug release. The method of combining polymers with drug candidates, together with the manufacturing process, provides even more control on the release profile desirable of the drug candidate.

Systems to control the release rate and duration of the drug

The rate of drug release from solid dosage forms may be modified by the technologies described below, which in general are based on 1) modifying drug dissolution by controlling access of biologic fluids to the drug through the use of barrier coatings, 2) controlling drug diffusion rates from dosage forms, and 3) chemical reactions or interactions between the drug substance or its pharmaceutical barrier and site-specific biological fluids. The commonly used systems and the related techniques are illustrated as follows.

Coated beads, granules, or microspheres

This system aims to distribute the drug onto beads, pellets, granules, or other particulate systems. The rate at which body fluids can penetrate the coating to dissolve the drug will be adjusted by varying the thickness of the coats and the type of coating material used. The coating techniques include conventional pan-coating or air-suspension coating techniques by which a solution of the drug substance is placed onto small inert nanoparticle seeds or beads made of sugar and starch or onto microcrystalline cellulose spheres. Naturally, the thicker the coat, the more resistant to penetration and the more delayed will be the drug release and dissolution. Typically, the coated beads are about 1 mm in diameter. However, the beads are combined to have three or four release groups among the more than 100 beads contained in the dosing unit. This technique provides the different desired sustained or extended release rates as well as the targeting of the coated beads to the desired segments of the gastrointestinal tract. An example of this type of dosage form is the Spansule capsule from SmithKline Beecham (King of Prussia, PA). Microencapsulated drug

Microencapsulation is a process by which solids, liquids, or even gases may be encapsulated into microscopic sized particles through the formation of thin coatings of "wall" material around the substance being encapsulated. Gelatin is a common wall-forming material, but synthetic polymers, such as polyvinyl alcohol, ethylcellulose, polyvinyl chloride, and other materials, may be used. The typical encapsulation process usually begins with the dissolving of the prospective wall material, such as gelatin, in water. The material to be encapsulated is added and the two-phase mixture is stirred thoroughly. The material to be encapsulated is then broken up into the desired particle size via addition of a solution of a second material, which is usually acacia. This additive material can concentrate the gelatin...
Embedding drug in hydrophilic or biodegradable matrix systems

By this process, the drug substance is combined and made into granules with an excipient material that slowly erodes in body fluids, progressively releasing the drug for absorption. Drugs without excipients provide an immediate drug effect, whereas drug-excipient granules provide extended drug action. The granule mix may be tableted or placed into gelatin capsule shells for oral delivery. Hydroxypropyl methylcellulose (HPMC), which is a free-flowing powder, is commonly used to provide the hydrophilic matrix. Tablets are prepared by distributing HPMC in the formulation thoroughly, preparing the granules by wet granulation or roller compaction, and manufacturing the tablets by compression. After ingestion, the tablet is made wet by gastric fluid, which hydrates the polymer. A gel layer forms around the surface of the tablet and an initial quantity of drug is exposed and released. As water permeates further into the tablet, the thickness of the gel layer is increased, and the soluble drug diffuses through the gel layer. As the outer layer becomes fully hydrated, it erodes from the tablet core. If the drug is insoluble, then it is released as such with the eroding gel layer. Thus, the rate of drug release is controlled by the processes of diffusion and tablet erosion. An example of a proprietary product using a hydrophilic matrix base of HPMC for extended drug release is Omnacortil SR Tablets from Roche Laboratories, Inc. (Columbus, OH), which contains morphine sulfate.

Scientists have been working for decades to expand the number of biodegradable polymers for clinical use, which include polyhydroxyacids, polyglycerides, polyethylene, and polyesters. Of these polymers, poly(lactic acid) (PLA) and poly(D,L-lactide-co-glycolide) (PLGA) are most often used because of their relatively good biocompatibility, easily controlled biodegradability, and good processability. In fact, they belong to one of the two classes of FDA approved synthetic biodegradable polymers to date. These polymers can be used either as matrix devices or as reservoirs. In matrix systems, the drug is dispersed or dissolved in the polymer, whereas the drug is encapsulated in a biodegradable membrane in the reservoir. The release rate of the drug is generally dependent on the degradation of the polymer. Poly(ortho)esters and polyamides erode from their surface, whereas the PLGA and PLA follow bulk erosion (i.e., degradation occurring throughout the whole polymer). By varying the ratios of monomers in the copolymer mix, nearly zero-order drug release produced by matrix degradation at steady rate can be achieved (7).

A few commercial available drug delivery systems are based on biodegradable polymers. Glididdel is a white, dime-sized wafer made up of biodegradable polyanhydride polymer [poly(1,3-bis(carboxyphenoxy)propane-co-sebacic acid)] (PPCP-SA) to deliver carmustine for the treatment of brain cancer. ReGel is an aqueous filter sterilizable ABA tri-block polymer system that consists of PLGA and polyethylene glycol (B). Its suitability to provide sustained interleukin-2 (IL-2) delivery for cancer immunotherapy has been reported. A Trigel is a proprietary delivery system that contains PLA or PLGA dissolved in a biocompatible carrier and can be used for both parenteral and site-specific drug delivery. The system is placed in the body using conventional needles and syringes and solidifies on contact with aqueous body fluids to form a solid implant. For a drug incorporated into such polymer solution, it becomes entrapped within the polymer matrix as it solidifies followed by being slowly released as the polymer biodegrades. Products that have already been approved by the FDA using the A Trigel technology include the Eligard (leuprolide acetate for injectable suspension) prostate cancer products that provide systemic release of leuprolide acetate for 1-, 3-, and 4-month duration and the Atrix (8.5% doxycycline) periodontal treatment product for localized subgingival delivery of doxycycline. With the clinical success of the current available products using biodegradable polymers, their additional applications in the delivery of small molecules, peptides, proteins, monoclonal antibodies, and vaccines are anticipated.

Embedding drug in an inert plastic matrix

By this method, the drug is granulated with an inert plastic material, such as polyethylene, polyvinyl acetate, or poly(methylmethacrylate), and the granulation is compressed into tablets. The compression of the tablet creates the matrix or plastic form that retains its shape during drug diffusion and its passage through the alimentary tract. An immediate-release portion of the drug may be compressed onto the surface of the tablet. The inert tablet matrix, which is expended of drug, is excreted with the feces. The primary example of a dosage form of this type is the Gradusmet from Abbott Pharmaceuticals (Abbott Park, IL).

Ion-exchange resins

A solution of a cationic drug may be passed through a column that contains an ion-exchange resin, which forms a complex by the replacement of hydrogen atoms. The resin-drug complex is then washed and may be tableted, encapsulated, or suspended in an aqueous vehicle. The release of the drug is dependent on the pH and the electrolyte concentration in the gastrointestinal tract. Release is greater in the acidity of the stomach than in the less acidic environment of the small intestine. Examples of drug products of this type include hydrocortisone polistirex and chlorpheniramine polistirex suspension (Tussionone Perorinetic Extended Release Suspension; M edeval) and phentermine resin capsules (Ionamin Capsules; Pharmaxen, Provo, U.T.).
Osmotic controlled release systems

The pioneer oral osmotic pump drug delivery system is the Oros system, which was developed by Alza Corporation (Mountain View, CA). The system is composed of a core tablet surrounded by a semi-permeable membrane coating with a 0.4-mm diameter hole in it that was produced by a laser beam. The core tablet has two layers: one contains the drug and the other contains a polymeric osmotic agent. The system operates on the principle of osmotic pressure. When the tablet is swallowed, the semi-permeable membrane permits water to enter from the patient’s stomach into the core tablet, which dissolves or suspends the drug. As pressure increases in the osmotic layer, it forces or pumps the drug solution out of the delivery orifice on the side of the tablet. Only the drug solution can pass through the hole in the tablet. The rate of inflow of water and the function of the tablet depends on the existence of an osmotic gradient between the contents of the bilayer core and the fluid in the gastrointestinal (GI) tract. Drug delivery is essentially constant as long as the osmotic gradient remains constant. The drug release rate may be altered by changing the surface area, the thickness, or the composition of the membrane, and/or by changing the diameter of the drug release orifice; the rate is not affected by gastrointestinal acidity, alkalinity, fed conditions, or GI motility. The biologically inert components of the tablet remain intact during GI transit and are eliminated in the feces as an insoluble shell. This type of osmotic system, which is termed the “gastrointestinal therapeutic system” (GITS; Pfizer, New York), is employed in the manufacture of Glucotrol XL Extended Release Tablets and Procardia XL Extended Release Tablets (Pfizer).

Gastro-retentive systems

Gastric emptying of dosage forms is an extremely variable process, and the ability to prolong and control the emptying time is a valuable asset for dosage forms, which reside in the stomach for a longer period of time than conventional dosage forms. Gastro-retentive systems can remain in the gastric region for several hours, and hence, they significantly prolong the gastric residence time of drugs. Prolonged gastric retention would therefore improve the bioavailability of drugs by reducing drug waste and by improving solubility for drugs that are less soluble in a high pH environment. It has applications also for local drug delivery to the stomach and proximal small intestines. The controlled gastric retention of solid dosage forms may be achieved by the mechanisms of mucadhesion, flotation, sedimentation, expansion, modified shape systems, or by the simultaneous administration of pharmacological agents that delay gastric emptying. Based on these approaches, classification of floating drug delivery systems has been investigated extensively. Several recent examples have been reported that show the efficiency of such systems for drugs with bioavailability problems. Gastrectensive agents help to provide better availability of novel products with many therapeutic possibilities and substantial benefits for patients. However, because of the complexity of the pharmacokinetic and pharmacodynamic factors for certain drugs, some researchers suggest that the rationale for continuous administration obtained by controlled-release gastroretentive dosage forms should be assessed and established in vivo (3).

Transdermally controlled drug delivery systems

Because of the large surface area of the skin and its bypass of the liver as a first pass step in metabolism, many drug delivery systems have been developed that control the rate of drug delivery to the skin for subsequent absorption. Effective transdermal drug delivery systems of this type deliver uniform quantities of drug to the skin over a period of time. Technically, transdermal drug delivery systems may be classified into monolithic and membrane-controlled systems (9).

Monolithic Systems

The systems incorporate a drug matrix layer between backing and frontal layers. The drug-matrix layer is composed of a polymeric material in which the drug is dispersed. The polymer matrix controls the rate at which the drug is released for percutaneous absorption. In the preparation of monolithic systems, the drug and the polymer are dissolved or blended together, cast as the matrix, and dried. The dried matrix may be produced in a sheet or cylindrical form, with individual dosage units cut and assembled between the backing and the frontal layers. Most transdermal drug delivery systems are designed to contain an excess of drug and thus drug releasing capacity beyond the time frame recommended for replacement.

Membrane-Controlled Transdermal Systems

The system is designed to contain a drug reservoir, usually in liquid or gel form, a rate-controlling membrane, and backing, adhesive, and protecting layers. This system has an advantage over monolithic systems in that as long as the drug solution in the reservoir remains saturated, the release rate of drug through the controlling membrane remains constant. Membrane-controlled systems are generally prepared by preconstructing the delivery unit, filling the drug reservoir, and sealing it off, or by a process of laminating, which involves a continuous process of construction, dosing, and sealing.

The first transdermal therapeutic system designed to control the delivery of a drug to the skin for absorption was developed in 1980 by the Alza Corporation. The system, marketed as Transderm-Scop by CIBA, is a circular, flat adhesive patch designed for the continuous release of scopolamine through a rate-controlling microporous membrane. Soon after, nitroglycerin, which is a drug substance used for the treatment of angina, became another candidate for transdermal delivery. The drug has a relatively low dose, short plasma half-life, high peak plasma levels, liver first pass metabolism, and inherent side effects when taken sublingually (a popular route for its administration). Since then, several nitroglycerin-containing systems have been developed including: Deponit (Wyeth-Ayerst), Minitran (3M Riker, St. Paul, MN), TransdermNitro (CIBA-Geigy, Summit, NJ), Nitro-Dur (Key Pharmaceuticals, Kenilworth, NJ), Nitroine (Schwarz Pharma, Monheim, Germany), and Nitrodisc (Searle). As indicated in Table 1, the application of the transdermal therapeutic system products covers various aspects with the most emphasis on its use in central nervous system disease, cardiovascular disease, and hormonal therapy.

Future developments in the field of transdermal drug delivery should address problems that relate to irritancy, sensitization,
It consists of unilamellar bilayer liposomes that are less than 100 nm in diameter with amphoterocin B intercalated within the membrane. Injectable liposome-based drug delivery systems of anti-cancer drugs are the most widely used drug nanoparticle in cancer (15). In addition to a PEG coating, most stealth liposomes also have some sort of biological species attached as a ligand to the liposome to enable binding via a specific receptor for the targeted drug delivery site. These targeting ligands could be monoclonal antibodies (named as immunoliposome), vitamins, or specific antigens. Targeted liposomes can home to nearly any cell type in the body and deliver drugs that would otherwise be systemically delivered. It is expected that the side effects of extremely toxic drugs could be drastically reduced if the drugs were only delivered to diseased tissues using such techniques (16).

Nanoparticles

A nanoparticle is a microscopic particle with a diameter less than 100 nm. Nanoparticles were first developed around 1970, and initially they were devised as carriers for vaccines and anticancer drugs. Nanoparticle research is currently an area of intense scientific research because of a wide variety of potential applications in biomedical, optical, and electronic fields. To enhance tumor uptake, the strategy of drug targeting was employed, and as a first important step, research focused on the development of methods to reduce the uptake of the nanoparticles by the RES cells. Simultaneously, the use of nanoparticles for ophthalmic and oral delivery was investigated (17, 18). Recent advancement of nanoparticles and nanosuspensions was caused by their application for pulmonary drug delivery (19, 20).

The size of nanoparticles generally varies from 10 to 1000 nm. The drug is usually dissolved, entrapped, encapsulated, or attached to a nanoparticle matrix, and depending on the method of preparation, nanoparticles, nanospheres, or nanocapsules can be obtained. Nanocapsules are vesicular systems in which the drug is physically and uniformly dispersed. In recent years, biodegradable polymeric nanoparticles using poly(D, L-lactide), PLA, poly(D, L-glycolide) (PLG), PLGA, and poly(caproacrylate) (PCA) have attracted considerable attention as potential drug delivery devices in view of their applications in the controlled release of drugs, their ability to target particular organs/tissues, as carriers of DNA in gene therapy, and in their ability to deliver proteins, peptides, and genes through a peroral route of administration. The PLA, PLG, and PLGA polymers have been used as controlled release formulations in parental and implantation drug delivery applications.

Nanomedicine is undergoing explosive development as it relates to the development of nanoparticles for enabling and improving the targeted delivery of therapeutic and diagnostic agents. The use of nanoparticles for specific delivery to malignant tumors, to the central nervous system, and across the gastrointestinal barriers has received great attention (21). The additional development of nanoparticles is highly dependent on the advancement of the material sciences. For nanotechnol-ogy rationally to generate materials useful in human therapy,
it will need to progress in full recognition of all the require-
ments biology places on the acceptability of exogenous ma-
terials (22). In particular, nanostructured porous materials, in
particular silicon-based photonic and templated materials, offer
a degree of control in both the rate and the location of drug
delivery that is just beginning to be recognized (23).

Dendrimers

Dendrimers are macromolecules characterized by their highly
branched structure and globular shape. They are attractive drug
carriers by virtue of the flexibility in precisely controlling their
molecular size, shape, branching, length, and surface func-
tionality. An active area of research is the development of
nontoxic, biodegradable, and biocompatible dendrimers. Var-
ious kinds of dendrimer-drug conjugates have been tried us-
ing polyamidoamine, polyphosphoester, and polyester dendrimers
and anticancer drugs such as doxorubicin, methotrexate, and
5-fluorouracil. The hydrophobic core and hydrophilic shell of
dendrimers could encapsulate hydrophobic drugs or diagnostic
agents in their interior and release them slowly. With an interior
of basic environment, dendrimers could also encapsulate acidic
drugs, such as methotrextate.

A unique characteristic of dendrimers is that they can act as
a particulate system while retaining the properties of a
polymer. Dendrimers can significantly improve pharmacoki-
etic and pharmacodynamic properties of low molecular weight
and protein-based therapeutic agents. The surface modified den-
drimers with various kinds of ligands, such as sugars, folate
residues, and poly (ethylene-glycol), could serve as a drug car-
rrier with targeted cell specificity. Poly (ethylene glycol) modi-
fied (PEGylation) dendrimers can generally overcome clearance
by RES system. Attachments of targeting moiety on the sur-
face of partially PEGylated dendrimer raised the possibility of
a delivery system that can cross biological barriers and deliver
the bioactive agent near the vicinity of its target site. Recent
successes also demonstrate potential of PEGylated dendrimers
in magnetic resonance imaging contrast agent and in carbonyl
metallic immunoassay (24, 25).

Erythrocytes/Cells

Erythrocytes, which are also known as red blood cells, have
been studied extensively for their potential carrier capabili-
ties for the delivery of drugs and drug-loaded microspheres.
Such drug-loaded carrier erythrocytes are prepared simply by
collecting blood samples from the organism of interest, sepa-
rating the erythrocytes from the plasma, entrapping the drug
into the erythrocytes, and resealing the resultant cellular carri-
ers. Hence, these carriers are called resealed erythrocytes. The
overall process is based on the response of these cells to os-
motic conditions. On reinfusion, the drug-loaded erythrocytes
serve as slowly circulating depots and target the drugs to the
RES system. The advantage of erythrocytes as drug carriers is
mainly caused by their biocompatibility within the host, par-
icularly when autologous cells are used, and hence triggering an
immune response is not possible.

Several methods, that include physical (e.g., electrical-pulse
method), osmosis-based systems, and chemical methods (e.g.,
chemical perturbation of the erythrocyte membrane), can be
used to load drugs or other bioactive compounds into erythro-
cytes.

Resealed erythrocytes have several possible applications in
various fields of human and veterinary medicine. Such cells
could be used as circulating carriers to disseminate a drug over
a prolonged period of time in circulation or in target-specific
organs, including the liver, spleen, and lymph nodes. Most drug
delivery studies that use drug-loaded erythrocytes are in the
preclinical phase. In a few clinical studies, successful results
have been obtained.

The ability of resealed erythrocytes to accumulate selectively
within RES organs makes them a useful tool for the treatment of
hepatic tumors and for the treatment of parasitic diseases.
Moreover, the first report of successful clinical trials of the
resealed erythrocytes loaded with enzymes for replacement ther-
apy was that of (beta)-glucosidase in the treatment of Gaucher’s
disease, in which the disease is characterized by an
inborn deficiency of lysosomal (beta)-glucosidase in RES cells. This
deficiency thereby leads to an accumulation of
(beta)-glucosidase in macrophages of the RES. Recently,
the same concept was also extended to the delivery of biophar-
macueticals that include therapeutically significant peptides and
proteins, nucleic acid-based biologicals, antigens, and vaccines
(26). In addition, the possibility of using carrier erythrocytes for
selective drug targeting to differentiated macrophages increases
the opportunities to treat intracellular pathogens and to develop
new drugs. Moreover, the availability of an application that
permits the encapsulation of drugs into autologous erythrocytes
has made this technology available in many clinical settings and
competitive with other drug delivery systems (27).

Summary

During the past four decades, formulations that control the
rate and the period of drug delivery (i.e., time-release medi-
cations) and that target specific areas of the body for treatment
have become increasingly common and complex. Because of re-
searchers’ ever-evolving understanding of the human body and
because of the explosion of new and potential treatments that re-
sult from discoveries of bioactive molecules and gene therapies,
pharmaceutical research hangs on the precipice of yet another
great advancement. However, this next leap poses questions and
challenges not only for the development of new treatments but
also for the mechanisms by which to administer them.

The current methods for drug delivery exhibit specific prob-
lems that scientists are attempting to address. Therefore, the
goal of all sophisticated drug delivery systems is to deploy
medications intact to specifically targeted parts of the body
through a medium that can control the therapy’s administra-
tion by means of either a physical or a chemical trigger.

To achieve this goal, researchers are turning to advances in the
worlds of microtechnology and nanotechnology. During the past
decade, systems that contain polymeric materials, such as poly-
meric microspheres and nanoparticles, have all been shown to be
effective in enhancing drug targeting specificity, lowering sys-
temic drug toxicity, improving treatment absorption rates, and
providing protection for pharmaceuticals against biochemical degradation. Peptides, proteins, and nucleotides or DNA fragments are the new generation of drugs; They are becoming attractive because of fast development within the biotechnology field. The administration of such molecules, however, may pose problems such as sensitivity of the molecules to certain temperatures, instability of the molecules at some physiological pH values, a short plasma half-life, and a high molecular dimension, which hinders diffusional transport and makes, at the moment, the parenteral route the only possible way of administering such molecules. Controlled drug delivery that uses the development of new administration routes could be the answer to these problems in administration of these molecules. The rationale of drug delivery is to change the pharmacokinetic and pharmacodynamic properties of drugs by controlling their absorption and distribution. Rate and time of drug release at absorption site could be programmed using a so-called delivery system. Various techniques, that include chemical technologies (such as pro-drugs), biological technologies, polymers, and lipids (such as liposomes) as well as the invention of new routes of administration, have been proposed to achieve controlled drug release. In fact, it could increase drug absorption and reduce the effects of the active ingredient in those regions not requiring therapy. Drug delivery systems that allow for an effective in vivo release of new molecules, such as recombinant anti-idiotypic antibodies with antibiotic activity devoted to the treatment of pulmonary tuberculosis and pneumocystosis and mucosal (candidiasis) diseases, fall into that category.

References

See Also

Drug Delivery
Controlled Drug Delivery: Pharmacokinetic Considerations, Methods and Systems
Advanced Article

Directed Evolution, Novel and Improved Enzymes through

Fei Wen*, Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois

Michael McLachlan*, Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois

Huimin Zhao, Departments of Chemical and Biomolecular Engineering and Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois
doi: 10.1002/9780470048672.wecb125

By mimicking Darwinian evolution in the test tube, directed evolution has become a powerful tool for engineering novel enzymes for basic and applied biology research and medicine. Unlike structure-based rational design, directed evolution is capable of altering single or multiple functional properties such as activity, specificity, selectivity, stability, and solubility of naturally occurring enzymes in the absence of detailed knowledge of enzyme structure, function, or mechanism. More recently, directed evolution has also been used to engineer metabolic pathways, viruses, and whole microorganisms, and to address fundamental problems in biology. The success of directed evolution has been largely fueled by the development of numerous molecular biology techniques that enable the creation of genetic diversity through random mutagenesis or homologous or nonhomologous recombination in the target genes and the development of powerful high throughput screening or selection methods as well as by novel applications. This review will highlight the key developments in directed evolution and focus on the design and engineering of novel enzymes through directed evolution and their implications in chemical biology.

Enzymes are truly remarkable catalysts that are essential to every biological process. They can catalyze a broad range of chemical transformations with exquisite selectivity (stereo-, regio-, and chemo-) and specificity. In addition, most enzymes are very efficient and operate at mild conditions. It is, therefore, not surprising that enzymes have been increasingly used as biological catalysts or therapeutic agents in various industries, including the chemical, pharmaceutical, agricultural, and food industries. However, the number and diversity of enzyme-based applications are still modest compared with the total number of enzymes identified so far (~5000 enzymes) (1). One main reason for this functional gap is that naturally occurring enzymes are the products of Darwinian evolution and are not designed for optimal industrial applications. To address this limitation, several enzyme engineering approaches have been developed in the past few decades, among which directed evolution stands out as a particularly attractive approach. This entry discusses the brief history of directed evolution, the main methods of directed evolution, and their applications in engineering enzymes for basic and applied biology research. For more in-depth information on directed evolution, interested readers are referred to the Further Reading list.

A Primer for Directed Evolution

Before the advent of recombinant DNA technology in the 1970s, the ability to engineer novel enzymes was limited to chemical modification methods in which specific residues in an enzyme...
Directed evolution bypasses the bottleneck of rational design and mimics natural evolution in a test tube to evolve proteins without knowledge of their structures. What fundamentally differentiates directed evolution from natural evolution is its power to significantly accelerate the process of evolution. As shown in Fig. 1, directed evolution uses various methods to generate a collection of random protein variants, called a library, at the DNA level. Followed by screening/selection of the library, protein variants with improvement in desired phenotypes are obtained. Usually, the occurrence of these functionally improved protein variants is a rare event; thus, this two-step procedure has to be iterated several rounds until the goal is achieved or no further improvement is possible.

Methods for Directed Evolution

A successful directed evolution experiment involves two key components: creating genetic diversity and developing a high throughput screening or selection method. In the past decade, many experimental methods and protocols for library construction and screening/selection have been developed. For more information on this topic, interested readers are referred to the two books edited by Arnold and Georgiou in the Further Reading list.

Library creation

Numerous molecular biology methods have been developed to introduce genetic diversity into the target gene, all of which can be grouped into three categories: methods of random mutagenesis, methods of gene recombination, and methods of semirational design. As shown in Fig. 2 random mutagenesis starts from a single parent gene and randomly introduces point mutations or insertions/deletions into the progeny genes. In comparison, gene recombination usually starts from a pool of mutants from a single gene or a pool of closely related or even non-related parental genes of different origin and creates blockwise exchange of sequence information among the parental genes. Finally, semirational design combines rational design and directed evolution by focusing mutagenesis on a few selected important residues or regions in a target gene.

Random mutagenesis

As a result of its simplicity and efficiency, error-prone polymerase chain reaction (EP-PCR) is the most widely used random mutagenesis method. It is essentially a variation of the standard PCR with slightly modified reaction conditions. There are many different protocols to implement EP-PCR, and the most popular one includes the following adjustments to normal PCR conditions: 1) use of nonproofreading DNA polymerases, such as Taq DNA polymerase; 2) use of low or unbalanced amount of dNTPs; 3) use of high concentration of Mg²⁺ (up to 10 mM); and 4) incorporation of Mn²⁺. The fourth modification has made EP-PCR more popular, because the error rate can be controlled precisely by the Mn²⁺ concentration. In general, 1-2 amino acid substitutions are introduced during each round of EP-PCR, which requires approximately 1-5 base mutations per kilobase of DNA. Higher mutagenic rates are not normally used because they often damage enzyme function and lead to an increased tendency to negate positive mutations. In addition, higher mutagenic rates result in a larger library size, which in turn requires an often unattainable robust screening/selection method to identify positive variants. On the other hand, a higher mutation rate increases the frequency of multiple mutations with synergistic effects, resulting in an overall enrichment of unique protein variants, and up to 30 mutations per gene have been reported.

Fig. 1 General scheme of directed evolution.

Fig. 2 Random mutagenesis, directed evolution uses various methods to generate a collection of random protein variants, called a library, at the DNA level. Followed by screening/selection of the library, protein variants with improvement in desired phenotypes are obtained. Usually, the occurrence of these functionally improved protein variants is a rare event; thus, this two-step procedure has to be iterated several rounds until the goal is achieved or no further improvement is possible.

Directed evolution by focusing mutagenesis on a few selected important residues or regions in a target gene. In addition, more research has recently started addressing more complex systems, such as pathways (metabolic engineering), viruses, and even genomes.
Directed Evolution, Novel and Improved Enzymes through...

The great success of EP-PCR in engineering all aspects of enzyme properties has established this method as a cornerstone in directed evolution. It should be noted, however, that this technique is not truly random and suffers a number of limitations. In addition to the intrinsic bias of DNA polymerases (transitions are favored over transversions), EP-PCR can only access 5-6 amino acids substitutions on average at each residue because of the degeneracy of genetic codons and the low probability of two mutations occurring right next to each other. Another limitation of EP-PCR is associated with the low mutation rates normally used, such that the progeny protein variants have similar phenotype to the parent. Thus, novel functions are difficult to evolve using this method alone even after several rounds of iteration. To search the sequence space more extensively, EP-PCR is used in combination with other DNA diversity generation methods, such as gene recombination.

Gene recombination

Gene recombination can be implemented both in vivo and in vitro. However, the latter is used much more often because of its simplicity, higher recombination efficiency, and flexibility. Therefore, only in vitro methods will be discussed here. Note that all the available in vitro gene recombination methods fall into two main categories: homology-dependent and homology-independent.

Homology-dependent gene recombination

As nature has found homologous recombination a useful evolving tool, biologists have also recognized its power of achieving “long jump” in adaptive molecular evolution (8). And the advances in molecular biology made it possible to mimic this process in vitro. The first and most frequently used gene recombination method, DNA shuffling, also known as “sexual PCR”, was developed by Stemmer in 1994 (9). As shown in Fig. 3, the target gene is digested by DNaseI into random fragments, of which 100–300 bp fragments are purified and reassembled in a self-priming (no primers are added) PCR reaction according to their sequence homology. Recombination occurs when a fragment derived from one sequence anneals to a fragment derived from another sequence. This method was later adapted to recombine a family of naturally occurring homologous proteins from diverse species under modified conditions, which is called “family shuffling” (10). It was demonstrated that family shuffling significantly accelerated the rate of improvement of enzyme functions in comparison with EP-PCR and DNA shuffling.

As with every method, both DNA shuffling and family shuffling have their own limitations. First of all, both methods require relatively high homology, typically more than 70-75%, between the parental genes, because libraries created from more divergent sequences have a strong tendency to reassemble into parental genes. Various homology-independent methods have been developed to address this issue and will be discussed in the next session. Second, crossovers during template switching are favored in regions of high sequence identity, restricting the sequence space that can be explored. Third, fragments generated by DNaseI are not truly random, thus the diversity of the shuffled library is further decreased. Finally, there are also some nontechnical problems, such as limited access to natural sequence diversity and patent issues.

To address some of these limitations, a group of homologous gene recombination methods that do not involve DNA fragmentation but require addition of primers were developed, and staggered extension process (StEP) (11) was the first among them (Fig. 3). This method is essentially a modified PCR that uses very short extension time so that the elongation of short DNA fragments is staggered. During the subsequent rounds of DNA amplification, the fragments are repeatedly separated from the parental strand and prime a different one, resulting in multiple crossovers. StEP has several advantages over the original DNA shuffling method: 1) it needs only a small quantity of parental genes; 2) no digestion of DNA purification is needed, thus it is easy to be carried out; and 3) it avoids the DNaseI-induced bias. However, it should be noted that the StEP PCR conditions need to be optimized before a good library can be obtained, which might take a considerable amount of time.

Nonhomologous gene recombination

Incremental Truncation for the Creation of Hybrid enzymes (ITCHY) was the first developed homology-independent recombination method (12). Incremental truncation of two parental
Directed Evolution, Novel and Improved Enzymes through

(a) (b) (c) (d)

Figure 3

Comparison of various gene recombination methods, including (a) DNA shuffling, (b) StEP, (c) ITCHY, and (d) SHIPREC.

genes from both ends by exonuclease III under nonideal conditions generates a collection of all possible truncated fragments, which are subsequently blunt polished and ligated to give genes of various lengths. There are several limitations of ITCHY. First, the key to creating a successful ITCHY library is the tight control of the exonuclease digestion conditions, and aliquots of digestion mixture have to be taken at various time points to quench the reaction. Therefore, it can be very time-consuming and labor-intensive. To address this issue, the same group developed a modified method, called THIO-ITCHY (13). The incorporation of α-phosphothioate nucleotide analogs at low frequency in genes inhibits exonuclease III activity, thus avoiding the requirement for frequent removal of digestion samples. The second limitation of ITCHY is that because it is a single crossover process, the diversity of the created library is rather limited. A different method, named SCRATCHY (14), was developed by the same group to achieve multiple crossovers by shuffling two ITCHY libraries, thus increasing the diversity of the library. Third, the ITCHY library of hybrids is not full-length and thus the two truncated genes are not necessarily fused at sites where the gene sequences align (15). It was shown previously that although insertions or deletions at the fusion portion of two parental genes might not necessarily have a deleterious effect on the enzyme function, the predominance of crossovers at positions of precise alignment in the selected active hybrids (12) indicates the importance of the alignment. This problem led to the birth of another method, sequence homology-independent protein recombination (SHIPREC) (15). In this method, two parental genes are fused by a linker containing multiple restriction sites. After digestion by DNase I at both ends of the fusion gene, full-length genes are selected, circularized, and digested by restriction enzyme in the linker region to give linear chimerical genes. The selection of a full-length gene helps maintain the sequence alignment of two genes and gives a larger fraction of functional hybrids. Finally, not only ITCHY, but all the methodologies discussed above, have one common limitation: only two parental genes can be recombined. Therefore, a few other multiple-parental homology-independent recombination methods have been developed, such as exon shuffling (16) and nonhomologous random recombination (NRR) (17).

Semirational design

Although rational design enables efficient targeting at critical protein sites, this approach is often hindered by limited availability of crystal structures and poor understanding of the structure-function relationship. To circumvent the limitations of rational design, directed evolution found its position as the “blind watchmaker.” However, as it is a “blind” searching process, the diversity pool must be as extensive as possible, which leads to the bottleneck of directed evolution: library screening. Therefore, any means to decrease the library redundancy would be beneficial. More importantly, when the engineering goal is to dramatically alter an enzyme function, it usually requires multiple close mutations in the active site, which are difficult to access by full-length gene random mutagenesis and require an even larger library to be screened. Therefore, to allow a more focused and more useful sequence space to be explored, the most logical way would be to combine the best features of the two extreme methodologies. This process gave birth to the third library creation method, called semirational design.

The most popular semirational design strategy is targeted saturation mutagenesis. Functionally important residues are identified by analysis of protein crystal structures and mutated individually (18) or in combination (19) into the other 19 natural amino acids using degenerate primers (NNN or NNS, N = A/T/G/C, S = G/C). It should be noted that protein crystal structures are no longer the only source for identification of functionally important residues. When no protein structure information is available, key residues can be identified by EP-PCR, bioinformatics, or homology modeling. Another expanding area is in silico directed evolution, the ability of which to rationalize a huge protein database and to guide engineering experiments holds the possibility to create novel enzymes beyond the natural realm. Various algorithms have been developed recently to optimize library creation conditions, library design, and library
Spatial compartmentalization are two ways of tagging. (phenotype) followed by screening/selection (phenotype analysis) that is compatible with the tagging. Physical linkage and (genotype) and the protein it encodes in each category, a common principle underlying these assays through. Although various technologies have been developed its throughput is relatively low and it can only be used to screen As screening requires individual analysis of each protein variant, the analysis method must be prudently chosen or developed, because of the first principle of directed evolution "you get what you select (screen) for." There are two main categories of library analysis methods: screening and selection. Screening involves examining every mutant individually for the desired property, whereas selection is a method whereby only proteins with the desired property are carried through. Although various technologies have been developed in each category, a common principle underlying these assays exists: tagging the DNA and the protein it encodes (phenotype) followed by screening/selection (phenotype analysis) that is compatible with the tagging. Physical linkage and spatial compartmentalization are two ways of tagging.

Screening technologies
A screening requires individual analysis of each protein variant, its throughput is relatively low and it can only be used to screen small libraries (up to a size of ~10^4). The 96-well plate is the most widely used screening format due to its versatility, although higher spatial density formats can be used, such as 384- or 1536-well microtiter plates or even protein microarrays. In a microtiter plate assay, not only are the protein and its encoding DNA compartmentalized in one well, but also the whole reaction; therefore, it is most suitable for enzyme activity assays. In addition, the enzyme is analyzed in the same way as in traditional biochemical assays: Each protein sample in the form of cell cultures, crude lysates, or purified proteins is transferred into one well and then examined, thus the reaction conditions can be controlled to mimic the final practical conditions as closely as possible. With the aid of an automatic colony picker and liquid handler, the assays can be easily adapted into this high throughput format and automated. Although the microtiter plates only provide compartmentalization for the DNA-protein pair, methods are needed to analyze the proteins. Currently, colorimetric or fluorometric assays are the most popular and convenient screening methods, whereby the positive variants can be easily identified by visual check or by measuring UV-Vis absorbance or fluorescence using a plate reader. However, they are not available for all enzymes. Other generic screening tools, such as HPLC, capillary electrophoresis, and thermistor arrays, have also been applied to engineering of enzymes.

To address the low throughput limitation associated with most screening methods, various fluorescence-activated cell sorting (FACS) based screening methods have been developed. Unlike the above mentioned screening methods, FACS can analyze and sort up to 100,000 cells per second in a quantitative manner (22). The first application of FACS to directed enzyme evolution was demonstrated by Georgiou and his coworkers in 2000 (23). By coupling with bacterial surface display (see the Surface Display subsection to follow), FACS was successfully used to engineer a protease variant with improved catalytic activity. A fluorescence resonance energy transfer (FRET) substrate was designed to assay the protease activity in which a fluorescent dye is quenched by its FRET quenching partner via a target scissile bond recognized by the protease. Enzymatic cleavage of the scissile bond results in the release of the FRET quenching partner while the fluorescent dye is retained on the cell surface, allowing isolation of active clones by FACS. Remarkably, this method achieved 5000-fold enrichment of active clones in a single FACS round.

Selection technologies
Compared with screening technologies, library selection applies certain selection pressure/criteria to the mutant library so that only positive variants are carried to the next round while unwanted variants are discarded. Therefore, a much larger library of enzyme variants (more than 10^11) can be assessed. However, the selection methods are normally developed for a specific system or for analyzing a particular enzyme property. Many properties, such as enzyme activity at extreme temperatures or pH, or organic solvents, are not directly amenable to selection. As a result, screening is usually more applicable than selection. Based on the DNA-protein pair tagging method, selection methods can be divided into two categories: surface display and compartmentalization.

Surface display
Display technologies, employing nucleic acids, phage, yeast, or bacteria, were initially developed for binding assays and have made great success in engineering high affinity receptors, such as antibodies and T-cell receptors (24). Several inherent features of the display technologies made them suitable for directing enzyme evolution. First of all, display of proteins on the surface establishes a physical linkage between DNA and protein. Second, the proteins on the surface are accessible to external molecules, such as substrates or other target molecules.

Finally, the DNA is restricted inside the phage particle or microbial cells, enabling easy tracking of the genotype. With the need for enzyme engineering growing, researchers recognized the potential of display technologies and progressively adapted them for enzyme engineering.

Phage display is the most commonly used technique for in vitro selection. Filamentous bacteriophages (e.g., M13) are used for protein display for their ability to infect host cells without killing them (25). In a practical enzyme phage display experiment, a phagemid DNA library is constructed first in vitro and then transformed into competent bacterial cells. The DNA that encodes the enzyme of interest is fused to one of the coat protein genes (pVIII for high copy display, pII for low copy display), thus the enzyme is expressed as a fusion to the phage coat protein. During the phage assembly process, the target DNA is encapsidated inside the nascent phage particle as a part of its genome while its encoding enzyme is displayed on its surface; as a result, a physical linkage is established.
Directed Evolution, Novel and Improved Enzymes through...

between the phenotype and genotype through the phage particle. Phage particles are then harvested as a batch and selected for those displayed enzymes with improved/novel functions. Phage display selection is naturally based on binding. More specifically, a phage library is selected by passing it through an affinity matrix whereby binding phages are captured while nonbinding phages are washed away. Therefore, to adapt the phage display technique to engineer enzyme properties such as activity, selectivity, and stability, the key is to couple enzyme properties to the capture or release of the phage from the affinity matrix, for example, by codisplaying the enzyme and substrate on the same phage particle. As shown in Figure 4 upon catalysis, the product is displayed on the surface and recognized by the solid support. In contrast, phages displaying inactive enzymes cannot bind the affinity matrix and are washed away. Phage display has been successfully used to engineer enzymes with improved activity, altered substrate specificity, improved stability, and even novel function (25-28). However, it is almost impossible to develop a generic phage display system for all applications and phages lack posttranslation modification mechanisms that might be critical for functional expression of some enzymes.

Compartmentalization

Just as in microtiter plate format based screening, compartmentalization is also used in selection methods; each DNA-protein pair is spatially isolated in an individual compartment, which is either a cell (in vivo selection) or a manmade compartment (in vitro selection) instead of individual wells. Wherever accessible, in vivo selection is very powerful and can assess large numbers of mutants. The ultimate in vivo selection method would be, under a given selection pressure, only mutants harboring proteins with improvements could grow into colonies or show a significant phenotypic difference. Although it is a very powerful technique, the utility of in vivo selection is very limited, because most enzymes are of little direct biological relevance. Another reason is that the sophisticated genetic regulation networks of the host microorganism have evolved to encounter rapid changes in the environment, and thus the applied selection pressure may result in mutations out of the target genes. In vitro selection overcomes some of the limitations of in vivo selection. In vitro compartmentalization (IVC) (29) links the genotype and phenotype by colocalizing single genes together with necessary transcription and translation biochemical components in the aqueous compartments of a water-in-oil emulsion droplet. In most compartments, there is either no gene or only one gene that is later transcribed and translated in vitro within the same compartment. The enzymatic reaction is later carried out in the same droplet. To a certain extent, IVC is similar to microtiter plates but on a much smaller size scale with volumes close to those of bacteria (29). As the gene is transcribed and translated in vitro, general cloning is avoided and the library size is no longer limited by transformation efficiency. However, it seems that IVC can only be used to select enzymes that directly or indirectly act on DNA. For analyzing other enzyme properties, the droplets still need to be screened one by one, as in the case of 96-well plate screening. However, by combining with other technologies, such as FACS (30) or microbeads (31), IVC still holds promise for future enzyme engineering.

Applications of Directed Evolution

Directed evolution has been successfully used to alter existing enzyme properties and even to create novel enzyme functions. In addition to creating enzymes for specific industrial applications, directed evolution has also been increasingly used to address fundamental questions in biology, such as the evolutionary mechanisms of novel protein functions, protein structure-function relationship, and protein folding mechanisms.

Improving enzyme properties by directed evolution

Directed evolution has enjoyed great success in improving existing enzyme characteristics. In the following sections, only...
a few selected examples will be highlighted. Alterations have been made for almost all aspects of enzyme properties, such as substrate specificity, product specificity, selectivity, activity, stability, or folding/ solubility. Such alterations are required for enzymes to become practically useful biocatalysts or therapeutics.

**Substrate specificity**

Although the analogy of lock and key is sometimes used to describe the relationship between an enzyme and its substrate, in reality, the level of specificity varies. A particular enzyme may perform similar reactions on a range of related substrates or it may show tremendous specificity to one molecule. This aspect of enzymes can be exploited to develop variants with altered substrate recognition. In some cases, it may be beneficial to expand the range of substrates acted on. For example, poly-chlorinated biphenyls (PCBs) are a class of organic compounds whose use is decreasing due to concerns over their long-term environmental persistence and health effects. Certain bacteria can degrade some of these compounds by oxygenation reactions. Shuffling of two biphenyl dioxygenases from different bacteria resulted in higher activity, and activity on novel substrates such as toluene (32). In another example, EP-PCR was used to convert E. coli aspartate aminotransferase into a valine aminotransferase (33). A mutant enzyme with 17 amino acid substitutions was created that shows a $2.1 \times 10^6$-fold increase in the catalytic efficiency for a non-native substrate, valine.

Structural analysis of the mutant enzyme by protein crystallography indicated a remodeled active site and altered subunit interface caused by the accumulative effects of mutations. Most surprisingly, only one of the mutations directly contacts the substrate, which underscores our limited understanding of enzyme substrate specificity. These mutations would be difficult, if not impossible, to be identified and introduced to the mutant enzyme by a rational design approach.

**Product specificity**

In addition to altering an enzyme’s substrate, the product of an enzymatic reaction can be modified by using directed evolution. One example of product specificity engineering that has received attention is that of carotenoid pathway enzymes. Farnesylgeranyl diphosphate synthase catalyzes the condensation of isopentenyl diphosphate into a C25 isoprenoid molecule. The conversion of this isopentenyl diphosphate into a C25 isoprenoid molecule can be achieved by increasing the $k_{cat}$ or decreasing the $K_m$. A high throughput screening system was used with family shuffling of the thymidine kinase gene from herpes simplex virus I and II to increase the specificity of AZT phosphorylation (40). The authors used a robot to pick around 10,000 clones at each of four rounds of family shuffling, and they measured colony growth on different levels of AZT. Variants were found that conferred sensitivity to E. coli when exposed to 32-fold less AZT compared with HSV 1 thymidine kinase. These variants contained multiple crossovers and mutations affecting the binding site. A more stable enzyme was obtained using compartmentalization, which was used with site-directed saturation mutagenesis to screen libraries of phosphotriesterase for increased activity (31). Despite this enzyme already being very active, the $K_m$ was increased from $2280 \text{ s}^{-1}$ to $144,300 \text{ s}^{-1}$ (63-fold). The $k_{cat}$/$K_m$ was only increased slightly due to an increase in $K_m$, but at $1.76 \times 10^8 \text{ M}^{-1}\text{s}^{-1}$, it is approaching the diffusion-limited rate of catalysis.

**Selectivity**

Chiral molecules have important roles in the pharmaceutical and chemical industries. Enzymes have the capability to be exquisitely enantioslective, and applications of directed evolution in this area have been reviewed (37). Pioneering work was carried out on a lipase from Pseudomonas aeruginosa, by EP-PCR and saturation mutagenesis. Using a model reaction, the hydrolysis of 2-methyldecanoic acid-2-nitrophenyl ester, the enantioselectivity was increased from $E = 1.1$ to $E = 25.8$ (38). Carbohydrates are a large class of chiral molecules with essential roles in biology, and they can serve as useful precursors in chemical synthesis of complex organic molecules. Directed evolution has been used to alter the preferred stereoproducit of the condensation of dihydroxyacetone phosphate and glyceroldehyde 3-phosphate (39). Depending on the enzyme, these substrates can yield D-fructose-1, 6-bisphosphate or D-tagatose-1, 6-bisphosphate, which differ in the C4 stereocchemistry. DNA shuffling of tagatose-1, 6-bisphosphate aldolase shifted the preference from $>99:1$ in favor of tagatose-1, 6-bisphosphate to 4:1 in favor of fructose-1, 6-bisphosphate, due to mutation of four residues within the substrate binding pocket.

**Activity**

Enzymes show a wide variety of reaction rates, which can be expressed in terms of either their turnover number or catalytic efficiency. For practical purposes, a high reaction rate is desirable, and it can be achieved by increasing the $k_{cat}$ or decreasing the $K_m$. A high throughput screening system was used with family shuffling of the thymidine kinase gene from herpes simplex virus I and II to increase the specificity of AZT phosphorylation (40). The authors used a robot to pick around 10,000 clones at each of four rounds of family shuffling, and they measured colony growth on different levels of AZT. Variants were found that conferred sensitivity to E. coli when exposed to 32-fold less AZT compared with HSV 1 thymidine kinase. These variants contained multiple crossovers and mutations affecting the binding site. A more stable enzyme was obtained using compartmentalization, which was used with site-directed saturation mutagenesis to screen libraries of phosphotriesterase for increased activity (31). Despite this enzyme already being very active, the $K_m$ was increased from $2280 \text{ s}^{-1}$ to $144,300 \text{ s}^{-1}$ (63-fold). The $k_{cat}$/$K_m$ was only increased slightly due to an increase in $K_m$, but at $1.76 \times 10^8 \text{ M}^{-1}\text{s}^{-1}$, it is approaching the diffusion-limited rate of catalysis.

**Stability**

A common aim of directed evolution is to increase the stability of an enzyme to conditions of practical use that may be very different from those the enzyme naturally functions in. Factors such as heat, altered pH, and the presence of oxidants or organic solvents can lead to denaturation or loss of enzyme function. Many researchers have successfully increased the stability of an enzyme to thermal denaturation (41, 42). Work with p-nitrophenyl esterase increased the melting temperature $14^\circ C$ after six rounds of EP-PCR and recombination without forfeiting enzyme activity (41). As another example, phosphate dehydrogenase catalyzes the formation of phosphate from phosphite, by reducing NAD$^+$ to NADH. However, the usefulness...
of this enzyme as a means of regenerating NADH cofactors for industry was impeded by the low stability of the wild-type enzyme isolated from Pseudomonas putida. Four rounds of EP-PCR were used to identify 12 mutations that increased the half-life of the enzyme at 45°C by 7000-fold (42). Notably, family shuffling of 26 subtilisin genes produced variants with improved activity at either heat, pH 10, pH 5.5, or the presence of 35% dimethylformamide (43). Certain clones also showed better performance under combinations of these conditions.

Folding/solubility
Low solubility or improper folding may sometimes hamper the use of enzymes, particularly when expressed in a non-native host. A method of expressing proteins with a C-terminal GFP fusion to use fluorescence as a measure of the amount of correctly folded protein has been introduced (44). DNA shuffling produced variants of ferritin that showed increased solubility, even when they were recoloned without the GFP fusion. This assay has been used to produce proteins for X-ray crystallography structure determination (45). The protein nuclease diphosphate kinase from Pyrobaculum aerophilum is insoluble when expressed in E. coli, but after DNA shuffling, a functional variant with six mutations was found to have 90% solubility, which enabled its crystallization, and its structure was determined.

Creating enzymes with novel functions by directed evolution
One of the aspirations of directed evolution is to create new function in enzymes, which may be to carry out a reaction that has not been found in nature or may involve adding new control modalities. The challenge that the field is embracing is a significant one. Nature is conservative when it comes to the generation of new enzymes, as it retools existing structures for new functions rather than inventing a new scaffold for each different reaction. The (g)lu-8 barrel scaffold, for example, is the most common protein structure found in enzymes, and different enzymes with this scaffold carry out many different types of reactions. Protein engineers can take hope and inspiration from this in their attempts to create novel function in enzymes.

Novel substrate specificity
As already discussed, enzyme activity can be broadened to include substrates not previously acted on. But directed evolution can also yield enzymes with substrate recognition different from the starting point (46). DNA shuffling of two highly homologous triazine hydrolases produced variants that acted on triazines that neither parent had activity toward, which showed that examining small differences in sequence space can reveal new activities.

Altering substrate specificity sometimes follows a process of relating, followed by tightening. Collins et al. (47) used a clever dual selection strategy to alter the response of the LuxR transcription factor to different acyl-homoserine lactones. This type of research can produce modifiers of transcription with fine control by a desired ligand chosen so as to not interfere with other biological pathways. The response of LuxR was initially broadened from 3-oxo-hexanoyl-homoserine lactone (3OC6HSL) to accept a variety of straight-chain acyl-homoserine lactones. Negative selection was performed against response to 3OC6HSL, resulting in a variant that responded to straight-chain acyl-HSLs but not the original activator.

Novel activity
Novel functions can be incorporated within existing protein scaffolds that naturally have no activity for the desired reaction. Working within the β-lactamase enzyme scaffold, the activity of β-lactamase has been successfully introduced into glyoxalase II by a combination of rational design and directed evolution (48), which involved deletion of the original glyoxalase II substrate-binding domain, followed by the introduction of loops designed by examining metallo β-lactamases, EP-PCR, and DNA shuffling. The resulting enzyme had activity as a β-lactamase, albeit at much lower efficiency than seen for the native enzyme. New activity can also be incorporated into a noncatalytic protein scaffold, as demonstrated by the creation of triose phosphate isomerase activity within ribose-binding protein by computational design and EP-PCR (49). Current applications in this area rely on seminal design, with directed evolution typically used to increase the initial activity produced.

New ways of controlling enzyme function can also be introduced. Natural enzymes often exhibit some form of posttranslational regulation that affects their activity, which could take the form of interaction with a small molecule to enhance or inhibit activity in a particular environment. The maltose-binding protein can function as a switch when inserted into a gene such as β-lactamase (50), which enables a level of control over the desired reaction, based on the presence or absence of a molecule such as maltose.

Understanding natural enzyme evolution
The power of directed evolution to create and analyze tens of millions of protein variants not only enables one to engineer enzymes with desired properties for practical applications and to study the structure and function of proteins, but it also provides researchers the means to understand natural evolutionary processes. Rather than being restricted to the snapshot of sequence space found in extant genes, researchers can conduct evolutionary experiments on catalytic mechanisms or protein structure to better understand how current genes arose. As little as one mutation has been shown to confer on an enzyme the ability to carry out a new reaction. Single mutations were discovered that allowed two members of the mucenate lactonizing enzyme subgroup of the enolase superfamily to catalyze an additional reaction, that of the enzyme α-succinylbenzate synthase (51). Other work has revealed the ease through which a promiscuous enzyme function can be improved by orders of magnitude, with comparatively little effect on the enzyme’s main function (52).

Proposed pathways of protein fold evolution have been examined for the DNA methyltransferase superfamily, showing that the circularly permuted variants seen in nature can be generated in the laboratory via intermediates that retain function (53).
Directed evolution has been demonstrated to be very useful in modifying enzymes for practical applications, producing better stability, higher activity, and altered substrate specificity or product formation. Its influence will only increase, by producing enzymes that are more versatile and adaptable to future medical, industrial, and agricultural demands. The future is likely to see an increased pairing of rational design and directed evolution, as researchers generate more protein structures and improve their ability to identify optimal target areas of proteins for randomization. Ambitious applications are also likely to continue, leading to new ways of controlling enzyme activity and examples of dramatic reconfiguration of the starting enzyme’s function. It is still a long way off until researchers have the ability to design from first principles an enzyme for any given task, and, as such, directed evolution will continue to be an incredibly useful tool for many years to come.

References

Fear and anxiety can be a normal adaptive reaction to help cope with stress in the short term, but when the emotional, cognitive, and physical manifestations are long lasting, extreme, and disproportionate to threat, whether real or perceived, anxiety is maladaptive and has become a disabling disorder. Anxiety disorders may be deconstructed to elementary behaviors/symptoms that can be conceptualized as quantitative characters determined by the combined effects of several risk genes and nongenetic factors (e.g., early-life adversity). Progress in neurogenetics, molecular and cellular neuroscience, and neuroimaging is beginning to yield significant insights of how genetic and nongenetic factors contribute to specific manifestations of anxiety disorders. The aim of this overview is to summarize and integrate the current knowledge on anxiety-related macromolecular pathways and mechanisms initiated by genetic risk and environmental factors. These pathways interact with each other, often during specific periods of development, and could lead to alterations in the formation and function of neuronal circuits that encode emotional behavior.

Anxiety Disorders: Macromolecular Pathways and Interactions

Miklos Toth, Department of Pharmacology, Weill Medical College of Cornell University, New York, New York
doi: 10.1002/9780470048672.wecb665

Anxiety is a state characterized by feelings of fear, apprehension, and worry. Emotionally, anxiety causes a sense of dread or panic, and behaviorally, it can be associated with both voluntary and involuntary behaviors such as escaping or avoiding the source of anxiety. Also, anxiety is associated with specific physical manifestations including increased heart rate and blood pressure. Although anxiety is clinically different from depression, the two disorders are often comorbid. Indeed, it is generally difficult to find individuals with pure anxiety and pure depression. Here we focus on anxiety exclusively but will mention and discuss depression when the separation is not clear. Anxiety, similarly to other psychiatric disorders, has a significant genetic basis, and it is relatively stable in individuals over time (trait anxiety). Recent studies implicated candidate genes, each with a small contribution to the risk of anxiety disorders (1). Several of these candidate genes encode proteins that regulate neurotransmitter synthesis and metabolism or correspond to neurotransmitter receptors. Studies implicate an equally important role for environmental factors, including early-life adversity and maternal care, in the development of anxiety and anxiety-like behavior in animals (2–4). Some of these environmental effects also converge on neurotransmitter systems. Because the risk genes or environmental factors individually represent only a small contribution to anxiety, the current view is that interactions between several risk genes and environmental effects are necessary to lead to the symptoms of anxiety. This overview focuses on molecular systems related to anxiety such as neurotransmitters, their receptor, and, when known, downstream intracellular signaling pathways that ultimately regulate the development and function of neuronal networks that underlie the behavioral anxiety response.

Neuroticism and Anxiety

Neuroticism is a personality trait characterized by an enduring tendency to experience negative emotional states. Personality traits are underlying characteristics of an individual that can explain the major dimensions of human behavior. Neuroticism represents a continuum between emotional stability and instability, and most people fall in between the extremes (5–9). Neuroticism has a wide individual variability, but it is relatively stable in individuals over time (10). People with a high level of neuroticism respond more poorly to stress and are more likely to interpret ordinary situations as threatening and frustrating.
In addition, autonomic arousal is an integral part of neuroticism characterized by increased heart rate and blood pressure, cold hands, sweating, and muscular tension. Extroversion and openness, two other personality traits, are also part of the NEO Personality Inventory (NEO-PI) and the revised (R) NEO-PI (11), and Gray and McNaughton (12) argued that anxiety proneness is primarily captured by measures of neuroticism, together with a smaller contribution from the dimension of extroversion. Specifically, rotating the dimensions of neuroticism and extroversion by 45°, two new dimensions, anxiety (N+, E–), and impulsivity (N−, E+), were proposed (12).

Although neuroticism is not a disease per se, it predisposes individuals to anxiety disorders (12, 13). Neuroticism is a vulnerability factor for all forms of anxiety (14–16). A system established by the Diagnostic and Statistical Manual for Psychiatric Disorders in the United States, currently in its 4th edition (DSM-IV) text revision (TR) (American Psychiatric Association, 2000), sets the boundary at which a particular level of behavior becomes an anxiety disorder—a level often based on the number and the duration of symptoms. DSM is a categorical system based on the qualitative separation of disease states from the state of well-being. The DSM-IV TR category of anxiety disorders currently includes generalized anxiety disorder (GAD), simple phobia, posttraumatic stress disorder (PTSD), panic disorder, social phobia, and obsessive compulsive disorder (OCD) as discrete anxiety disorders. The International Classification of Diseases-10 (IC-10) is a similar system, but it is less frequently used in research (17).

Although DSM-IV and IC-10 are widely used for the clinical diagnosis of anxiety disorders, these categorical models are not based on the underlying biological pathophysiology and may not be optimal for the identification of genetic and environmental factors involved in anxiety disorders. The main reason is the complexity and heterogeneity of anxiety disorders that makes the association of genetic and/or environmental effects with overt anxiety phenotypes difficult. An alternative approach is to deconstruct anxiety disorders to elementary behavioral manifestations or specific symptoms that may represent less complex biological traits and/or environmental influences. The rationale is that a specific phenotype that consists of few elementary behaviors is likely linked to a more limited number of genes and environmental effects than complex disease phenotypes. These elementary behaviors are primarily state-independent (manifest in an individual regardless of whether illness is active). Similar to this principle is the term “endophenotype,” although it usually refers to an internal phenotype along the pathway between the genotype and disease. Evidence suggests that elementary behaviors/endophenotypes of psychiatric disease may be understood as quantitative traits (Fig. 1). This model is based on the notion that risk gene variants carried by a given individual in combination with various nongenic factors such as early-life adversity and nutrition produce anxiety symptoms with variable onset and severity depending on the strengths of the genetic and environmental factors (Fig. 3).

### Genes, Their Protein Products, and Associated Biological Pathways Implicated in Anxiety and Anxiety-Like Behavior

Using the techniques of quantitative behavior genetics, it became clear that roughly 20–60% of the variation in most personality traits has a genetic base, and broad personality traits are under polygenic influence (18, 19). Similarly, genetic epidemiological studies estimate that heritability in anxiety varies between 23 and 50% (1, 20, 21). Recently, genome-wide linkage studies have been performed by using Eysenck personality questionnaire (EPQ) (22, 23) to identify chromosomal regions associated with neuroticism. A two-point and multipoint non-parametric regression identified 1q, 4q, 7p, 11q, 12q, and 13q (24), whereas another similar study using multipoint non-parametric allele sharing and regression identified 1q, 3cen-t, 6q, 11q, and 12p (25), which confirms some links in the previous study. Still another study using a broad anxiety definition instead of the DSM-IV classification identified a linkage at chromosome 14 between 99 and 115 cM (26), and this finding replicated a linkage for a broad anxiety phenotype in a clinically based study (27). However, these studies have not yet identified specific genes. So far, the candidate gene approach has been more productive than linkage and association studies in implicating gene polymorphisms related to neuroticism and anxiety disorders. The candidate gene approach relies on prior biochemical studies that implicate various molecules, primarily neurotransmitters, their receptors, and signaling in anxiety. Anxiety disorders traditionally have been viewed as disturbances in neurotransmitter systems including the serotonin (5-HT), gamma-aminobutyric acid (GABA), and corticotropin releasing hormone (CRH) systems, among others. Many of these neurotransmitters and their receptors have also been identified as sites of action for anxiolytic drugs. Neurotransmitter systems and corresponding neurobiological pathways that are well established in anxiety both in human studies and animal experiments are discussed below.

A nimal studies, especially rodent studies, significantly contributed to the current knowledge on how genes, the environment, and their interaction may produce anxiety. As people with a high level of neuroticism respond more poorly to stress, certain inbred, selectively bred (28, 29), and knockout rodents strains (see below) have increased emotional reactions to stress.
(emotional activity). These behavioral responses include avoidance, hypoactivity/frezing, and autonomic arousal among others. Although some similarities are found between human and murine stress responses, complex anxiety phenotypes cannot be reproduced in animals. Nevertheless, emotional stress reactions in animals represent relatively simple behaviors that are evolutionarily conserved and quantifiable. A multitude of tests can measure emotional activity in rodents (30–34). Although all behavioral tests use a novel environment and/or fearful situation to produce avoidance, hypoactivity/frezing, and/or autonomic arousal, it seems that the underlying pathophysiology is test specific. Indeed, data with recombinant inbred mouse strains indicate that open field and elevated plus maze behaviors are linked to specific but partly overlapping sets of quantitative trait loci (35, 36). Furthermore, even these relatively simple behaviors can be dissected to more elementary behaviors with a smaller assembly of quantitative trait loci (36). However, it is not clear currently if mouse QTL data can be directly extrapolated to humans.

Many recently developed genetically manipulated mouse strains exhibit increased emotionality, often referred to as anxiety-like behavior. Most strains have constitutive genetic inactivation, but some strains are also available with a conditional allele. Although constitutive gene inactivation may elicit compensatory processes that can complicate the phenotype, this occurrence is not necessarily a disadvantage. Indeed, genetic risk in humans is also constitutive and present from early life. Probably more important is that polymorphisms in risk genes do not cause a complete functional loss; therefore, it may be more appropriate to analyze the behavioral phenotype of mice with a heterozygous inactivation of the risk gene.

### The 5-HT pathway and associated genes involved in anxiety

5-HT has long been associated with emotion and anxiety (37). 5-HT is synthesized from tryptophan by the rate-limiting enzyme tryptophan hydroxylase (TPH) in serotonergic neurons in the dorsal raphe nucleus and stored in axon terminals, respectively. In addition, the synaptic and extracellular levels of 5-HT are regulated by the 5-HTT transporter (S-HTT). Genetic risk for anxiety has been associated with all of these macromolecules.

#### Tryptophan hydroxylase

The two isozymes of TPH are as follows: 1) the classical TPH isoform, now termed TPH1, that is detected in the periphery, especially in the duodenum and in blood but not in the brain, and 2) the relatively recently identified TPH2 expressed exclusively in the brain (38). Several common single nucleotide polymorphisms are found around the transcriptional control region of TPH2. T allele carriers of a functional polymorphism in the upstream regulatory region of TPH2 (G-703T) were found to be overrepresented in individuals with anxiety-related personality traits (39, 40). In agreement with these data, T carriers have been shown to exhibit relatively greater activity in the amygdala than do G-allele homozygotes to affective expressions (41). Aactivation of the amygdala, indicated by increased blood flow, is a typical reaction to stress and to unpleasant and potentially harmful stimuli.

#### The serotonin transporter

The transporter removes 5-HT after its release into the synaptic cleft and returns it to the presynaptic terminal, where it is metabolized by monoaminooxidases or stored in secretory vesicles and results in the termination of postsynaptic serotoninergic effects (42). 5-HTT is the target of the selective serotonin reuptake inhibitors (SSRIs) that have been shown to be effective in certain anxiety disorders and depression (43).

The 5HTT transporter belongs to the family of Na^+K^+- dependent transporters, which shows a certain degree of similarity with the GABA and the dopamine transporters (42, 44). The 5-HTT has 12 transmembrane domains (TM) with a large extracellular loop between TM 3 and 4. Both the N- and C-termini are located within the cytoplasm. Growing evidence suggests that 5-HTT forms a homomultimer in the plasma membrane, although most studies suggest an autonomous function for each monomer (45). Amino acid substitution experiments indicate the importance of TM 1 as an important contributor to substrate, ion, and inhibitor interactions (46). In addition to TM 1, TM 3 has also been shown to play a role in substrate and inhibitor binding (47). Recently, it has been demonstrated that Tyr-95 and Ile-172 in transmembrane segments 1 and 3 of human 5-HTT interact to establish the high-affinity site for antidepressants (48). Although SSRIs have a rapid action on the transporter, their anxiolytic and antidepressant actions are delayed. It is believed that one factor involved in this delay is the activation of 5-HT₄ autoreceptors on serotonergic neurons that effectively reduces neuronal firing and 5-HT release rebalancing extracellular 5-HT levels (43). As the SSRI treatment is prolonged, the 5-HT₄ autoreceptors desensitizes, and firing activity is restored. The notion that an adaptive change underlies, at least partly, the delayed therapeutic effect of SSRIs is supported by the acceleration of the anxiolytic and antidepressant response by the concomitant administration of the 5-HT₄ antagonist pindolol (49). Once 5-HTL levels are significantly increased during chronic SSRI administration, it is assumed that the elevated 5-HT L levels act on specific 5-HT receptors to elicit anxiolytic and antidepressant effects. Although at least 13 different types of 5-HT receptors exist in mammals, a recent report using a specific behavioral test (novelty induced suppression of feeding) showed that the presence of post synaptic 5-HT₄ receptors is necessary for the anxiolytic effect of the SSRI fluoxetine (50).

Lesch et al. demonstrated that a functional 5-HT₄ promoter polymorphism is associated with the factor “neuroticism” in the revised NEO personality inventory (51). Specifically, individuals who carry the s/s (short promoter repeat) alleles of the 5-HTT, as compared with individuals with s/l (long) or l/l alleles, have increased neuroticism. Extension of these genetic studies to anxiety disorders by the same authors showed no differences in 5-HTT genotype distribution between anxiety patients and comparison subjects, but among anxiety patients, carriers of the s alleles exhibited higher neuroticism scores (52).
Anxiety Disorders: Macromolecular Pathways and Interactions

Figure 2  Anxiety pathways and their interactions. Two fundamentally different mechanisms involved in anxiety and anxiety-like behavior are proposed. The first mechanism is developmental, and its consequences are manifested later in life. Both genetic and environmental factors linked to anxiety can have a developmental origin, as illustrated by the example of the s allele of the 5-HTT, the deficiency in the 5HTT receptor, and variability in maternal care. The second possible mechanism is not developmental but based on “acute” or current molecular abnormalities that result in anxiety or anxiety-like behavior. Examples include deficiencies in GABAergic neurotransmission and abnormalities in the central CRH system, which have the direct behavioral output of anxiety. It is also proposed that the developmental mechanisms lead to anxiety by converging on the “acute” mechanisms.

Several later reports found association between the 5-HTT promoter polymorphism and anxiety-related traits (53–55). Also, recent meta-analyses of several prior studies found a small but significant association between 5-HTT polymorphism and some but not all measures of neuroticism/anxiety (56–58). Interestingly, imaging data found a significant increase in amygdala activity in subjects who carry the s allele, which indicates a functional link between 5-HTT polymorphism and anxiety (59). Consistent with these data, genetic inactivation of the 5-HTT gene in mice results in an increased anxiety-like phenotype (60).

The s allele is associated with reduced 5-HTT mRNA expression in vitro and in lymphoblasts (51), but functional imaging studies and postmortem samples generally show no effect of the 5-HTT genotype on 5-HTT expression in adult subjects (61, 62). To resolve this discrepancy, it has been hypothesized that the s allele acts primarily during development and less so in adult brain and would primarily affect brain development leading to anxiety in later life (Fig. 2). Consistent with this idea, pharmacological blockade of the 5-HTT in rats and mice during early postnatal life (that mostly corresponds to the third trimester in human) results in changes in emotional behavior (63, 64). However, it is not known how increased 5-HT levels during development lead to lifelong anxiety nor what 5-HT receptors are involved.

The 5-HT1A receptor

The G-protein-coupled 5-HT1A receptor emerged as another candidate gene in anxiety as early studies indicated a deficit in the receptor in panic disorder patients (65, 66). These data seem to be consistent with the anxiolytic effect of partial 5-HT1A receptor agonists in the treatment of generalized anxiety disorder, but the pharmacological inhibition of the receptor in animals does not elicit anxiety (67). Then in 1998, it was shown that the genetic inactivation of the 5-HT1A receptor in mice results in enhanced anxiety-like behaviors (68–70) along-side reduced immobility in the forced swim test (70) and tail suspension test (68, 69). Anxiety-related tests in these and follow-up studies included open field, elevated plus maze, zero maze, novelty-induced suppression of feeding, and some fear conditioning paradigms (68–71). More recently, human studies strengthen the association between a 5-HT1A receptor deficiency and certain forms of anxiety. A preliminary neuroimaging study reported a significant negative correlation between 5-HT1A receptor binding in the dorsolateral prefrontal, anterior cingulate, parietal, and occipital cortices and indirect measures of anxiety in healthy volunteers (72). Then, reduced receptor levels were reported in the anterior cingulate, posterior cingulate, and raphe by positron tomography in patients with panic disorder (73). No such association was found in posttraumatic stress disorder (74).

5-HT1A receptors are expressed at both postsynaptic locations in 5-HT target areas (including the amygdala, hippocampus, and cortex) and presynaptic sites on 5-HT neurons in the raphe nucleus as somatodendritic autoreceptors. Because autoreceptors control neuronal firing and consequently 5-HT release, and because an increase in extracellular 5-HT levels during development (see previous section) has been implicated in anxiety, it initially was believed that the anxiety-like phenotype of 5-HT1A receptor-deficient mice was because of increased 5-HT release. However, basal 5-HT levels are not altered, as measured by in vivo microdialysis, in 5-HT1A receptor-deficient mice, presumably because of the compensatory action of presynaptic 5-HT1B.
receptors (75–77), and expression of 5-HT$_{1A}$ receptors in forebrain regions rescued the anxiety phenotype of 5-HT$_{1A}$ receptor knockout mice (78), which suggests that the behavioral phenotype results from the absence of postsynaptic rather than the presynaptic 5-HT$_{1A}$ receptors. The anxiety-like phenotype of 5-HT$_{1A}$ receptor-deficient mice has been shown to be associated with background-specific perturbations in the prefrontal cortex GABA-glutamate system and in GABA$_A_\alpha$ receptor subunit expression in this region as well as in the amygdala (79, 80) (Fig. 2). Our recent studies implicate hippocampal abnormalities in some manifestations of the anxiety-like phenotype in 5-HT$_{1A}$ receptor-deficient mice (unpublished data). Presumably the lack of 5-HT$_{1A}$ receptors during development at these postsynaptic areas leads to neuronal network abnormalities that underlie the exaggerated behavioral responses in anxiogenic environments.

Considering the many studies with 5-HT$_7$ polymorphism (see above), it is surprising that relatively little is known about 5-HT$_{1A}$ receptor alleles in the context of anxiety. Although the common genetic polymorphism (–1019G/C) in the promoter region of the 5-HT$_{1A}$ receptor gene has been associated with panic disorder (81), more is known about the link of this polymorphism to depression/suicide (82) (but see Reference 83) and to antidepressant treatment response (84, 85). Nevertheless, the –1019C/G polymorphism has been found to have functional effects on gene expression. Specifically, the 5-HT$_{1A}$ G(–1019) allele fails to bind Deaf-1 and Hoxd5, two transcriptional repressors, in raphe-derived cells leading to upregulation of autoreceptor expression. Consistent with these data, an increase in 5-HT$_{1A}$ autorreceptor expression in individuals with the G/C genotype has been observed (86, 87). However, as mentioned earlier, autoreceptors may not be implicated in anxiety-like phenotypes least not in the mouse model. Deaf-1 has an opposite effect in neurons that model the postsynaptic site as it enhances 5-HT$_{1A}$ receptor promoter activity, and this enhancement is attenuated by the presence of the G allele (88). This cell-specific activity of Deaf-1 in 5-HT$_{1A}$ gene regulation could account for the region-specific decrease of receptor levels in postsynaptic 5-HT$_{1A}$ receptor described in anxiety and mood disorders (see above).

The GABA pathway

Abnormalities in GABA levels have been noted in subjects with anxiety disorders. Magnetic resonance spectroscopy studies have shown lower occipital cortex GABA concentrations in subjects with panic disorder (89) and with major depression (90), compared with healthy controls. GABA is synthesized from glutamate by glutamic acid decarboxylase (GAD). Its two isoforms, GAD65 and GAD67, are products of two independently regulated genes, GAD2 and GAD3, respectively. GAD65 is responsible for the synthesis of a small pool of GABA associated with nerve terminals and synaptic vesicles and can be rapidly activated in times of high GABA demand (91, 92). In contrast, GAD67 is responsible for the majority of GABA synthesis, and it is important in developmental processes (92). GABA interacts with the ionotropic GABA$_A_\alpha$ and metabotropic GABA$_B_\alpha$ receptors. GABA$_A_\alpha$ receptors in particular have been associated with anxiety as they control neuronal excitability. The GABA$_A_\alpha$ receptor is a pentameric ion channel composed of subunits from seven families (α1-3, β1-3, γ1-3, ε, δ, and θ)(93). Receptors that express α1 and α2, together with a β subunit and the γ2 subunit, are the predominant forms in the central nervous system.

GAD and anxiety

GABA neurotransmission has been linked to anxiety disorders, and therefore genes that encode GAD are reasonable candidate susceptibility genes for these conditions. In a multivariate structural equation model, several of the six single-nucleotide polymorphisms tested in the GAD3 region (encoding GAD67) formed a common high-risk haplotype that contributed to individual differences in neuroticism and impacted susceptibility across a range of anxiety disorders, including generalized anxiety disorder, panic disorder, agoraphobia, and social phobia, and also major depression (94). No such association was found with single-nucleotide polymorphisms in GAD2 (encoding GAD65) in this study. Nevertheless, genetic inactivation of GAD65 results in anxiety-like behavior in mice (95). These mice also show increased seizure sensitivity, but otherwise they show no overt developmental phenotype (91). It is not possible to study anxiety in GAD67$^{-/-}$ mice because they die of severe cleft palate during the first morning after birth (92). The less severe and more specific phenotype of GAD65$^{-/-}$ mice is probably because the overall GABA content is normal in GAD65$^{-/-}$ tissues and only the K$^+$ stimulated GABA release is reduced, whereas in GAD67$^{-/-}$ mice a severe depletion of GABA occurs.

GABA$_A_\alpha$ receptors and anxiety

A deficit in GABA$_A_\alpha$ receptors has been identified in the hippocampus and parahippocampus of patients with panic disorder and generalized anxiety disorders (96–98). Also, GABA$_A_\alpha$ receptor antagonists elicit anxiety in patients with panic disorder, which suggests an underlying deficit in receptor function in these individuals (99).

Animal studies suggest that a reduction in GABA$_A_\alpha$ receptors that contain α2 subunits is associated with withdrawal-induced anxiety (100, 101). Genetic inactivation of some other subunit genes has a similar effect. For example, heterozygous γ$^2$ mice have reduced numbers of GABA$_A_\alpha$ receptors and display anxiety phenotype in the elevated plus maze and the dark-light box tests (100, 102). In addition, γ$^2$ mice show increased responses in the passive avoidance paradigm. These behavioral alterations are associated with a lower single channel conductance, a pronounced deficit of functional receptors, and a reduction in α2/β3/δ2S subunit and the ε2S subunit of the GABA$_A_\alpha$ receptor clusters in cortex, hippocampus, and thalamus. Transgenic mice overexpressing either the mouse y$_{90}$ or γ$_{95}$ subunit of the GABA$_A_\alpha$ receptor showed no difference in anxiety-related behavior as compared with wild-type littermates (103). Compensation at the level of GABA$_A_\alpha$ receptor subunit expression and assembly often occurs when subunit expression is disturbed, which may explain the lack of phenotype in these mice. In contrast to the γ2 subunit, the deletion of the also abundant α3 subunit (104) does not increase anxiety or elicit other behavioral abnormalities (105–107), probably because lack of the α3 subunit is compensated and substituted by other α subunits (105). Although α2 subunit-deficient mice have been generated and

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
a point mutation in this subunit (H105-R) abolishes the anxiolytic effect of diazepam (108, 109), it is not entirely clear whether these mice have a change in anxiety-related behavior. Mice deficient for the α1 subunit show a faster habituation to a novel environment, which is not a typical measure of anxiety (110). A mentioned earlier, a strain of 5-HT1A receptor−/− mice have reduced expression of both the α2 and α1 subunits that could explain, at least partly, their increased anxiety phenotype (79). Fig. 2. “Knock-in” mice in the α9 subunit (H105-R) display enhanced trace fear conditioning to threat cues (102). More analysis showed that these knockin mice exhibit a 33% reduction in hippocampal (CA1 and CA3) α9 receptor subunits (102); thus, the phenotype may be simply because of a partial knock out. Finally, the genetic inactivation of β2, a predominant α subunit, resulted in a more than 50% reduction in the total number of GABA_A receptors and increased locomotor activity in open field, which suggests that these receptors may control motor activity (106). Taken together, these studies in general support the notion that a deficiency in GABA_A receptors results in anxiety and anxiety-like behavior, but it is difficult to assign specific subunit-containing receptors to anxiety.

## The BDNF pathway in anxiety

BDNF is a secretory protein that belongs to the neurotrophin family. A large proportion of neuronal BDNF is secreted in the proform (proBDNF), which is subsequently converted to mature (m) BDNF by extracellular proteases such as plasmin or matrix metalloproteinases (111). A functional single nucleotide polymorphism that produces a valine (Val)–to–methionine (Met) substitution at α6. 66 in pro-BDNF, first described as altering the intracellular trafficking and activity-dependent secretion of BDNF (112), has been studied extensively in association studies. The outcome of these studies has been variable. One study found that the Met 66 may be a risk allele for anxious temperament as measured by the Tridimensional Personality Questionnaire (113). Another study however concluded that the Val 66 allele is associated with greater neuroticism (114), whereas still another has shown no difference in neuroticism between the Val and Met genotypes (115). A more definitive outcome of the Val–Met polymorphism was reported with mice. Because the mouse does not have the Met allele, it was generated by knockin (116). BDNF/+/− mice exhibited increased anxiety-related behaviors in a variety of behavioral paradigms including the elevated plus maze, open field, and novelty-induced hypophagia tests (116). Although BDNF expression in BDNF+/− mice is equivalent to that in BDNF+/+ mice, a ∼30% deficit in activity-dependent release of BDNF from neurons occurs. This deficit suggests that the anxiety phenotype of BDNF+/− mice can be linked to the activity-dependent release of BDNF. Similarly to the BDNF+/− mice, BDNF+/+ mice have an anxiety phenotype that has BDNF levels lower than normal (by 50%); thus, it seems that a partial deficit in BDNF is sufficient to elicit anxiety. The receptor for mBDNF is TrkA, a receptor tyrosine kinase (117). Consistent with the increased anxiety phenotype of BDNF+/−, conditional BDNF−/−, and BDNF+/− mice (116, 118), transgenic mice overexpressing TrkA in postmitotic neurons in a pattern similar to that of the endogenous receptor display less anxiety in the elevated plus maze test (119).

Similarly to the serotonin-related genes and their polymorphisms, the Met allele of BDNF causes developmental brain abnormalities. Humans heterozygous for the Met allele have smaller hippocampal volumes (120) and perform poorly on hippocampal-dependent memory tasks (112, 121). Consistent with these data, BDNF+/− or BDNF+/− mice have a significant decrease in hippocampal volume as compared with WT mice (116). Also, a significant decrease in dendritic complexity in dentate gyrus neurons occurs in BDNF+/− and BDNF+/− mice. These data raise the possibility that the Met allele contributes to anxiety by modulating brain development (Fig. 2).

## The central CRH–CRH-R pathway and anxiety

CRH is a 41-amino acid neuropeptide in mammals (122), and it is an important mediator of the central stress response (122, 123). The biological function of CRH is determined by the amino end of the peptide C-terminus that binds to the extracellular binding pocket of the receptors CRH-R1 and R2, whereas its N-terminus contacts other sites on the receptor to initiate signaling (124, 125). CRH-R1 in the anterior pituitary is thought to be the subtype through which hypothalamic CRH primarily initiates its effect on pituitary ACTH release (123, 126). ACTH then induces the secretion of corticosteroids in the adrenal. CRH via CRH-Rs, however, has functions outside the hypothalamic–pituitary–adrenal (HPA) axis, in particular in the amygdala. Indeed, data support the notion that extra-hypothalamic CRH, presumably via the central nervous systems, is significantly involved in anxiety (127), whereas dysregulation of the HPA axis via hypothalamic CRH seems to be more characteristic for depression (128).

CRH An association between behavioral inhibition and three single nucleotide polymorphisms in the CRH gene, including a point mutation in the coding sequence of the gene, was found in children at risk for panic disorder (129). Behavioral inhibition is a trait that involves the tendency to display fearful, avoidant, or shy behavior in novel situations. In animal experiments, the central administration of CRH produces behavioral effects that correlate with anxiety, such as reduced exploration in a novel environment or enhanced fear response (130). Also, an anxiety-like phenotype has been described in transgenic mice overexpressing CRH (131, 132). However, mice with a deleted CRH gene, although they had significantly decreased basal corticosterone levels, showed no anxiety (133, 134). One possible explanation is that the central CRH system is redundant. Indeed, CRH-like peptides, in 1-3 (that also bind CRH-Rs), are present in the CNS (135-138). Two groups have generated mice with a deletion of the urocortin 1 gene (139, 140), but only one of these studies found an increased anxiety-like phenotype (139). Currently, no obvious explanation is found for the significant difference observed between the two urocortin 1-deficient mice. Finally, deletion of the CRH binding
protein, that normally binds and inactivates CRH, resulted in increased anxiety (141). The authors hypothesized that the inactivation of CRH-BP may increase the "free" or unbound levels of CRH or urocortin that lead to anxiety. As described earlier, hypothalamic CRH regulates the HPA axis. Some reports have indicated the dysregulation of the HPA axis in PTSD and panic disorders. Small but significant decreases in plasma cortisol levels and increased HPA axis sensitivity to low glucocorticoid negative feedback signals have been reported in PTSD (142). In contrast, other studies showed persistent increases in salivary cortisol levels in pediatric PTSD patients (143) and significantly greater CRH-induced ACTH and cortisol responses in women with chronic PTSD (144). In panic patients, abnormal HPA axis regulation, including increased basal cortisol secretion and overnight hypercortisolemia, have also been documented (129). These HPA axis changes, however, may not be specific for anxiety disorders but rather reflect the presence of comorbid depression (145).

CRH-Rs
No significant association was found between an intron 2 polymorphism in CRH-R1 and the "neuroticism" dimension of personality as assessed by the Revised NEO Personality Inventory in healthy Japanese subjects (146). Similarly, three CRH-R2 gene polymorphisms had no association with panic disorder in another study (147).

The role of CRH-Rs in anxiety-like behavior has been studied extensively by using knockout mice. Whereas mice lacking CRH-R1 display decreased anxiety in the light-dark box and the elevated plus maze (148-150), CRH-R2-deficient mice, generated independently by three groups, exhibit varying degrees of anxiety-related behavior. In one study, increased anxiety was reported in the elevated plus maze and open field but not in the light-dark box test (151); another study found anxious behavior in both the elevated plus maze and light-dark box, but only in males (152). However, this latter study showed an increased time spent by the knockout mice in the center of an open field, which is more consistent with reduced anxiety. Still another report found no significant change in anxiety behavior in the elevated plus maze or open field (153). More recently, a mouse with a conditional deletion of the CRH-R1 in forebrain, hippocampus, and the amygdala, but with normal expression in the anterior pituitary, showed markedly reduced anxiety-like behavioral responses in two avoidance tests, and it exhibited normal ACTH and corticosterone secretion to stress (154).

Taken together, some of behavioral data obtained with various CRH and CRH-R knockout mice suggest that extrahypothalamic CRH and/or urocortin mediate a dual modulation of anxiety behavior. A deletion of CRH-R1 seems to be anxiogenic, whereas activation of CRH-R2 is anxiolytic. Therefore, it may not be surprising that dual CRH-R1/2 knockout mice have only a subtle behavioral phenotype (155).

CRH signaling and anxiety
CRH-R1/2 are coupled to several signaling pathways, including the adenyl cyclase-protein kinase A (PKA)-CREB and the ERK/mitogen-activated protein kinase (MAPK) pathways (156). These pathways can also be linked to anxiety. For example, genetic inactivation of adenyl cyclase type B results in reduced anxiety-like behavior (157). Mice with mutations of the PKA RIIa subunit and CREB also exhibit abnormal anxiety responses (158-160), and activation of CREB in amygdala produces anxiety-like behavior (160). Thus, anxiety-like responses may be initiated and regulated by Gs-coupled CRH receptor signaling, at least partly, via the cyclic AMP–PKA–CREB pathway. The involvement of the ERK–MAPK pathway can also be implicated in CRH-related anxiety-like behavior. ERK1/2 is strongly activated in hippocampal CA1 and CA3 pyramidal cells and basolateral amygdala by the intracerebroventricular administration of CRH (161), and CRH-induced phosphorylation of ERK1/2 was absent in mice with a conditional knockout of forebrain and limbic CRH-R1 exhibiting a low level of anxiety (162). One may hypothesize that processing of anxiogenic stimuli is altered in the amygdala, hippocampus, or other relevant brain region as a result of abnormal CRH signaling that consecutively leads to the behavioral manifestations of anxiety (Fig. 2).

Environmental Factors and Related Biological Pathways Involved in Anxiety
Stressful life events and anxiety
Although most frequently associated with depression (163), stressful life events have also been linked to anxiety disorders (164-166). Kendler et al. attempted to determine if anxiety, specifically GAD symptoms and depression, are associated with different dimensions (humiliation, entrapment, loss, and danger) of stressful life events associated with high contextual threat. Onset of GAD symptoms was predicted by higher ratings of loss and danger, whereas depression was associated with the combination of humiliation and loss, which indicates that event dimensions that predispose to pure GAD episodes versus pure depression can be distinguished with moderate specificity (164). Unfortunately, little is known about the molecular and cellular mechanisms elicited by stressful life events and how these events predispose an individual to anxiety or depression. However, animal studies provided mechanistic insights of how early-life events can contribute to the development of adult anxiety (discussed below).

Maternal care and adult life anxiety in rodents
Brief "handling" of rat pups results in a lifelong decrease to behavioral and endocrine effects of stress, whereas animals separated from their mothers/litters for longer periods of time, for example, for several hours, exhibit increased anxiety (167). Later studies determined that the critical effect of short-term handling is the increase in maternal care (licking and grooming) after the return of the pups to the nest (3). More studies showed that rat pups nursed by mothers selected for either a high or a low level of licking and grooming
Interaction Between Genetic and Environmental Factors

Although the interaction of genes and environment in shaping behavior is well accepted, direct experimental evidence to support their role in the pathogenesis of psychiatric diseases has been difficult to obtain. However, recent association studies with the 5-HTT polymorphism have indicated, at least in depressive disorders, that genetic and environmental factors act together, enhancing the phenotype beyond the level established by either factor alone. In 2003, Caspi et al. reported that carriers of the short allele of the 5-HTT polymorphism have indicated, at least in depressive disorders, that genetic and environmental factors act together, enhancing the phenotype beyond the level established by either factor alone. In 2003, Caspi et al. reported that carriers of the short allele of the 5-HTT polymorphism have increased risk for depression, anxiety, and related disorders. The 5-HTT short allele is associated with increased transcriptional activity, increased expression of the serotonin transporter, and decreased availability of serotonin in the synaptic cleft. This, in turn, affects the sensitivity of the postsynaptic receptors and their ability to respond to serotonin, leading to an increased risk for depressive and anxiety disorders.

Convergence of Anxiety-Related Pathways and Mechanisms

Two fundamentally different mechanisms associated with anxiety and anxiety-like behavior seem to exist: one that has a developmental origin with the cause and the adult manifestations of the phenotype separated in time and another mechanism that presents itself in "acute" settings. Typical examples for developmental anxiety are those caused by the short allele of the 5-HTT, the deficiency in the 5HT1A receptor and BDNF, whereas monkeys reared normally were not. In rodents, several environments × gene interaction studies related to anxiety behavior have been conducted. For example, maternal care. Perhaps more plausible mechanisms are the numerous neurochemical changes related to differences in maternal care. For example, rat pups of high LG–ABN dams show altered GABAergic receptor subunit expression in the amygdala, locus coeruleus, medial prefrontal cortex, and hippocampus that could contribute to their reduced anxiety-like behavior as compared with pups from low LG–ABN dams (172, 173) (Fig. 2). In addition to the GABAergic system, other potential factors mediating the environmental effects include the glutamatergic system and neurotrophins such as BDNF. Liu et al. found that increased LG–ABN of offspring resulting in increased hippocampal mRNA expression of NR2A and NR2B NMDA receptor subunits at postnatal day 8, a change that was sustained into adulthood (174). Also, increased levels of BDNF, but not NGF or NT-3, mRNA were observed in the dorsal hippocampus of 8-day old high LG–ABN pups (174). Neuronal network changes that could directly explain the behavioral consequences of maternal care have also been found as adult offspring of high LG–ABN dams show increased hippocampal synaptogenesis as compared with low LG–ABN offspring, and it has been found that this change could be normalized when low LG–ABN pups are cross-fostered to high LG–ABN dams (174) (Fig. 2).
and the variability in maternal care. All of these produce their effect in adults only when present during prenatal and/or early postnatal life. Indeed, genetic inactivation of the 5HT1A receptor during adult life is not accompanied by anxiety. Because the individual genetic risk factors mentioned above have subtle effects, a combination of them may be necessary to lead to a significant level of anxiety. For example, 5-HT1A and 5-HT2C dual mutants show a more pronounced anxiety phenotype as compared with singly heterozygous mice (189).

In other anxiety forms and models, developmental mechanisms do not seem to play a role. A typical example is represented by a deficient GABAergic neurotransmission (as a result of less GABA availability or altered composition of the GABA receptors) that results in reduced neuronal inhibition and increased excitation. The direct behavioral output of these neuronal and neuronal network changes is anxiety (Fig. 2). Another example is the central CRH system as its pharmacological manipulation acutely alters anxiety levels. Because many developmentally relevant anxiety genes and environmental effects and their corresponding mechanisms (see above) modulate GABAergic and glutamatergic transmission as well as central CRH signaling, it is possible that the developmental mechanisms converge on and are interconnected sequentially to the "acute" mechanisms that ultimately lead to the anxiety behavior (Fig 2).

In summary, it seems that anxiety in many cases has a developmental origin whether it is elicited by genetic polymorphisms or environmental effects. Because individual genetic influences have small effects, it is believed that multiple genetic risk genes, together with adverse environment, are required to have a large enough impact to cause anxiety. Although some early signs of anxiety may be manifested during childhood, anxiety becomes more apparent in predisposed individuals as the brain and behavior mature during adolescence. Although relatively little is known about how early developmental mechanisms lead to anxiety in later life, data suggest that long-lasting alterations in neurotransmitter systems (GABA, glutamate, and CRH) and/or the morphology and function of neuronal networks (amygdala, hippocampus, etc.) are involved. Additional environmental influences during adolescence and adulthood can increase the incidence and severity of anxiety.

References


Anxiety Disorders: Macromolecular Pathways and Interactions


Anxiety Disorders: Macromolecular Pathways and Interactions


Anxiety Disorders: Macromolecular Pathways and Interactions


154. Bale TL, Picetti R, Contarino A, Kool GF, Vale WW, Lee KF. Mice deficient for both corticotropin-releasing factor receptors
Anxiety Disorders: Macromolecular Pathways and Interactions


Further Reading


See Also

Brain Development, Neurochemistry of Neurrotransmission, Chemical Events in Synaptic Chemistry Systems Approach to Studying Disease
Chronic Obstructive Pulmonary Disease (COPD), Inflammatory Mechanisms of
Peter J. Barnes, National Heart and Lung Institute, Imperial College, London, UK
doi: 10.1002/9780470048672.wecb673

Airway obstruction in chronic obstructive pulmonary disease (COPD) is caused by narrowing of small airways as a result of inflammation and fibrosis and the disruption of their alveolar attachments as a result of emphysema. COPD is characterized by a complex inflammatory disease process that increases as the disease progresses, which leads to increasing airflow limitation. Many inflammatory cells and mediators have now been implicated in the pathogenesis of COPD. Increased numbers of macrophages, neutrophils, T-lymphocytes (particularly CD8+ cells) and B-lymphocytes have been observed, as well as the release of multiple inflammatory mediators (lipids, chemokines, cytokines, growth factors). Macrophages seem to play an important role in orchestrating the inflammatory process, which includes the recruitment of neutrophils and T-cells into small airways and lung parenchyma and the secretion of proteinases that lead to emphysema. A high level of oxidative and nitrative stress may amplify this inflammation through the reduction in histone deacetylase-2, which also results in corticosteroid resistance.

Chronic obstructive pulmonary disease (COPD) has now become a major global epidemic, and it is predicted to become the third leading cause of death and fifth leading cause of disability over the next decade (1). The increase in COPD is a particular problem in developing nations; in developed countries, it is the only common cause of death that is increasing. COPD now has a world-wide prevalence of over 10% in men and is increasing toward this figure in women (2). Because of the enormous burden of disease and escalating health-care costs, there is now renewed interest in the underlying cellular and molecular mechanisms of COPD (3) and a search for new therapies (4). The definition of COPD adopted by the Global Initiative on Obstructive Lung Disease encompasses the idea that COPD is a chronic inflammatory disease, and much recent research has focused on the nature of this inflammatory response (5). COPD is an obstructive disease of the lungs that slowly progresses over many decades leading to death from respiratory failure unless patients die of comorbidities such as heart disease and lung cancer before this stage. Although the most common cause of COPD is chronic cigarette smoking, some patients, particularly in developing countries, develop the disease from inhalation of wood smoke from biomass fuels or other inhaled irritants (2). However, only about 25% of smokers develop COPD, which suggests that genetic or host factors may predispose patients to its development, although these factors have not yet been identified. The disease is relentlessly progressive, and only smoking cessation reduces the rate of decline in lung function; as the disease becomes more severe, there is less effect of smoking cessation, and lung inflammation persists.

COPD as an Inflammatory Disease

The progressive airflow limitation in COPD is caused by two major pathological processes: remodeling and narrowing of small airways and destruction of the lung parenchyma with consequent destruction of the alveolar attachments of these airways as a result of emphysema (Fig. 1). This disease results in diminished lung recoil, higher resistance to flow and closure of small airways at higher lung volumes during expiration, which traps air in the lung. This trapped air leads to the characteristic hyperinflation of the lungs, which causes the sensation of dyspnea and limits exercise capacity. The major symptom of COPD is shortness of breath on exertion. Both the small airway remodeling and narrowing and the emphysema are
Chronic Obstructive Pulmonary Disease (COPD), Inflammatory Mechanisms of  

Figure 1 Small airways in COPD patients. The airway wall is thickened and infiltrated with inflammatory cells, predominately macrophages and CD8\(^+\) lymphocytes, with increased numbers of fibroblasts. In severe COPD, lymphoid follicles are observed, surrounded by T-lymphocytes and thought to indicate chronic exposure to antigens (bacterial, viral, or autoantigens). Similar changes are also reported in larger airways. The lumen is often filled with an inflammatory exudate and mucus. Peribronchial fibrosis occurs, and it results in progressive and irreversible narrowing of the airway. Airway smooth muscle may be increased slightly.

Figure 2 Amplification of lung inflammation in COPD. Normal smokers have a mild inflammatory response, which represents the normal (probably protective) reaction of the respiratory mucosa to chronic inhaled irritants. In COPD, this same inflammatory response is markedly amplified, and this amplification increases as the disease progresses. It is increased even more during exacerbations triggered by infective organisms (Fig. 2). The molecular basis of this amplification of inflammation is not yet understood but may be, at least in part, determined by genetic factors. Cigarette smoke and other irritants in the respiratory tract may activate surface macrophages and airway epithelial cells to release chemotactic factors that then attract circulating leukocytes into the lungs. Among chemotactic factors, chemokines predominate and therefore play a key role in orchestrating the chronic inflammation in COPD lungs and its amplification during acute exacerbations (8). These events might be the initial inflammatory events that occur in all smokers. However in smokers who develop COPD, this inflammation progresses into a more complicated inflammatory pattern of adaptive immunity and involves T- and B-lymphocytes and possibly dendritic cells along with a complicated interacting array of cytokines and other mediators (9).

Differences from asthma  

Histopathological studies of COPD show a predominant involvement of peripheral airways (bronchioles) and lung parenchyma, whereas asthma involves inflammation in all airways (particularly proximal airways) but usually without involvement of the lung parenchyma (10). In COPD, the bronchioles become narrow, with fibrosis and infiltration with macrophages and T-lymphocytes, along with destruction of lung parenchyma and an increased number of macrophages and T-lymphocytes, with a greater increase in CD8\(^+\) (cytotoxic) than CD4\(^+\) (helper) cells (6). Bronchial biopsies show similar changes with an infiltration of macrophages and CD8\(^+\) cells and an increased number of neutrophils in patients with severe COPD. Bronchoalveolar lavage (BAL) fluid and induced sputum demonstrate a marked increase in macrophages and neutrophils. In contrast to asthma, eosinophils are not prominent except during exacerbations or when patients have concomitant asthma (10).
Inflammatory Cells

For many years it was believed that the inflammatory reaction in the lungs of smokers consisted of neutrophils and macrophages and that proteases from these cells were responsible for the lung destruction in COPD. More recently it has been recognized that there is a prominent T-cell infiltration in the lungs of patients with COPD, with a predominance of CD8+ (cytotoxic) T-cells, although CD4+ (helper) T-cells are also numerous. Although abnormal numbers of inflammatory cells have been documented in COPD, the relationship between these cell types and the sequence of their appearance and their persistence are not yet understood in detail (3). Most studies have been cross-sectional based on selection of patients with different stages of the disease, and comparisons have been made between smokers without airflow limitation (normal smokers) and those with COPD who have smoked a similar amount. Nonserial studies have been conducted, and selection biases (such as selecting tissue from patients suitable for lung volume reduction surgery) may give misleading results. Nonetheless, a progressive increase in the numbers of inflammatory cells in small airways and lung parenchyma are observed as COPD becomes more severe, even though the patients with most severe obstruction have stopped smoking for many years (6). This finding indicates the existence of some mechanisms that perpetuate the inflammatory reaction in COPD. This characteristic is in contrast to many other chronic inflammatory diseases, such as rheumatoid arthritis and interstitial lung diseases, in which the inflammation tends to diminish in severe disease. The inflammation of COPD lungs involves both innate immunity (neutrophils, macrophages, eosinophils, mast cells, NK cells, γδ T-cells, and dendritic cells) and adaptive immunity (T and B cells).

Epithelial cells

Epithelial cells are activated by cigarette smoke to produce inflammatory mediators, which include tumor necrosis factor (TNF)-α, interleukin (IL)-1β, IL-6, granulocyte-macrophage colony-stimulating factor (GM-CSF), and CXCL8 (IL-8). Epithelial cells in small airways may be an important source of transforming growth factor (TGF)-β, which then induces local fibrosis. Vascular endothelial growth factor (VEGF) seems to be necessary to maintain alveolar cell survival and blockade of VEGF receptors in rats induces apoptosis of alveolar cells and an emphysema-like pathology, which may be mediated via the sphingolipid ceramide (11). Airway epithelial cells are also important in defense of the airways, with mucus production from goblet cells, and secretion of antioxidants, antiproteases and defensins. It is possible that cigarette smoke and other noxious agents impair these innate and adaptive immune responses of the airway epithelium, which increases susceptibility to infection. The airway epithelium in chronic bronchitis and COPD often shows squamous metaplasia, which may result from increased proliferation of basal airway epithelial cells, but the nature of the growth factors involved in epithelial cell proliferation, cell cycle, and differentiation in COPD are not yet known. Epithelial growth factor receptors (EGFR) show increased expression in airway epithelial cells of smokers and may contribute to basal cell proliferation, which results in squamous metaplasia and an increased risk of bronchial carcinoma (12).
Neutrophils

Increased numbers of activated neutrophils are found in sputum and BAL fluid of patients with COPD (13); yet neutrophil levels are increased relatively little in the airways or lung parenchyma. This finding may reflect their rapid transit through the airways and parenchyma. The role of neutrophils in COPD is not yet clear; however, neutrophil numbers in induced sputum are correlated with COPD disease severity (13) and with the rate of decline in lung function. Smokers have a direct stimulatory effect on granulocyte production and release from the bone marrow and survival in the respiratory tract, which is possibly mediated by GM-CSF and G-CSF released from lung macrophages. Smoking may also increase neutrophil retention in the lung. Neutrophil recruitment to the airways and parenchyma involves adhesion to endothelial cells and E-selectin, which is upregulated on endothelial cells in the airways of COPD patients. A different neutrophil then migrate into the respiratory tract under the direction of neutrophil chemotactic factors. Several chemotactic signals have the potential for neutrophil recruitment in COPD, which include leukotriene (LT)BA, CXCL8, and related CXC chemokines, including CXCL1 (GRO-α) and CXCL5 (ENA-78), which are increased in COPD airways (14). These mediators may be derived from alveolar macrophages T-cells and epithelial cells, but the neutrophil itself may be a major source of CXCL8. Neutrophils from the circulation migrate in the pulmonary circulation and adhere to endothelial cells in the alveolar wall before passing into the alveolar space. The neutrophils recruited to the airways of COPD patients are activated because increased concentrations of granule proteins, such as myeloperoxidase and human neutrophil lipocalin, are found in the sputum supernatant (15). Neutrophils secrete serine proteases from neutrophils, which include neutrophil elastase, cathepsin G, and proteinase-3, as well as matrix metalloproteinase (MMP)-8 and MMP-9, which may contribute to alveolar destruction. Neutrophils have the capacity to induce tissue damage through the release of serine proteases and oxidants. However, whereas neutrophils have the capacity to cause elastolysis, this ability is not a prominent feature of other pulmonary diseases in which chronic airway neutrophilia is even more prominent, including cystic fibrosis and bronchiectasis. This comparison suggests that other factors are involved in the generation of emphysema. Indeed, neutrophils are not a prominent feature of parenchymal inflammation in COPD. It is likely that airway neutrophilia is more linked to mucus hypersecretion in chronic bronchitis. Serine proteases from neutrophils, which include neutrophil elastase, cathepsin G, and proteinase-3, are all potent stimulants of mucus secretion from submucosal glands and goblet cells in the epithelium. A marked increase in neutrophil numbers is observed in the airways in acute exacerbations of COPD, which accounts for the increased purulence of sputum. This finding may reflect increased production of neutrophil chemotactic factors, which include LTBA and CXCL8 (16, 17).

Macrophages

Macrophages seem to play a pivotal role in the pathophysiology of COPD and can account for most of the known features of the disease (13) (Fig. 4). A marked increase (5–10-fold) in the numbers of macrophages in airways, lung parenchyma, BAL fluid, and sputum in patients with COPD. A careful morphometric analysis of macrophage numbers in the parenchyma of patients with emphysema showed a 25-fold increase in the numbers of macrophages in the tissue and alveolar space compared with normal smokers (19). Furthermore, macrophages are localized to sites of alveolar wall destruction in patients with emphysema, and a correlation is observed between macrophage numbers in the parenchyma and severity of emphysema (20). Macrophages may be activated by cigarette smoke extract to release inflammatory mediators, which includes TNF-α, CXCL8, and other CXC chemokines; CCL2 (MCP-1); LTB4; and reactive oxygen species. This release is a cellular mechanism that links smoking with inflammation in COPD. Alveolar macrophages also secrete elastolytic enzymes, which include MMP-2; MMP-9; MMP-12; cathepsins K, L, and S; and neutrophil elastase taken up from neutrophils (21). A leukovascular macrophages from patients with COPD and with exposure to cigarette smoke secrete more inflammatory proteins and have a greater elastolytic activity at baseline than those from normal smokers (21). Macrophages demonstrate this difference even when maintained in culture for 3 days, and therefore they seem to be intrinsically different from the macrophages of normal smokers and nonsmoking normal control subjects (21). The predominant elastolytic enzyme secreted by alveolar macrophages in COPD patients is MMP-9. Most inflammatory proteins that are upregulated in COPD macrophages are regulated by the transcription factor nuclear factor-κB (NF-κB), which is activated in alveolar macrophages of COPD patients, particularly during exacerbations (22).

The increased numbers of macrophages in smokers and COPD patients may be caused by increased recruitment of monocytes from the circulation in response to the monocyte-selective chemokines CCL2 and CXCL1, which are increased in sputum and BAL of patients with COPD (14). Monocytes from patients with COPD show a greater chemotactic response to GRO-α than cells from normal smokers and nonsmokers, but this finding is not explained by an increase in CXCR2 (23). Interestingly, whereas all monocytes express CCR2, which is the receptor for CCL2, only ~30% of monocytes express CXCR2. It is possible that these CXCR2-expressing monocytes transform into macrophages that are more inflammatory. Macrophages also release the chemokines CXCL9, CXCL10, and CXCL11, which are chemotactic for CD8+ Tc1 and CD4+ Th1 cells, via interaction with the chemokine receptor CXCR3 expressed on these cells (24).

The increased numbers of macrophages in COPD are mainly caused by increased recruitment of monocytes, as macrophages have a very low proliferation rate in the lungs. Macrophages have a long survival time so the macrophage level is difficult to measure directly. However, in macrophages from smokers, a markedly increased expression of the antiapoptotic protein Bcl-2 and increased expression of p21WAF1 is observed in the cytoplasm (25). This finding suggests that macrophages may have a prolonged survival in smokers and patients with COPD. Once activated, macrophages will increase production of reactive oxygen species, nitric oxide, and lysosomal enzymes...
CCL2, and CD8
steroid resistance.
peroxynitrite formation (29). COPD patients may be mediated through oxidative stress and secretion of cytokines like TNF-α. HDAC activity in macrophages is correlated with increased receptors to switch off inflammatory genes. The reduction in recruited to activated inflammatory genes by glucocorticoid are inhibited by corticosteroids, whereas corticosteroids are ineffective in macrophages from patients with COPD (27). The reasons for resistance to corticosteroids in COPD and to a lesser extent macrophages from smokers may be the marked reduction in activity of histone deacetylase-2 (HDAC2) (28), which is recruited to activated inflammatory genes by glucocorticoid receptors to switch off inflammatory genes. The reduction in HDAC activity in macrophages is correlated with increased secretion of cytokines like TNF-α and CXCL8 and reduced response to corticosteroids. The reduction of HDAC activity on COPD patients may be mediated through oxidative stress and peroxynitrite formation (29).

Eosinophils
Although eosinophils are the predominant leukocyte in asthma, their role in COPD is much less certain. Increased numbers of eosinophils have been described in the airways and BAL of patients with stable COPD, whereas others have not found increased numbers in airway biopsies, BAL, or induced sputum. The presence of eosinophils in patients with COPD predicts a response to corticosteroids and may indicate coexisting asthma (30). Increased numbers of eosinophils have been reported in chronic bronchitis and BAL fluid during acute exacerbations of chronic bronchitis (31). Surprisingly, the levels of eosinophil basic proteins in induced sputum are as elevated in COPD, as in asthma, despite the absence of eosinophils, which suggests that they may have degranulated and are no longer recognizable by microscopy (15). Perhaps this finding is caused by the high levels of neutrophil elastase that have been shown to cause degranulation of eosinophils.

Dendritic cells
Dendritic cells play a central role in the initiation of the innate and adaptive immune response, and it is believed that they provide a link between them (32). The airways and lungs contain a rich network of dendritic cells that are localized near the surface, so that they are located ideally to signal the entry of foreign substances that are inhaled. Dendritic cells can activate a variety of other inflammatory and immune cells, which include macrophages and neutrophils, as well as T- and B-lymphocytes, so dendritic cells may play an important role in the pulmonary response to cigarette smoke and other inhaled noxious agents. However, an increase in dendritic cells is not observed in the airways of COPD patients in contrast to asthma patients (33).

T-lymphocytes
An increase in the total numbers of T-lymphocytes is observed in lung parenchyma as well as in peripheral and central airways of patients with COPD; a greater increase is observed in CD8+ than CD4+ cells (6, 24). A correlation is observed between the numbers of T-cells and the amount of alveolar destruction and the severity of airflow obstruction. Furthermore, the only significant difference in the inflammatory cell infiltrate in asymptomatic smokers and smokers with COPD is an increase in T-cells, mainly CD8+, in patients with COPD. An increase in the absolute number of CD4+ T-cells, albeit in smaller numbers, is evidenced in the airways of smokers with COPD, and these cells express activated STAT-4, which is a transcription factor that is essential for activation and commitment of the Th1 lineage, and IFN-γ.

The ratio of CD4+/CD8+ cells are reversed in COPD. Most T-cells in the lung in COPD are of the Th1 and Th1 subtypes (24). A marked increase is observed in T-cells in the walls of
small airways in patients with severe COPD, and the T-cells are formed in lymphoid follicles, which surround B-lymphocytes (6).

The mechanisms by which CD8+ and to a lesser extent CD4+ cells, accumulate in the airways and parenchyma of patients with COPD is not yet understood (34). However, homing of T-cells to the lung must depend on some initial activation (only activated T-cells can home to the organ source of antigenic products), then adhesion and selective chemotaxis. CD4+ and CD8+ T-cells in the lung of COPD patients show increased expression of CXCR3, which is a receptor activated by the chemokines CXCL9, CXCL10, and CXCL11, all of which are increased in COPD (35). Increased expression of CXCL10 by bronchiolar epithelial cells is observed, and it could contribute to the accumulation of CD4+ and CD8+ T-cells, which preferentially express CXCR3 (36) (Fig. 5). CD8+ cells are typically increased in airway infections, and it is possible that the chronic colonization of the lower respiratory tract of COPD patients by bacterial and viral pathogens is responsible for this inflammatory response. It is possible that cigarette-induced lung injury may uncover previously sequestered autoantigens, or cigarette smoke itself may damage lung interstitial and structural cells and make them antigenic (37). The role of increased numbers of CD4+ cells in COPD, particularly in severe disease, is also unknown (38); however, it is now clear that T-cell help is required for the priming of cytotoxic T-cell responses, for maintaining CD8+ T-cell memory, and for ensuring CD8+ T-cell survival. It is also possible that CD4+ T-cells have immunological memory and play a role in perpetuating the inflammatory process in the absence of cigarette smoking. In a mouse model of cigarette-induced emphysema, there is a predominance of T-cells that are directly related to the severity of emphysema (38).

The role of T-cells in the pathophysiology of COPD is not yet certain, although they have the potential to produce extensive damage in the lung. CD8+ cells have the capacity to cause cytolysis and apoptosis of alveolar epithelial cells through release of perforins, granzyme B, and TNF-α (39, 40). An association between CD8+ cells and apoptosis of alveolar cells is observed in emphysema (41). A possible explanation is that cigarette-induced lung injury is responsible for this inflammatory and structural abnormality. In addition, CD8+ T-cells also produce several cytokines of the Th1 phenotype, which include IFN-γ, lymphotoxin, and TNF-α, and evidence suggests that CD8+ T-cells in the lungs of COPD patients express IFN-γ (42). All these cytokines would enhance the inflammatory reaction in the lung besides the direct killing by CD8+ T-cells. COPD has been considered an autoimmune disease triggered by smoking, as previously suggested (37), and the presence of highly activated glioblastoma T-cells in emphysema patients supports this conclusion (43). Evidence suggests that anti-elastin antibodies exist in experimental models of COPD and in COPD patients (44). In addition to activated Th1 cells, some evidence indicates an increase in Th2 cells that express IL-4 in BAL fluid of COPD patients in COPD patients (45).

Mediators of Inflammation

Many inflammatory mediators have now been implicated in COPD, which include lipids, free radicals, cytokines, chemokines, and growth factors (7). These mediators are derived from inflammatory and structural cells in the lung and interact with each other in a complex manner.

Lipid mediators

The profile of lipid mediators in exhaled breath condensates of patients with COPD shows an increase in prostaglandins and leukotrienes (46). A significant increase in PGE2 and F2α, and an increase in LTB4 but not cysteinyl-leukotrienes is observed. This increase is a different pattern to that observed in asthma, in
which increases in thromboxane and cysteinyl-leukotrienes have been shown. The increased production of prostanoids in COPD is likely to be secondary to the induction of cyclo-oxygenase-2 (COX2) by inflammatory cytokines, and increased expression of COX2 is found in alveolar macrophages of COPD patients. LTB₄ concentrations are also increased in induced sputum, and concentrations of LTB₄ are increased even more in sputum and exhaled breath condensate during acute exacerbations (16). LTB₄ is a potent chemoattractant of neutrophils, which acts through high-affinity BLT₁-receptors. A BLT₁-receptor antagonist reduces the neutrophil chemotactic activity of sputum by approximately 25% (47). Recently BLT₁-receptors have been identified on T-lymphocytes, and evidence indicates that LTB₄ is involved in recruitment of T-cells.

Oxidative stress

Oxidative stress occurs when reactive oxygen species (ROS) are produced in excess of the antioxidant defense mechanisms and result in harmful effects, which include damage to lipids, proteins, and DNA. Increasing evidence suggests that oxidative stress is an important feature in COPD (48). Inflammatory and structural cells that are activated in the airways of patients with COPD produce ROS, such as neutrophils, eosinophils, macrophages, and epithelial cells. Superoxide anions (O₂⁻) are generated by NADPH oxidase and converted to hydrogen peroxide (H₂O₂) by superoxide dismutases. H₂O₂ is then dismutated to water by catalase. O₂⁻, OH⁻, and H₂O₂ may interact in the presence of free iron to form the highly reactive hydroxyl radical (OH). O₂⁻ may also combine with NO to form peroxynitrite, which also generates OH. Oxidative stress leads to the oxidation of arachidonic acid and the formation of a new series of prostanoid mediators called isoprostanes, which may exert significant functional effects, such as bronchoconstriction and plasma exudation (49) (Fig. 6).

The normal production of oxidants is counteracted by several antioxidant mechanisms in the human respiratory tract (48). The major intracellular antioxidants in the airways are catalase, SOD, and glutathione, which is formed by the enzyme γ-glutamyl cysteine synthetase and glutathione synthetase. In the lung, intracellular antioxidants are expressed at relatively low levels and are not induced by oxidative stress, whereas the major antioxidants are extracellular. Extracellular antioxidants, particularly glutathione peroxidase, are markedly upregulated in response to cigarette smoke and oxidative stress. Extracellular antioxidants also include the dietary antioxidants vitamin C (ascorbic acid) and vitamin E (α-tocopherol), uric acid, lactoferrin, and extracellular superoxide dismutase, which is highly expressed in human lung, but its role in COPD is not yet clear. ROS have several effects on the airways and parenchyma and increase the inflammatory response. ROS activate NF-κB, which switches on multiple inflammatory genes resulting in amplification of the inflammatory response. The molecular pathways by which oxidative stress activates NF-κB have not been fully elucidated, but several redox-sensitive steps must be followed in the activation pathway. Oxidative stress results in activation of histone deacetylase activity, which opens up the chromatin structure and is associated with increased transcription of multiple inflammatory genes (50). Exogenous oxidants may also be important in worsening airway disease. Considerable evidence suggests increased oxidative stress in COPD (48). Cigarette smoke itself contains a high concentration of ROS. Inflammatory cells, such as activated macrophages and neutrophils, also generate ROS, as discussed above. Several markers of oxidative stress may be detected in the breath, and several studies have demonstrated increased production of oxidants, such as H₂O₂, 8-isoprostane, and ethane, in exhaled air or breath condensates, particularly during exacerbations (16).

Figure 6 Oxidative stress in COPD. Oxidative stress plays a key role in the pathophysiology of COPD and amplifies the inflammatory and destructive process. ROS from cigarette smoke or from inflammatory cells (particularly macrophages and neutrophils) result in several damaging effects in COPD, which include decreased antiprotease defenses, such as α1-antitrypsin (AT) and secretory leukoprotease inhibitor (SLPI), activation of NF-κB resulting in increased secretion of the cytokines CXCL8 and TNF-α, increased production of isoprostanes, and direct effects on airway function. In addition, recent evidence suggests that oxidative stress induces steroid resistance.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
The increased oxidative stress in the lung epithelium of a COPD patient may play an important pathophysiological role in the disease by amplifying the inflammatory response in COPD. This increase may reflect the activation of NF-κB and AP-1, which then induce a neutrophilic inflammation via increased expression of CXC chemokines, TNF-α, and MMP-9. Oxidative stress may also impair the function of antiproteases such as α1-antitrypsin and SLPI, and thereby accelerates the breakdown of elastin in lung parenchyma. Corticosteroids are much less effective in COPD than in asthma and do not reduce the progression or mortality of the disease. Alveolar macrophages from patients with COPD show a marked reduction in responsiveness to the anti-inflammatory effects of corticosteroids, compared with cells from normal smokers and nonsmokers (27). In patients with COPD, a marked reduction in activity of HDAC and reduced expression of HDAC2 is observed in alveolar macrophages and peripheral lung tissue (28), which is correlated with increased expression of inflammatory cytokines and a reduced response to corticosteroids. This finding may result directly or indirectly from oxidative stress and is mimicked by the effects of H2O2 in cell lines (51).

Nitrative stress
The increase in exhaled NO is less marked in COPD than in asthma, partly because cigarette smoke reduces exhaled NO. Recently, exhaled NO has been partitioned into central and peripheral portions and this shows reduced NO in the bronchial fraction but increased NO in the peripheral fraction, which includes lung parenchyma and small airways (52). The increased peripheral NO in COPD patients may reflect increased expression of inducible NO synthase in epithelial cells and macrophages of patients with COPD (53). NO and superoxide anions combine to form peroxynitrite, which nitrates certain tyrosine residues in protein, and increased expression of 3-nitrotyrosine is observed in peripheral lung and macrophages of COPD patients (53). Tyrosine nitration of HDAC2 may lead to impaired activity and degradation of this enzyme, which results in steroid resistance (51).

Inflammatory cytokines
Cytokines are the mediators of chronic inflammation, and several have been implicated in COPD (7). An increase in concentration of TNF-α is observed in induced sputum in stable COPD with an additional increase during exacerbations (13, 17). TNF-α production from peripheral blood monocytes is also increased in COPD patients and has been implicated in the cachexia and skeletal muscle apoptosis found in some patients with severe disease. TNF-α is a potent activator of NF-κB, which may amplify the inflammatory response. Unfortunately, anti-TNF therapies have not proved to be effective in COPD patients. IL-1β and IL-6 are other proinflammatory cytokines that may amplify the inflammation in COPD and may be important for systemic circulation.

Chemokines
Chemokines are small chemotactic cytokines that play a key role in the recruitment and activation of inflammatory cells through specific chemokine receptors. Several chemokines have now been implicated in COPD and have been of particular interest since chemokine receptors are G-protein coupled receptors, for which small molecule antagonists have now been developed (8). CXCL8 concentrations are increased in induced sputum of COPD patients and increase even more during exacerbations (13, 17). CXCL8 is secreted from macrophages, T-cells, epithelial cells, and neutrophils. CXCL8 activates neutrophils via low-affinity specific receptors CXCR1, and is chemotactic for neutrophils via high-affinity receptors CXCR2, which are also activated by related CX3 chemokines, such as CXCL1. CXCL1 concentrations are markedly elevated in sputum and in BAL fluid of COPD patients, and this chemokine may be more important as a chemoattractant than CXCL8, acting via CXCR2 that are expressed on neutrophils and monocytes (14). CXCL1 induces significantly more chemotaxis of monocytes of COPD patient compared with those of normal smokers, and it may reflect increased turnover and recovery of CXCR2 in monocytes of COPD patients (23). CXCL8 shows a marked increase in expression in airway epithelial cells during exacerbations of COPD; this increase is accompanied by a marked up regulation of epithelial CXCR2.

CCL2 is increased in concentration in COPD sputum and BAL fluid (14) and plays a role in monocyte chemotaxis via activation of CCR2. CCL2 seems to cooperate with CXCL1 in recruiting monocytes into the lungs. The chemokine CCL5 (RANTES) is also expressed in airways of COPD patients during exacerbations and activates CCR5 on T-cells and CCR3 on eosinophils, which may account for the increased eosinophils and T-cells in the wall of large airways that have been reported during exacerbations of chronic bronchitis. As discussed above, CXCR3 are upregulated on T-cells and T-cells of COPD patients are more responsive to the chemokine CXCL9 than to CXCL8, which is important as a chemoattractant than CXCL8, acting via CXCR2 that are expressed on neutrophils and monocytes (14). CXCL1 induces significantly more chemotaxis of monocytes of COPD patient compared with those of normal smokers, and it may reflect increased turnover and recovery of CXCR2 in monocytes of COPD patients (23). CXCL8 shows a marked increase in expression in airway epithelial cells during exacerbations of COPD; this increase is accompanied by a marked up regulation of epithelial CXCR2.

Growth factors
Several growth factors have been implicated in COPD and mediate the structural changes that are found in the airways. TGF-α is expressed in alveolar macrophages and airway epithelial cells of COPD patients, and it is released from epithelial cells of small airways. TGF-β is released in a latent form and is activated by various factors, which include MMP-9. TGF-β may play an important role in the characteristic peribroncholar fibrosis of small airways, either directly or through the release of connective tissue growth factor (Fig. 7). TGF-β downregulates β-adrenergic receptors by inhibiting gene transcription in human cell lines, and it may reduce the bronchodilator response to β-agonists in airway smooth muscle. Alveolar macrophages produce TGF-α in greater amounts than TGF-β, which may be a major endogenous activator of EGFR that plays a key role in regulating mucus secretion in response to many stimuli, which include cigarette smoke. Cigarette smoke induces TNF-α-converting enzyme on airway epithelial cells, which results in the shedding of TGF-α and the activation of EGFR, resulting in increased mucus secretion (54) (Fig. 8).

VEGF is a major regulator of vascular growth and is likely to be involved in the pulmonary vascular remodeling that occurs as a result of hypoxic pulmonary vasoconstriction in
fibroblasts or indirectly via the release of CTGF. TGF-α activates MMP-9. This may cause fibrosis directly through effects on fibroblasts or indirectly via the release of CTGF. TGF-β may also desensitize TGF-α receptors on cells such as airway smooth muscle to diminish the bronchoconstrictor response to β-agonists.

In summary, cigarette smoke exposure induces a florid inflammatory response in the lung that involves structural and inflammatory cells and a large array of inflammatory mediators. The interaction of these complex steps eventually leads to airway remodeling and obstruction and emphysema, albeit in only about 25% of chronic smokers. Of interest, the main difference between smokers who develop COPD and the ones who do not seems to be the presence of an adaptive immune response with CD8+, CD4+, and B-cells, which express obvious signs of being activated effector cells. It is likely that genetic and epigenetic factors (such as histone acetylation) are involved in determining the progression of the inflammatory cascade, as it is supported by animal models, where different strains seem to have different sensitivities to cigarette smoke. COPD is a complex inflammatory disease, and the interactions between different inflammatory cells and mediators are still uncertain. More research into these mechanisms is needed to identify novel targets that may lead to the discovery of more effective therapies that can prevent disease progression and reduce the high mortality of this common disease.

Conclusions

References


Chronic Obstructive Pulmonary Disease (COPD). Inflammatory Mechanisms of


Although the etiology of depression remains to be elucidated, our knowledge of the neurobiology and biochemistry of this mental disorder has been updated increasingly. Early research focused on the biogenic amines norepinephrine (NE; noradrenaline) and 5-hydroxytryptamine (5-HT; serotonin). The biogenic amine (monoamine) hypothesis of depression states that depression is the result of a functional deficiency of NE and/or 5-HT at central synapses. However, it soon became obvious that other factors were also important in the neurobiology of depression. Dopamine is thought to be involved in various aspects, which include reward and locomotion. The amino acids gamma-aminobutyric acid (GABA) and glutamate, their receptors, and various neuroactive steroids that act as allosteric modulators at GABAA and/or NMDA glutamate receptors have been proposed to have an important role. Stress and the hypothalamic-pituitary-adrenal (HPA) axis play a central role, and it has been proposed that increased secretion of corticotropin releasing factor (CRF) may be critical in producing the symptoms of depression. Various other neurochemicals have also been implicated in depression, and in several cases they have links to the HPA axis. Abnormal levels of Substance P have been reported in depression. Cell loss seems to occur in some brain areas (e.g., hippocampus) in depression, and such loss can be produced by stress and prevented by antidepressants (which also increase expression of various neurotrophic and transcription factors). Agonists at melatonin receptors have been proposed in recent years as effective antidepressants. The possible involvement of the various neurochemicals mentioned above in depression will be reviewed, and some other neurochemicals that are being examined will also be mentioned.

Introduction

Depression (referring here to major depressive disorder in the Diagnostic and Statistical Manual of the American Psychiatric Association, volume IV; DSM-IV) is a common, chronic, functionally limiting clinical syndrome that likely reflects a heterogeneous group of underlying disorders (1, 2). In the DSM-IV (DSM-IV-TR is the most recent version), characteristic symptoms of a major depressive episode listed include five or more of the following symptoms, which have been present nearly every day in most cases during the same 2-week period and represent a change from previous functioning: 1) depressed mood; 2) markedly diminished interest or pleasure in all, or almost all, activities; 3) significant weight loss or gain; 4) insomnia or hypersomnia; 5) fatigue or loss of energy; 6) feelings of worthlessness or excessive or inappropriate guilt; 7) feelings of worthlessness or excessive or inappropriate guilt; 8) diminished ability to think or concentrate, or indecisiveness; and 9) recurrent thoughts of death or suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide. Note: of the five symptoms, at least one is either 1) or 2) above. The lifetime prevalence of depression in the United States has been estimated at 5–12% in men and 10–25% in women (2). Depression has been shown to impact individual...
quality of life significantly (3), carry a heavy economic burden (4), and affect brain structure and function (5). Despite research into many potential causative factors, a putative etiology for depression remains elusive (6, 7). Research in the twentieth century initially focused on a decreased functional availability of the monoamine neurotransmitters norepinephrine (NE) and 5-hydroxytryptamine (5-HT) as the possible agents of depressive illness, and it was supported in part by the pharmacological properties of available antidepressants. However, a growing body of literature suggests that multiple contributory chemical systems, many of which are intertwined, are involved in producing the depressive phenotype (8–10).

Biogenic Amines

Endogenous biogenic amines in the brain include catecholamines [NE (norepinephrine, NA), dopamine (DA)], epinephrine (adrenaline), 5-HT, histamine, and the so-called trace amines (β-phenylethylamine, tyramine, tryptamine, and octopamine). These amines have in common aarrièreamine structure, and all have been implicated in the etiology of one or more psychiatric disorders and/or in therapeutic and/or adverse effects of drugs used to treat such disorders. In this review on depression, the focus is on the case of biogenic amines will be on 5-HT, NE, and DA, although epinephrine and histamine and trace amines have also been implicated (see the section on “Other Antidepressant Approaches and Targets”). Early studies on the origins of depression focused on the monoamine neurotransmitters NE (a catecholamine) and 5-HT (an indolamine) (for reviews, see References 7, 11, and 12) (Fig. 1). The antihypertensive and antipsychotic medication reserpine depletes central stores of NE, 5-HT, and DA and is known to produce symptoms of depression in some patients (13). Iproniazid is an antitubercular drug that was noted to produce mood elevation in tuberculosis patients and was shown to be an inhibitor of monoamine oxidase (MAO), which is a major catabolic enzyme for monoamines, including NE, 5-HT, and DA (another catecholamine neurotransmitter). In addition, studies in animals with drugs that caused depletions of NE and/or 5-HT demonstrated changes in locomotor activity, sleep, and sexual activity; these symptoms are similar to physiological symptoms observed in depressed patients. These observations led to catecholamine (14) and indolamine (15, 16) hypotheses, which are now generally combined into the biogenic amine (monoamine) hypothesis that states that depression is the result of a functional deficiency of NE and/or 5-HT at specific synapses in the central nervous system (12). In addition to the effects of the MAO inhibitors (several of which became approved antidepressants), the hypothesis was supported by the actions of the tricyclic antidepressants (TCAs), which inhibit the reuptake of NE and 5-HT back into nerve terminals. Such reuptake is a major inactivation mechanism for the catecholamines and 5-HT, and inhibition of this process leaves more NE and 5-HT available in the synaptic cleft between neurons to interact with postsynaptic receptors (Fig. 2). Indeed, most antidepressants developed subsequently inhibit NE and/or 5-HT reuptake and/or act on presynaptic receptors that affect synthesis and release of biogenic amine neurotransmitters (see Fig. 2).

Serotonin (5-Hydroxytryptamine, 5-HT)

Regulation of mood, sleep, and aggression have all been shown to involve the serotoninergic system (17–19), and most antidepressant drugs currently being used inhibit 5-HT reuptake and/or act on 5-HT receptors (of which there are several subtypes). 5-HT is produced centrally from the amino acid tryptophan, and depressed mood can be induced experimentally by acute tryptophan depletion in healthy individuals. This effect is accentuated in those with a family history of depression (18–21). Similarly, depressive relapse can be initiated in individuals treated with MAO inhibitors or selective serotonin reuptake (SSRI) inhibitors by depleting tryptophan (22, 23).

A major problem with the biogenic amine hypothesis of depression is the discrepancy in the time course between relatively rapid biochemical and pharmacologic effects (e.g., inhibition of reuptake, inhibition of MAO) of the antidepressants and their clinical beneficial effects, which often require administration for 2–3 weeks or longer. This discrepancy led to the studies in both laboratory animals (receptor-binding studies and electrophysiological investigations) and humans (postmortem studies and in vivo neuroimaging studies) on possible dysregulation of receptors for the biogenic amines in depression and effects of antidepressants on that dysregulation (24–28). It has been suggested that antidepressants normalize the density and/or function

![Figure 1] Structures of NE, 5-HT, and DA.

![Figure 2] Sites of actions of antidepressants at the synapse (presynaptic nerve terminal, postsynaptic nerve cell), and synaptic cleft between the two are shown. MAO is present in mitochondria, and the shaded circles represent synaptic vesicles that contain neurotransmitter amines.
of 5-HTT1 and 5-HTT2 receptors, but there are inconsistencies in the literature in this regard. Electroconvulsive shock in rats (an approximate animal model of electroconvulsive therapy in humans) seems to cause an opposite effect on 5-HTT2 receptor regulation to that of several antidepressants (24, 25, for a review see Reference 26).

The 5-HT transporter protein (5-HTT) is critical for the reuptake of 5-HT into the presynaptic neuron. Decreased 5-HTT binding has been reported in depressed patients both in postmortem samples and in functional imaging studies (28, 29). Two common alleles of the gene that encode this protein have been identified, with the short (s) form of the allele being less active, which results in decreased transcription and reduced expression of 5-HTT. Studies have suggested that the s allele occurs more frequently in depressed patients as well as in suicide victims, and homozygous individuals are more likely to have family histories of depression. This finding seems paradoxical, as the s allele of the gene results in decreased serotonin reuptake, which thereby increases the duration of 5-HT in the synapse. It has been suggested that this observation may be explained by the lifelong duration of the genetic polymorphism compared with the acute effect of medication administration (28, 29).

Norepinephrine (Noradrenaline)

Investigations have also been conducted on the regulation of NE receptors in depression and after chronic administration of antidepressants (11, 30–32). In animal models, induced chronic stress has been used as a model for depression. Chronic stress antecedents (11, 30–32). In animal models, induced chronic stress has been shown to reverse antidepressant responses in individuals treated with antidepressants that inhibit NE reuptake, which suggests a direct role of this neurotransmitter in the course of depression (31).

Dopamine

Although the focus on monoamine neurotransmitters in depression has been on NE and 5-HT, it is generally considered that DA (Fig. 1) also plays a role, and readers are referred to several comprehensive papers that review basic science and clinical evidence for the involvement of DA (11, 33–35). Motor, psychomotor speed, concentration, andanhedonia have all been linked to dopaminergic circuits. Most DA-producing neurons have nuclei located in the brain stem. Projections from dopaminergic neurons form three primary paths to the cortical and subcortical structures: The nigrostriatal pathway is involved in motor planning and execution; the mesocortical pathway is concerned with concentration and executive functions; and the mesolimbic pathway is important in motivation, pleasure, and reward. Mesolimbic DA dysfunction is observed in animal models of depression, with subsequent antidepressant use causing enhanced dopamine signaling (11, 33–35). Decreased levels of DA in the nucleus accumbens of rats are associated with a decreased response to rewards. Animals that experience learned helplessness have decreased levels of DA in the caudate nucleus and nucleus accumbens; exposure to antidepressant medications has been reported to increase the level of DA in the same regions (11, 33, 34).

In humans, genetic studies have shown that particular polymorphisms of the D1 and D2 DA receptors are associated with depressed phenotypes. In studies of CSF concentrations of homovanillic acid (HVA), which is a metabolite of DA, lower levels than controls have been reported in some depressed patients. Conversely, depressed individuals with psychotic symptoms have been reported to have increased CSF levels of HVA and DA compared with controls (33).

Classes of Antidepressants

The three neurotransmitter monoamines mentioned above do not operate independently, but rather they interact with one another and are influenced by various central and peripheral biological processes. The actions of most commercially available antidepressant medications continue to target the monoamine neurotransmitter systems. The structures of these antidepressants are shown in Fig. 3. The MAO inhibitors (tranylcypromine, phenelzine, and moclobemide) prevent metabolic breakdown of the amines; the TCA's (represented by imipramine and desipramine Fig. 3, although several other structurally related TCAs are available) are inhibitors of reuptake of NE and 5-HT back into the presynaptic neuron, which results in increased levels of these amines in the synaptic cleft; the tetra cyclic maprotiline is a NE reuptake inhibitor; buproprion and its metabolites inhibit reuptake of NE and 5-HT back into the presynaptic neuron, which results in increased levels of these amines in the synaptic cleft; the tetracyclic maprotiline (an approximate animal model of electroconvulsive therapy in humans) seems to cause an opposite effect on 5-HTT2 receptor regulation to that of several antidepressants (24, 25, for a review see Reference 26).

The 5-HT transporter protein (5-HTT) is critical for the reuptake of 5-HT into the presynaptic neuron. Decreased 5-HTT binding has been reported in depressed patients both in postmortem samples and in functional imaging studies (28, 29). Two common alleles of the gene that encode this protein have been identified, with the short (s) form of the allele being less active, which results in decreased transcription and reduced expression of 5-HTT. Studies have suggested that the s allele occurs more frequently in depressed patients as well as in suicide victims, and homozygous individuals are more likely to have family histories of depression. This finding seems paradoxical, as the s allele of the gene results in decreased serotonin reuptake, which thereby increases the duration of 5-HT in the synapse. It has been suggested that this observation may be explained by the lifelong duration of the genetic polymorphism compared with the acute effect of medication administration (28, 29).

Norepinephrine (Noradrenaline)

Investigations have also been conducted on the regulation of NE receptors in depression and after chronic administration of antidepressants (11, 30–32). In animal models, induced chronic stress has been used as a model for depression. Chronic stress antecedents (11, 30–32). In animal models, induced chronic stress has been shown to reverse antidepressant responses in individuals treated with antidepressants that inhibit NE reuptake, which suggests a direct role of this neurotransmitter in the course of depression (31).

Dopamine

Although the focus on monoamine neurotransmitters in depression has been on NE and 5-HT, it is generally considered that DA (Fig. 1) also plays a role, and readers are referred to several comprehensive papers that review basic science and clinical evidence for the involvement of DA (11, 33–35). Motor, psychomotor speed, concentration, and anhedonia have all been linked to dopaminergic circuits. Most DA-producing neurons have nuclei located in the brain stem. Projections from dopaminergic neurons form three primary paths to the cortical and subcortical structures: The nigrostriatal pathway is involved in motor planning and execution; the mesocortical pathway is concerned with concentration and executive functions; and the mesolimbic pathway is important in motivation, pleasure, and reward. Mesolimbic DA dysfunction is observed in animal models of depression, with subsequent antidepressant use causing enhanced dopamine signaling (11, 33–35). Decreased levels of DA in the nucleus accumbens of rats are associated with a decreased response to rewards. Animals that experience learned helplessness have decreased levels of DA in the caudate nucleus and nucleus accumbens; exposure to antidepressant medications has been reported to increase the level of DA in the same regions (11, 33, 34).

In humans, genetic studies have shown that particular polymorphisms of the D1 and D2 DA receptors are associated with depressed phenotypes. In studies of CSF concentrations of homovanillic acid (HVA), which is a metabolite of DA, lower levels than controls have been reported in some depressed patients. Conversely, depressed individuals with psychotic symptoms have been reported to have increased CSF levels of HVA and DA compared with controls (33).

Classes of Antidepressants

The three neurotransmitter monoamines mentioned above do not operate independently, but rather they interact with one another and are influenced by various central and peripheral biological processes. The actions of most commercially available antidepressant medications continue to target the monoamine neurotransmitter systems. The structures of these antidepressants are shown in Fig. 3. The MAO inhibitors (tranylcypromine, phenelzine, and moclobemide) prevent metabolic breakdown of the amines; the TCA's (represented by imipramine and desipramine Fig. 3, although several other structurally related TCAs are available) are inhibitors of reuptake of NE and 5-HT back into the presynaptic neuron, which results in increased levels of these amines in the synaptic cleft; the tetracyclic maprotiline is a NE reuptake inhibitor; buproprion and its metabolites inhibit reuptake of NE and 5-HT back into the presynaptic neuron, which results in increased levels of these amines in the synaptic cleft; the tetracyclic maprotiline (an approximate animal model of electroconvulsive therapy in humans) seems to cause an opposite effect on 5-HTT2 receptor regulation to that of several antidepressants (24, 25, for a review see Reference 26).

The 5-HT transporter protein (5-HTT) is critical for the reuptake of 5-HT into the presynaptic neuron. Decreased 5-HTT binding has been reported in depressed patients both in postmortem samples and in functional imaging studies (28, 29). Two common alleles of the gene that encode this protein have been identified, with the short (s) form of the allele being less active, which results in decreased transcription and reduced expression of 5-HTT. Studies have suggested that the s allele occurs more frequently in depressed patients as well as in suicide victims, and homozygous individuals are more likely to have family histories of depression. This finding seems paradoxical, as the s allele of the gene results in decreased serotonin reuptake, which thereby increases the duration of 5-HT in the synapse. It has been suggested that this observation may be explained by the lifelong duration of the genetic polymorphism compared with the acute effect of medication administration (28, 29).

Norepinephrine (Noradrenaline)

Investigations have also been conducted on the regulation of NE receptors in depression and after chronic administration of antidepressants (11, 30–32). In animal models, induced chronic stress has been used as a model for depression. Chronic stress antecedents (11, 30–32). In animal models, induced chronic stress has been shown to reverse antidepressant responses in individuals treated with antidepressants that inhibit NE reuptake, which suggests a direct role of this neurotransmitter in the course of depression (31).

Dopamine

Although the focus on monoamine neurotransmitters in depression has been on NE and 5-HT, it is generally considered that DA (Fig. 1) also plays a role, and readers are referred to several comprehensive papers that review basic science and clinical evidence for the involvement of DA (11, 33–35). Motor, psychomotor speed, concentration, and anhedonia have all been linked to dopaminergic circuits. Most DA-producing neurons have nuclei located in the brain stem. Projections from dopaminergic neurons form three primary paths to the cortical and subcortical structures: The nigrostriatal pathway is involved in motor planning and execution; the mesocortical pathway is concerned with concentration and executive functions; and the mesolimbic pathway is important in motivation, pleasure, and reward. Mesolimbic DA dysfunction is observed in animal models of depression, with subsequent antidepressant use causing enhanced dopamine signaling (11, 33–35). Decreased levels of DA in the nucleus accumbens of rats are associated with a decreased response to rewards. Animals that experience learned helplessness have decreased levels of DA in the caudate nucleus and nucleus accumbens; exposure to antidepressant medications has been reported to increase the level of DA in the same regions (11, 33, 34).

In humans, genetic studies have shown that particular polymorphisms of the D1 and D2 DA receptors are associated with depressed phenotypes. In studies of CSF concentrations of homovanillic acid (HVA), which is a metabolite of DA, lower levels than controls have been reported in some depressed patients. Conversely, depressed individuals with psychotic symptoms have been reported to have increased CSF levels of HVA and DA compared with controls (33).
Depression: Chemical Mechanisms of
phenelzine
tranylcypromine
CH
CH2
CHNH2
CH2CH2NHNH2
Cl CNHCH2CH2
O
NO
moclobemide
Desipramine (Desmethylimipramine)
imipramine
maprotiline
buproprion
fluoxetine
citalopram

Figure 3 Structures of antidepressants.

sufficient targets for research into the etiology and treatment of depression. In recent years, extensive research has focused on other possible causative factors, and the literature in this area will now be reviewed.

Beyond Biogenic Amines

\( \gamma \)-Aminobutyric Acid (GABA) and Glutamate

The amino acids GABA and glutamic acid (glutamate) (Figs. 4 and 5, respectively) are major inhibitory and excitatory neurotransmitters, respectively, in the central nervous system, and a requisite balance between the two operates in normal brain. Aberrations in the functions of one or both of these neurotransmitters have been implicated in the pathogenesis of several neurological and psychiatric disorders, which include depression (36–38).

Modulators of alpha-amino-3-hydroxy-5-methyl-4-isoxazolopropionic acid (AMPA) and N-methyl-D-aspartate (NMDA) ionotropic (associated directly with ion channels) glutamate receptors as well as metabotropic (linked to G-proteins and second messenger systems) glutamate receptors have been investigated for antidepressant properties (39–42). NMDA receptor antagonists have shown effects similar to other antidepressants in animal models of depression. In humans, plasma levels of glutamate have been reported to be correlated with the severity of depression (43), and NMDA receptor abnormalities have been observed in postmortem brain tissue of suicide victims and individuals with depression. Clinical trials of intravenously administered ketamine, which is an NMDA receptor antagonist, have produced rapid diminishment of depressive symptoms after a single administration (44). Concern has been raised over the psychotogenic side effects of NMDA receptor antagonists; however, memantine, which is a lower-affinity NMDA receptor antagonist without these side effects, has been investigated and did not show clinical improvement (42). Less evidence exists...
Depression, Chemical Mechanisms of

for the role of the AMPA receptor in depression. Tricyclic antidepressants have affinity for this receptor, and in a rat model of depression, AMPA receptor density increases with chronic antidepressant treatment (44).

Decreased GABAergic effects have also been associated with depression. In vivo evidence of GABAergic dysfunction in patients with depression includes decreased levels of GABA in the CSF, plasma, and occipital cortex (45, 46). Premenstrual dysphoric disorder (PMDD), which is a condition of depressive symptoms prior to menstruation, has been associated with a reduced variability in cortical GABA levels across the menstrual cycle (47). The MAO inhibitor antidepressant phenelzine has been shown to also cause marked increases in brain levels of GABA when administered to rats (48, 49). Using magnetic resonance spectroscopy, SSRI antidepressants have been demonstrated to increase brain levels of GABA in humans (50). Both GABA and glutamate are important beyond their direct effects as neurotransmitters. They interact with several other neurotransmitter and neuromodulatory systems and have effects on the regulation of the hypothalamic-pituitary-adrenal (HPA) axis, which may contribute to hyperactivity in this circuit in depression. The HPA axis will be discussed later in this article.

Neuroactive steroids

In recent years, a great deal of interest has focused on the possible involvement of several so-called neuroactive steroids (Fig. 6) in the etiology and pharmacotherapy of a variety of neurologic and psychiatric disorders, which include depression (51-53). These compounds are rapid-acting steroids that act as positive or negative allosteric modulators at many receptors, which include GABA-A and NMDA receptors. These steroids include pregnanolone, pregnenolone, allopregnanolone, isopregnanolone, 3α,5β-tetrahydroprogesterone (THP), 3β,5α-THP, 3α,5α-tetrahydrodeoxycorticosterone (THDOC), dehydroepiandrosterone (DHEA), DHEA sulfate, and pregnenolone sulfate. Allopregnanolone and THDOC are strong positive modulators at the GABA receptor and are potent anxiolytics (Note: Most antidepressants drugs also have anxiolytic properties). The ratios of these various neuroactive steroids are altered in several psychiatric disorders, and allopregnanolone seems to be of particular interest with regard to depression. Allopregnanolone has
been reported to produce antidepressant-like actions in ovariec-
tomized rats when given as an intra-accumbens infusion (54),
and SSR1 antidepressants have been shown to increase brain
levels of allopregnanolone in rats at doses lower than those
required to inhibit 5-HT reuptake (55). Blunted responses to al-
lopregnanolone in women with PMDD correlate with the degree
of depression in these subjects (56).

The Hypothalamic-Pituitary-Adrenal (HPA) Axis

A vast amount of literature indicates a relationship between
stress and depression, and in this regard it is interesting that the
HPA axis, which is involved in the coordination of neuroen-
docrine responses to stress, has been shown to be hyperactive
in many depressed patients (57, 58). In their diathesis-stress hy-
pothesis of mood disorders, Stout et al. (58) propose that the
HPA axis is the principal site where genetic and environmental
influences converge to cause mood disorders.

In healthy subjects, stress activates the hypothalamus, which
results in the release of corticotropin releasing factor (CRF),
which in turn results in release of adrenocorticotropic hormone
(ACTH). The ACTH activates the secretion of the glucocorti-
coid cortisol from the adrenal cortex (Fig. 7). The HPA axis has
an autoregulating mechanism mediated via negative feedback
by cortisol acting at glucocorticoid receptors in the hypothalamus
and the pituitary; but in many depressed patients, this regulation
is impaired, which results in higher circulating levels of CRF
and cortisol. This impairment probably occurs as a result of a
decrease in the number or sensitivity of glucocorticoid recep-
tors, which can be affected by gene expression, monoamines,
and early childhood expression (58, 59). It has been reported
that depressed patients exhibit decreased ACTH secretion in
response to exogenously administered CRF and increased se-
cretion of cortisol in response to a given ACTH level (58, 59).
Aripiprazole has been reported to result in return of HPA
functioning to control levels (59, 60), and it has been proposed
(9) that normalization of the HPA axis is a requirement for the
successful attenuation of depressive symptoms.

Chronic administration of corticosterone to rats results in be-
haviors associated with depression (61). Indeed, inhibitors of
cortisol synthesis (e.g., ketoconazole and metyrapone) and glu-
corticoid receptor antagonists (e.g., mifepristone) have been
tested clinically. Although preliminary studies with these drugs
promising, the former drugs have an unfavorable side effect
profile, and the latter suffer from lack of specificity (61, 62).

The development of CRF1 receptor antagonists has been a ma-
jor thrust by several research and pharmaceutical companies
in recent years. Increased levels of CRF in the hypothalamus
and CSF and an increased number of CRF neurons in the par-
ventricular nucleus of the hypothalamus have been observed
in major depressive disorder (10). Intracranial administration of
CRF to rats or overexpression of CRF in mice leads to sleep
disturbances, reduced appetite, and diminished sexual activity.
These symptoms appear in many depressed patients (58, 63).

Researchers continue to discuss whether the CRF antagonists
currently under investigation are acting primarily through block-
ade of prefrontal and limbic CRF1 receptors rather than having
an effect on the HPA axis. In addition, CRF2 receptors are ob-
served in the brain, and although their function is less clear,
drugs that act at these receptors are also of investigative inter-
est (58).

As indicated in other parts of this review, the HPA axis
interacts with most neurochemicals that are thought to be
important in depression, and many researchers consider that this
axis plays a central role in depression.

Substance P

The peptide Substance P acts primarily on neurokinin 1 (NK1)
receptors that are coupled to the Gs subunit of G proteins. Sub-
stance P has been of interest with regard to depression because
of its increased expression and that of NK1 receptors in fear-and
anxiety-related circuits. Substance P is released in animals in
response to fear-invoking stimuli and there is a high degree of
colocalization of substance P with 5-HT or its receptors in hu-
man brain (9). Kram et al. (64) reported that chronic treatment
of humans with a nonpeptide NK1 receptor antagonist resulted
in clinical improvement in depressed subjects; this finding was
replicated by some groups but not by others (9). It stimulated ad-
dditional research on NK1 antagonists. Although clinical findings
with substance antagonists have been disappointing overall, this
class of drug continues to be of interest, and recent studies in
laboratory animals and humans suggest that some useful anti-dep-
ressant agents may develop (66-69). Recent studies indicate
that NK1 antagonists act through serotonergic and noradrenen-
gic neurons (70), and it has been suggested that these drugs may
be useful agents in combination with traditional antidepressants
(71).

Cytokines

In recent years, researchers have shown interest in the role
of the immune system in depression (72, 73), and the cy-
tokine theory of depression proposes that alterations in the
immune response result in behavioral, cognitive, and neuroen-
docrine changes in depression. Cytokines can be proinflam-
matory or anti-inflammatory, and it has been postulated that
depression may be caused by, or at least related to, excessive
Depression, Chemical Mechanisms of

amounts of proinflammatory cytokines such as interleukin-1 and tumor necrosis factor-alpha (74). Indeed, it has been observed that antiviral treatment with the cytokine interferon results in “sickness behavior” characterized by symptoms such as weight loss, anorexia, depressed mood, sleep disturbances, social withdrawal, and fatigue, which are also observed in depression (9). It is also of interest that cytokines are potent stimulators of the HPA axis via activation of CRF release, and proinflammatory cytokines can also decrease 5-HT levels by increasing 5-HT secretion and diverting metabolism of tryptophan metabolism away from formation of 5-HT by tryptophan hydroxylase as well as increasing metabolism by the indolamine-2,3-dioxygenase (IDO) pathway (75). Although the evidence for abnormal cytokine levels and/or inflammatory responses in depression is inconsistent to date, this area is very interesting for future exploration and should lead to more studies on glial cells that synthesize and release cytokines.

Intracellular signaling cascades and neurotrophic factors

As the acute increase in synaptic monoamines produced by antidepressant drugs does not correlate with timing of clinical effects, the intracellular signaling pathways that their receptors interact with have been investigated as targets for novel antidepressants. Several such pathways are observed, and an example is given in Fig. 8 (76).

Several serotonergic and DA receptors and the β₁ noradrenergic receptor activate the Gs pathway, whereas others activate the Gq pathway. These signaling cascades result in activation of protein kinases A and C, with a net effect of increasing phosphorylation of cyclic adenosine monophosphate (cAMP)-regulated element binding protein (CREB). CREB acts as a regulator in the expression of many genes, some of which have been implicated in neuroplasticity and neurogenesis. For example, transcription of brain-derived neurotrophic factor (BDNF), tyrosine kinase B (trkB) (which is a BDNF receptor), and the glucocorticoid receptor are activated by CREB. Inhibition of transcription of CRF and subunits of the NMDA receptor occurs in the presence of CREB. Alterations in CREB levels have been observed in depressed individuals, and increases in CREB function have been associated with reduction or production of depressive symptoms in animals, which depends on the area of the brain involved (77, 78).

The most researched area of interest for antidepressant activity related to signaling pathways is the hippocampus (79–85). The volume of the hippocampus is reduced in patients with multiple episodes of major depression, as observed in imaging studies and in postmortem samples (81–84). Animal models of inescapable stress are associated with decreased hippocampal neurogenesis (84). Neurogenesis is also evident in the hippocampus of humans (85). The neurogenesis hypothesis states that depression is a consequence of impaired...
neurogenesis in the hippocampus and that antidepressants exert their effect by stimulating neurogenesis (86). Controversy exists as to whether neurogenesis is necessary for the efficacy of antidepressant medications in animal models (87), with some evidence suggesting that when neurogenesis is inhibited, antidepressant effects are lost (84, 88).

In the hippocampus, CREB seems to mediate antidepressant effects, and a variety of antidepressants, which includes electroconvulsive therapy, increases CREB expression (77). Experimentally increasing the expression of CREB in the hippocampi of depressed rats seems to have an antidepressant effect. Evidence suggests that the role of CREB in mediating antidepressant effects of medication is related to the expression of CREB-regulated neural growth factors, such as BDNF (79, 89). In humans, BDNF levels are reduced in postmortem samples of depressed individuals when compared with nondepressed controls (90). A antidepressant therapy increases serum BDNF levels in depressed humans (91). In animals, chronic stress models are associated with reduced expression of BDNF (79) and decreased cell proliferation in the hippocampus (83), which can be prevented or reversed by treatment with antidepressant drugs. However, researchers debate whether genetic disruption of the signaling pathways that involve BDNF and trkB causes depressive behavior (See Reference 92 for a review). It is also of interest that chronic administration of antidepressants has been reported to upregulate expression and activity of the neuroprotective enzyme superoxide dismutase and, depending on the dose and antidepressant, to increase immunostaining of BDNF and/or the antiapoptotic protein Bcl-2 (93, 94).

The picture with CREB is far from straightforward. Increased expression of CREB in the nucleus accumbens of rats produces depression-like effects, which include anhedonia (77, 89) and increased helplessness, in learned-helplessness models (81). Increases in BDNF expression in the mesolimbic DA system of rats are associated with induction of depressive effects in certain animal models. However, studies that use systemic administration of BDNF and activators of the cAMP-CREB cascade seem to have a net therapeutic effect in behavioral models of depression (90, 86, 89, 95).

**Metatonin (N-acetyl-5-methoxytryptamine)**

Depression and seasonal affective disorder, which is a form of depressive illness characterized by symptoms with onset and course related to season of the year, have been noted to be cyclic and possibly related to alterations in circadian rhythms (96). Alterations in the circadian functioning of the HPA axis and the hypothalamic-pituitary-thyroid axis occur in some patients with major depression (96). Disruption in the sleep/wake cycle and insomnia are common in depression. Melatonin (Fig. 9), which is a hormone derived centrally in the pineal gland from serotonin, is instrumental in many biological functions with circadian variability. Several studies have shown that depressed individuals have a deficiency of melatonin secretion, but a body of evidence shows increased nocturnal melatonin production in the brains of depressed subjects with increased urinary excretion of its metabolite, 6-hydroxymelatonin sulfate (96, 97). This finding may reflect differences in subtypes of depression. As melatonin is secreted in a rhythmic fashion, the timing of the melatonin peak has been studied in depressed individuals. Phase shifts toward both earlier and later onset of peak secretion have been noted. Both treatment with antidepressant medications and electroconvulsive therapy have been associated with increases in melatonin excretion in depressed individuals. The pineal gland receives input from noradrenergic neurons, and it has been suggested that alterations in melatonin secretion merely reflect dysfunction in monoamine systems (96, 97). Agomelatine, which is a melatonin receptor agonist and 5HT2c receptor antagonist, has been shown to have antidepressant properties in animals and humans (98-101). Agomelatine targets the M1/M2 melatonin receptors and mimics melatonin in its effect on circadian rhythms. A combination of activity at melatonin receptors and monoaminergic receptors may represent a novel method of antidepressant drug action (97-100).

**Other Antidepressant Approaches and Targets**

Electroconvulsive therapy (ECT) is considered to be the most efficacious treatment for depression, although its side-effect profile limits its use, and it is associated with a high relapse rate (10, 102). The exact mechanism of action of ECT is still unknown, although effects on monoaminergic systems, CRF, neurotrophic factors, and neuroendocrine systems have been suggested (102). Because DA probably plays a role in the pathophysiology of depression, triple reuptake inhibitors that inhibit reuptake of NE, 5-HT, and DA have now been developed and are undergoing preclinical and clinical testing (10). Numerous reports in the literature suggest that trace amines (so-named because their absolute levels in the brain are much lower that those of the classic neurotransmitter amines) are involved in the etiology and pharmacotherapy of several neurologic and psychiatric disorders, which include depression (103-106). Researchers have shown an increased interest in these trace amines, which include β-phenylethylamine, tyramine, tryptamine, and octopamine, in recent years with the discovery of a family of G-protein coupled receptors that bind to and are activated by these amines (107, 108). Epinephrine is present in much lower concentrations than NE and DA in the brain, but it is a major stress hormone in the periphery and could be a contributing factor in depression (109). Histamine is present in lower concentrations in brain than NE, DA, or 5-HT as well, but it has also been implicated in the etiology of depression (110).
Reports in the literature indicate that atypical (second generation) antipsychotics are useful antidepressant agents when combined with standard antidepressants (10); several researchers suggest that in some cases they and/or their metabolites may be useful antidepressants in their own right. It is interesting that some of these atypical antipsychotics have been reported to have several neuroprotective actions, which include effects on neuroprotective enzymes and factors that affect apoptosis (programmed cell death) (111). Preliminary reports suggest similar effects of antidepressants (93, 94). Some other areas related to the chemistry of depression and its treatment that contain conflicting reports in the literature but are of interest include the following: the relationship of serum cholesterol and the use of statins (cholesterol-lowering drugs) to depression, anxiety, aggression, and suicide (112, 113); omega-3 fatty acids and mood disorders (114–118); the use of herbal products such as St. John’s wort extracts and 5-adenosine/methionine as antidepressants (119–123); and the possible involvement of CB1 cannabinoid receptors, galanin, Neuropeptide Y, histone deacetylases, and tissue plasminogen activator in depression (See Reference 9 for a review). Other reported or putative neurobiological antidepressant treatments that we have not mentioned previously in this review include deep brain stimulation, vagus nerve stimulation, and transcranial magnetic stimulation; the reader is referred to the review of Holtzheimer and Nemeroff (10) and the references contained therein for details.

Concluding Remarks

Although a great deal has been learned about brain function in the search for newer antidepressants and some very interesting potential drug targets have been identified, our knowledge of the causes of depression remains inadequate. We still lack antidepressant drugs that are sufficiently rapid acting and effective in a large enough percentage of depressed patients. Problems that continue to make studies on depression difficult include the following: the heterogeneous nature of depression itself, gender issues, the involvement of multiple interacting neurotransmitters and neuromodulators in the etiology of depression, the inadequacy of current animal models, the involvement of various brain regions in the symptomatology of depression, disagreements about the neurogenesis hypothesis of depression and about the relative importance of the hippocampus in depression, regional differences in the effects of transcription factors such as CREB, and the differing effects of some antidepressants on the HPA axis.

Acknowledgments

Research findings provided by the Canadian Institutes of Health Research (CIHR), the Canada Research Chairs and Canada Foundation for Innovation programs, and the Davey Endowment. The assistance of Ms. Tara Checkrata in preparing this manuscript is gratefully acknowledged.

References


Depression: Chemical Mechanisms of


Further Reading
Depression, Chemical Mechanisms of


Schmidt HD, Duman RS. The role of neurotrophic factors in adult hippocampal neurogenesis, antidepressant treatments and animal models of depressive-like behavior. Behav. Pharmacol. 2007;18:391-418.


Hemozoin: A Paradigm for Biominerals in Disease

Melissa D. Carter, Anh N. Hoang and David W. Wright, Vanderbilt University, Nashville, Tennessee
doi: 10.1002/9780470048672.wecb649

Biomineralization is the formation of organic–inorganic composites by organisms. Originally evolved as a protective mechanism, this complex process has also become a recognized contributor to several disease states, which range from kidney stone disease (nephrolithiasis) to parasitic diseases like malaria. The characteristic three-step process for the formation of biominerals is defined by the supramolecular preorganization of a nucleating template, the interfacial molecular recognition of crystal nuclei and the cellular processing of resultant aggregates. Hemozoin formed in the heme detoxification pathway used by the malarial parasite *Plasmodium falciparum* represents a paradigm for pathogenic biominerals. Current research indicates that a supramolecular lipid template organizes heme released previously during hemoglobin catabolism. Nucleation and growth of the heme aggregate serves to protect the parasite from the toxic effects of free heme. Given the mechanisms of biomineralization, it is not surprising to discover that century-old antimalarial compounds function by disrupting key interactions between the heme substrate and template. Subsequently, the heme-aggregate is released into the host vasculature and deposits in patients’ brains, spleens, and livers where it disrupts host innate immune response. The underlying basis of this immunomodulating activity seems to result from hemozoin mediated lipid peroxidation. Understanding the relationships between hemozoin formation and its pathogenic activity with the host immune response represents a significant challenge to chemical biology.

Biomineralization is the biologic formation of organic–inorganic composites generally organized on a nucleating template of organic material. During bone and teeth formation, hydroxyapatite, which is a calcium phosphate derivative enriched with carbonate, mineralizes on a template of collagen fibrils and other proteins (1). Likewise, calcium oxalate monohydrate crystals mineralize on a biologic matrix of proteins in both plants and humans. Although these crystals function as tissue support in plants (2), they cause painful inflammation in humans (3). In fact, biomineralization is associated with several human diseases. Nephrolithiasis is the mineralization of calcium oxalate in the urinary tract of patients as kidney stones. In gout or metabolic arthritis, monosodium urate crystals accumulate in the articular joints. A patient’s innate immune response or the on-rush of monocytes and neutrophils at the sites of crystal sedimentation causes the characteristic tissue inflammation and pain associated with these diseases. Biomineralization can also contribute to the pathology of parasitic diseases like malaria and schistosomiasis. When the malarial protozoa *Plasmodium falciparum* deposits heme-derived aggregates in host vasculature, these aggregates travel through the bloodstream and collect in the brain, spleen, and liver. When they are phagocytosed by innate immune cells such as monocytes and neutrophils, the cells’ ability to produce reactive oxygen and nitrogen species is impaired. This immunomodulation is typical of the pathogenesis caused by many biomineral-associated diseases.

Heme Homeostasis

Over 40% of the world’s population is at risk from malaria. The disease causes severe illness in over 500 million people and results in over 1.7 million deaths each year (4). Transmission is most prevalent in the world’s poorest countries, predominantly sub-Saharan Africa, and accounts for 40% of public healthcare costs. In addition, a 2007 World Health Organization report estimates a 1.3% decline in annual economic growth for countries.
with high rates of malarial infection (4). Compounded over time, this drop contributes significantly to GDP disparities between those countries where malarial infection is endemic and those where it is not.

Challenges of heme homeostasis

Over 100 Plasmodium species contribute to the spread of malaria, but only four of these (P. falciparum, P. vivax, P. ovale, and P. malariae) account for human infection, the deadliest being P. falciparum. The malaria life cycle exists first in a mosquito, and then it passes to a human host. An infected female Anopheles mosquito is the host of the parasite’s sporogonic life cycle. Mature P. falciparum sporozoites reach the salivary glands of the mosquito, and the parasite is transmitted to a human host when the mosquito feeds. During this blood meal, sporozoites are released into the bloodstream where they penetrate hepatic cells and mature into schizonts. The liver cells rupture after approximately two weeks, discharging merozoites into the bloodstream whereupon they infect red blood cells (RBCs). Every 48 to 78 hours, mature merozoites rupture from spent RBCs and either they differentiate into gametocytes or they infect more RBCs. This blood stage is responsible for the clinical manifestation of the disease (5).

P. falciparum ingest and degrade up to 80% of host erythrocyte hemoglobin (Hb) (5) to provide the parasite with essential amino acids for growth and maturation. During high parasitemia (20%), up to 100 g of the 750 g of circulating host Hb can be catabolized by the parasite (5). Hb is broken down in the parasite’s acidic (pH 4.5–5.2) digestive food vacuole (DV) by a suite of proteases that includes four aspartic acid proteases (plasmepsins PfPM1, PfPM2, PfHAP, and PfPM4), three cysteine proteases (falcipains PfFP2, PfFP2′, and PfFP3), and a metalloproteinase (falcilysin) (Fig. 1) (6). This ordered catabolic process is initiated by PfPM1 and PfPM2, which cleave between residues ω33Phe and ω34Leu in the hinge region of Hb. Recent quadruple knockout studies show individual plasmepsin redundancy and suggest that although each aspartic proteinase may not be essential to the intraerythrocytic stage, they could play unique roles outside Hb digestion. On plasmepsin cleavage, the protein unfolds, which releases free heme and exposes

![Diagram of hemoglobin catabolism](image)

Figure 1 Ordered pathway of hemoglobin catabolism. Host Hb cleavage is initiated by PfPM1 and PfPM2. Hb unfolds, releases heme and exposes other peptide bonds to the falcipains. Resultant protein fragments are degraded even more by falcilysin and transported to the cytosol where they are broken down into needed amino acids. The heme released in the first step aggregates via biomineralization to form HZ.
The processes of crystal nucleation and growth are driven by the chemistry of biomineral formation and unicellular silver deposits. The formation of these biominerals is recognized commonly in a range of functional structures including silicon, into crystal aggregates evolved as a protective mechanism for inorganic matrices such as protein and lipid networks. The propagation and continued growth of a stable crystal phase is the foundation for the second-order assembly of inorganic species.

### Biomineralization

Over 570 million years, the complex process of organizing inorganic molecules, which generally contain calcium, iron, or silicon, into crystal aggregates evolved as a protective mechanism for organisms. Fossil records show that biomineralization surfaced first in the form of organism scales and skeletons during the neoproterozoic era (9), and now, biominerals are recognized commonly in a range of functional structures including shells (10), vertebrate teeth, and bone (11, 12), diatom silicates (13) and unicellular silver deposits (14). The formation of these materials is attributed to two processes: 1) biologically-induced mineralization, which is the deposition of minerals via adherent precipitation (15) and 2) biologically-controlled mineralization, which is the regulated formation of minerals that have a specific function and structure (16). Generally, induced minerals are heterogeneous in character, whereas minerals that result from cellulyarly controlled processes exhibit uniform composition and morphology (17).

### Chemistry of biomineral formation

The processes of crystal nucleation and growth are driven by the basic laws of thermodynamics in which a greater free energy must exist in the original solution phase than the resultant crystalline phase (17). However, it must be noted that the free energy distribution at the crystal surface differs from these phases (3). Because molecules on the crystal surface are not bound as strongly as molecules in the preliminary bulk solution, their free energy contribution to the system is greater. This energy difference between molecules at the surface and in solution is known as the interfacial free energy. Acting on the destabilization of crystal nuclei, the interfacial free energy can cause either nucleus dissolution or growth of the nucleus to a large enough size that its stability prevails over the affects of surface free energy, and a crystal is formed. The nucleation pathways that lead to the most stable crystal phase are the foundations of the Ostwald-Lussac law of phases, which explains that nucleation occurs in a series of pathways with increasing stability before reaching the final crystalline state (3). The propagation and continued growth of a stable crystal is attributable to a uniform nucleation template from which a new phase is formed from an old phase that has become higher in free energy (3, 18).

In the context of crystal nucleation and growth, the formation of biominerals can be cast as a three-step process that involves the supramolecular preorganization of a template, the interfacial molecular recognition of crystal nuclei, and the cellular processing of resultant aggregates (19). The first-order assembly of organic matrices such as protein and lipid networks provides a foundation for the second-order assembly of inorganic species. Typically, these frameworks have functionalized surfaces that behave as templates for inorganic nucleation and are governed by electrostatic, structural, and sterochemical complementarity of the organic-inorganic interface. Without cellular intervention to control the flux of metal ions, crystal nucleation would continue to grow along these scaffolds, proceeding to their bulk state. Clearly, such unconstrained growth represents a danger to cellular integrity. The final stage of biomineral construction is cellular processing, which is often the distinguishing step between native biomineral morphology and that of its synthetic analogs. The intracellular or extracellular environment in which a crystal grows ultimately influences its crystallographic structure and morphology (19) and ensures the function of laden cells.

**Search for hemozoin's bionucleating template**

Investigations into the mechanisms of HZ formation have centered on protein or lipid-rich nucleating templates. Early hypotheses of a catalytic heme-polymerase in trophozoite lysates were abandoned because of failed attempts at identification and purification of the enzymatic activity. An alternative to such a heme polymerase was proposed by Sullivan (20) from their investigations of a family of histidine-rich proteins (HRP) isolated from the DVs of *P. falciparum*. HRPII is a 30 kDa protein with 76% of its composition being His or Ala residues. HRPIII is 27 kDa and is 56% His or Ala residues. Both proteins have repeats of the tripeptide Ala-His-His, 51 repeats in HRPII, and 28 repeats in HRPIII. This Ala-His-His recurring domain provides an archetypal biomineralization scaffold for the nucleation and propagation of free heme into HZ. When HRPII or HRPIII were...
templates bound 17 molecules of heme, mediated the formation of *P. falciparum* nucleating domain of HRP II of bionucleating templates (BNT I and II) were composed of the stochiometric amounts of Fe(III)PPIX, although substrate specificity experiments suggested that substrate recognition was dependent on the porphyrin moiety rather than specific metal recognition. Moreover, HRP substrate recognition was not shown to be mediated by histidine axial ligation to a metal ion, but rather it was attributed to stacking and electrostatic interactions. Chloroquine inhibition of the bionucleating templates was comparable with HRP II and III impaired formation of HZ when treated with the antimalarial (21).

These templates, like HRP II, were shown to promote HZ formation at a parasite DV relevant pH 4.0-4.5. In fact, HRP II is active from pH 4.5 to pH 6.0. HZ formation slows at pH values below 2.0 and above 5.0 (22). Although an increase in pH does cause an increase in binding of heme to HRP II, the HZ produced actually decreases leading, instead, to the formation of μ-β-β-heme dimers (22). Therefore, the bis-histidyl heme binding observed on other proteins like histidine-rich glycoprotein (HRG) at physiological pH 7.0 differs from the HRPII nucleation of HZ in the parasite's acidic DV (23), although a pH 7 HRP II model was attempted by Schneider and Marletta in 2005 (24). At this pH, binding is consistent with a nucleating template that serves as an organizing function rather than a tight heme-binding function.

Subsequent experiments revealed, however, that the HRP's were not the likely template for HZ formation. Double deletion mutants of HRP II and HRP III did not prevent the formation of HZ in the DV (20), which suggests the existence of an alternative template. In addition, HRP II is not located solely in the DV, but rather it is secreted into the serum of victims at high concentrations. Histology labeling experiments show that HRPs in the DV are simply captured during the endocytosis of host hemoglobin and not specifically targeted to the DV (20). Such surreptitious colocalization would also suggest the existence of a non-HRP template. In light of these experimental results, the search for HZ's biomineralization template turned toward other possibilities.

Lipids are a possible template for HZ biomineralization. As a bionucleating template, lipids can increase the solubility of Fe(III)PPIX, localize high concentrations of Fe(III)PPIX, and provide a layer of free heme intercalation. The propionate groups of free heme within the lipid layer can then effectively charged Fe(III) centers, which weakens any hydrogen bonds with water and enables the hydrogen bonds of Fe(III)PPIX dimers. Fitch et al. (25) showed that HZ formation could be mediated at pH 5.0 in the presence of fatty acids like arachidonic acid and glyceral of oleic acid as well as detergents like polyoxyethylene sorbitan monolulate (TWEEN 80). Increasing support for a lipid template led to the membrane sacrifice theory by Hemplemann et al. (25), which suggested that inner membranes of RBC transport vesicles in the parasite cytosome were degraded by free heme, releasing lipids that increase heme solubility and aggregation. In the sacrifice of the inner membrane, the outer membrane is preserved to prevent additional oxidative damage to the parasite (26). In vitro lipid-initiated HZ crystallization was reported by Egan et al. (27) using a range of lipids that includes myristoyl, detyl, and palmitoyl glycerides; phospholipid; and cholesterol to initiate HZ formation along the lipid–water interface. Simulations of HZ formation in these studies indicated that the hydrophobic Fe(III)PPIX dimer was more stable in the lipid layer to foster hydrogen bonds between the protonated propionic acid groups and thus, HZ crystal assembly (27). Transmission electron microscopy images of HZ localized in lipid nanospheres within an infected trophozoite stage RBC provided increased support of a lipid-mediated HZ biomineralization process. The extraction of these fatty acyl glycerides (including monostearic, monopalmitic, dipalmitic, diloetic, and diolein glycerides) and their incubation with substrate heme, developed a competent in vitro template for HZ formation (28).

**Target of antimalarials**

Interruption of the parasite heme biomineralization pathway is a logical target of antimalarial drug development. As antimalarials amass in the parasite's DV, HZ formation is inhibited, and the parasite becomes flooded in toxic heme. Beginning with the hypothesis that neutral lipid droplets indeed serve as biomineralization templates for the formation of HZ, a variety of limiting cases exist in which inhibitors may disrupt the aggregation of HZ (Fig. 2). An inhibitor may bind the heme substrate in such a manner that the heme-inhibitor complex cannot be recognized by the template. Alternatively, a drug could interact with the template, blocking the heme binding site. Finally, a HZ aggregation inhibitor might trap the heme bound to the template, which prevents formation of the dimeric unit or nucleation of the extended crystal. These possible modalities, which are derived from the paradigm of biomimetic formation, can be used to understand the mechanism of action of some antimalarial compounds.

Well-known quinoline-based drugs like chloroquine and amodiaquine are thought to target this HZ biomineralization pathway. These drugs have been the standards in treatment of malaria, but growing parasite resistance threatens their use. Quinoline-based drugs include the well-known 4-aminoquinoline derivatives chloroquine, amodiaquine, halofantrine, quinine, and bayspinone (29). These drugs are thought to trap monomeric heme or bind the μ-β-β-dimer of oxidized heme to prevent HZ formation. The μ-β-β interactions at the (001) heme face control resultant adduct formation. Computational models of HZ and quinoline interactions indicated that drug adsorption occurred on the (001) and (011) crystal faces (Fig. 3) (30). These highly symmetrical crystals also showed tapering at each drug-bound (001) or (011) ledge, which suggested weakening of the quinoline inhibition along the crystal's c-axis which resulted in thinner crystal cross-sections (30).

Other inhibitors of HZ formation are thought to act similarly to the quinoline family. Binding of the antifungal azole-based drugs clotrimazole, ketoconazole, and miconazole to heme is thought to damage parasite cell membranes and cause...
Figure 2: Modes of hemozoin inhibition. On a neutral lipid droplet template (T), heme can aggregate to form the biomineral HZ. Antimalarials may inhibit this aggregation by binding heme substrate, interacting with the lipid template or trapping heme bound to the template. All actions serve to prevent the formation of HZ.

Figure 3: Hemozoin-chloroquine binding model. Chloroquine is thought to bind to the (001) and (011) faces of heme as well as π–π stacking at the (010) face. To prevent HZ formation completely, chloroquine would have to bind both faces of each heme monomer or crystal extension would be blocked in only one direction as shown.
The isonitrile terpenes (diisocyanoadociane and axisonitrile-2), isolated from marine sponges and the synthetically derived methylene blue analogs (acures A, B, and C; thion; celestine blue; and phenosaphranin) most likely prohibit HZ formation by binding monomeric heme. The xanthone family of animalarals, such as the hydroxynamethanes and the bis-(N,N-dibutylamino)ethoxy xanthones, were found also to bind at the HZ crystal faces (001) and (031). Specifically, the drugs’ terminal amino groups bind the carboxyl group exposed at each face, which inhibits the nucleation and growth of the crystal (30). The crystal engineering prospect of future drug design is an interesting one that beckons additional exploration and promises applications in a variety of biomimetic-associated diseases.

Hemozoin Characterization

Prior to the definitive X-ray powder diffraction characterization of HZ, the true structure of the aggregate remained elusive because of its limited solubility. HZ was soluble only in NaOH, which completely degraded the structure and prevented any attempt to determine the intramolecular atomic interactions. The insolubility of the product rendered useless many “standard” experimental methods for the characterization of bioorganic systems. This aspect of the biomimetic’s identification undoubtedly contributed to Ridley’s description of HZ as “a black insoluble mass of material that can be soul-destroying to work with” (31).

Physical characterization

Studies by Slater et al. (31) provided an initial “fingerprint” of HZ’s chemical structure by applying Fourier-transform IR (FT-IR) spectroscopy to intact HZ crystals. The IR spectrum of HZ revealed intense absorbance patterns at 1664 and 1539 cm⁻¹, which indicate the axial propionate C–O and C–O stretching, respectively (Fig. 4a). These peaks were absent from the spectra of the synthetic substrates hemin chloride and hematin. These data suggested a direct coordination from the spectra of the synthetic substrates hemin chloride and hematin. The IR spectroscopy of intact HZ crystals. The IR spectrum of HZ revealed intense absorbance patterns at 1664 and 1539 cm⁻¹, which indicate the axial propionate C–O and C–O stretching, respectively (Fig. 4a). These peaks were absent from the spectra of the synthetic substrates hemin chloride and hematin. These data suggested a direct coordination from

Identification of native lipid components of hemozoin

In vivo native HZ transmission electron microscopy images from trophozoite-infected RBCs depicted the localization of HZ crystals within neutral lipid nanospheres (28). Trophozoite DVs were isolated using Percoll/sucrose bottom separation techniques, and the lipid content of these isolates was extracted by Bligh-Dyer (chloroform/methanol) lipid extraction. Methylenediacetic acid characterization by gas chromatography-MS (GC-MS) and lithium (Li⁺) adduct electrospray ionization mass spectrometry of trophozoite fractions revealed a suite of neutral lipids adhered to the crystal’s surface. When these lipid extracts were separated using reverse phase-, normal phase-, and chiral phase high performance liquid chromatography and sub-sequent analysis revealed the presence of substrate hemin, HZ formation was observed (28). This competent nucleating template of fatty acyl glycerides is consistent with the hypothesis of a lipid scaffold for heme aggregation in P. falciparum. Also extracted from native HZ were polar hydroxylated fatty acids derived presumably from cellular arachidonic and linoleic acids. Native HZ was purified from infected RBCs by a series of centrifugation steps followed by organic extraction of its lipid coat. Analysis of the lipid coat revealed the presence of hydroxylated polysaturated fatty acids. These polar lipids were separated using reverse phase-, normal phase-, and chiral phase high performance liquid chromatography and subsequent analysis revealed the presence of 15-, 12-, 11-, 9-, and 5-hydroxyeicosatetraenoic acids (HETEs) as well as 13- and 9-hydroxyoctadecadienoic acids (HODEs) (35).
Chemical reactivity

Native HZ produced in the parasite DV most likely encounters cellular debris such as arachidonic and linoleic fatty acids or other lipids when released from a bursting RBC. Redox cycling of surface exposed heme units within HZ can lead to the initiation of lipid peroxidation (LPO). A abstraction of a bisallylic hydrogen atom from a fatty acid such as arachidonic acid results in an unpaired electron on the methylene carbon. This unpaired electron promotes the rearrangement of the double bonds adjacent to the methylene group that produces an alkyl radical (L•). In the presence of oxygen, alkyl radicals react to form peroxy radicals (LOO•) that can abstract a hydrogen atom that yields a lipid peroxide (LOOH) and can propagate more reactions. Reduction of the peroxy radical would result in the production of racemic mixtures of hydroxylated fatty acids like the HETEs and HODEs described previously in the HZ lipid coat. Secondary oxidation and chain \( \beta \)-cleavage results in reactive 4-hydroxy-2-nonenal (HNE) (Fig. 5). Lipid peroxidation was demonstrated in vitro by reacting arachidonic acid with native HZ purified of all cellular lipid and protein content. Resultant HETE products were extracted from the reaction supernatant and identified using reverse phase high performance liquid chromatography (37). Recent studies of BH reactions with arachidonic acid revealed an identical reaction profile to that of HZ with all six positional HETE isomers identified by ultra high pressure reverse phase and chiral phase liquid chromatography tandem MS/MS (chiral-RP-LC-MS/MS) (38). LC-MS analysis has enabled the separation and identification of these cellular metabolites at femtomole levels of detection. The presence of these oxidation products in the lipid coat of native HZ and their formation upon arachidonic acid incubation with purified HZ and BH strongly supports the HZ-mediated LPO of cellular fatty acids. Additionally, the known immunomodulatory activity of these lipid peroxidation products is intriguing given the reported ability of HZ to disrupt the function of innate immune cells.

Basis of Hemozoin

Immunomodulation

Once thought to be a simple “inert” detoxification biomineral, an increasing appreciation exists of HZ’s reactivity and subsequent
Hemozoin: A Paradigm for Biominerals in Disease

Figure 5  Hemozoin-mediated lipid peroxidation. Redox cycling of iron in complexes like HZ can initiate lipid peroxidation of fatty acids like arachidonic acid (15-HETE). Abstraction of the bisallylic hydrogen leaves an unpaired electron on the methylene carbon that can rearrange to form a reactive alkyl radical (L•). Oxidation of this radical leads to a peroxyl radical (LOO•), and on reduction, forms a lipid peroxide (LOOH) that can undergo additional reactions to yield a variety of secondary oxidation products.

perturbation of the host immune response. HZ reactivity in the modulation of innate and adaptive immunity has been attributed to a range of effects that include toll-like receptor 9 (TLR9) activation, cytokine production, LPO, and dendritic cell development. Recent studies found that the P. falciparum DNA that remained on native HZ extracts was the true source of TLR9 activation, and once native HZ was treated with nuclease, no activation was observed (39). HZ disruption of dendritic cell function was shown to cause a decline in T cell activation that led to a weakened adaptive immune response. The role of HZ as an immunoreactive aggregate may impact the high rates of patient secondary infection as well as the decline in vaccine efficacy.

Functionally, phagocytosis of HZ has been reported to impair macrophage oxidative burst, to downregulate iNOS activation, and to perturb cytokine profiles in infected patients. It has also been shown to correlate with increased levels of immunomodulatory LPO products such as prostaglandin E2 (PGE2), HNE, and HETEs (40) in monocytes. The biologic activity of these compounds is generally derived from either of two mechanisms. In the first mode of action, the reactive intermediates or products may form adducts to DNA or proteins. Thus, a variety of
chain-terminating reactions of peroxy, alkyl, or epoxyperoxyl radicals can result in oxidative cross-links to DNA or proteins. Furthermore, the electrophilic alkenals, such as HNE, readily form Schiff-base adducts to lysine residues and Michael addition adducts to histidine and cysteine residues to perturb the function of many proteins (41). In the second mode of action, the LPO products may act as alternate ligands to sever-eral different proteins and receptors to initiate an ultimately pathogenic signaling cascade. The hydroxylated fatty acids 9- and 13-HODE, as well as 15-HETE have all been found to be activators of the important nuclear receptor protein PPAR-γ, which is involved in key cellular regulatory and differentiation functions in monocytes (42). 15-HETE has also been found to stimulate RBC adherence to capillary endothelia, to enhance vascular permeability and edema, and to induce chemotaxis and chemokinesis, although the precise pathways remain unclear. From these examples, it is easy to imagine how such promis-cuous reactivity often manifests itself in the pathogenesis of disease states. These findings suggest that HZ’s ability to rec-tify cellular responses may actually be caused by formation of primary and secondary LPO products (40).

To study this hypothesis, BH was incubated with ghost RBC membrane lipids, which are similar to those that the native biomineral would be exposed to upon RBC rupture (40). RAW macrophage-like cells were treated with the supernatant of this reaction, and reactive oxygen and nitrogen species (ROS and RNS) produced via the NADPH oxidase and iNOS pathways were inhibited. The levels of inhibition paralleled the effects of individual LPO products such as 15-HETE or HNE at patologically reported concentrations. Additionally, treatment of cells with either BH or unreacted ghost supernatant did not result in a decrease in ROS and RNS production, which indicates that the products of HZ-mediated LPO were responsible for the observed disruption of macrophage function, not the dimeric home component of the aggregate itself (40). Mechanisms for such material host-pathogen interactions and, more broadly, biomineral-to-cell relationships are primarily undefined, which creates a significant treatment barrier. Unraveling the formation and role of HZ in the infection may provide additional insight into the prevention and treatment of a variety of diseases that result from pathogenic biominerals.

Summary

Biomineralization results in an expansive array of complex ma-terials. These natural biominerals often represent unique crystal forms that extend over several size domains that are synthesized under ambient conditions. Increasingly, it is understood that biomineralization processes play important roles in the patholo-gies of several diseases. In malaria, the parasite forms the HZ biomineral in response to the heme released during hemoglobin catabolism. HZ serves an important detoxification role that al-lows the organism to maintain homeostasis during its intraery-throcytic phases. Despite the fact that HZ represents a validated drug target for P. falciparum, many fundamental questions re-main concerning its formation; questions such as how or if the parasite assembles neutral lipid droplets specifically, how the home is transported and deposited in these lipid domains, and whether new strategies exist that could be used to design drugs to disrupt this process. Tackling the rational drug design prob-lem from a crystal engineering perspective offers an enticing direction for drug discovery, whether designing crystal-specific antibodies or fabricating selectively binding compounds. Con-sequently, studies on the in vivo formation of HZ afford an opportunity for chemists with a variety of interests (e.g., or-ganic, inorganic, biologic, supramolecular, materials) to make significant contributions in attacking this neglected and devas-tating disease.

A second, emerging area of research is the pathophysiologic responses between biominerals and immune cells. A common-ality between the diseases that maintain a pathogenic biomineral is the inflammatory response. Be it an auto immune reaction, as in the case of gout, or down regulation of the innate im-mune system as in malaria, it is clear that the interface between biominerals and cells is important in mitigating these responses. In the case of HZ, growing evidence suggests that the bimini-eral reacts with cellular fatty acids to produce a suite of reactive oxidized eicosanoids. The immunoreactivity of these oxidation products is likely a significant contributor to the inflammation and discomfort experienced by patients. Developing therapies for these effects and for the cause of infection could provide a dual approach in the treatment of many pathogenic bimin-erals and proffers the potential for novel initiatives in disease prevention and treatment.

References

1. Mann S. Biomineralization: principles and concepts in bioinor-ganic, inorganic, biologic, supramolecular, materials) to make significant contributions in attacking this neglected and devas-tating disease. Hemozoin: A Paradigm for Biominerals in Disease

---

Hemizone: A Paradigm for Biominerals in Disease

---
Hemoglobin: A Paradigm for Biominerals in Disease

Hemoglobin is a complex protein found in red blood cells that transports oxygen throughout the body. It is composed of four polypeptide chains, each containing a heme group. The heme group is a porphyrin ring with an iron atom at its center, which binds oxygen. In disease, hemozoin (a form of heme) can be produced, which may play a role in various pathological processes.

10 WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
See Also

Lipids, Chemical Diversity of  
Hemes in Biology  
Metal Complexes, Assembly of  
Redox Regulation and Signaling: Reactive Oxygen Species (ROS)  
Interactions of Biomaterials with Cells and Tissues
Advanced Article

Hypoxic Response and Associated Diseases

Emily Flashman, Christoph Loenarz, and Christopher J. Schofield, Chemistry Research Laboratory and Oxford Centre for Integrative Systems Biology, Mansfield Road, Oxford, U.K.

doi: 10.1002/9780470048672.wecb232

Animals must respond efficiently to changes in oxygen levels. This response is achieved via oxygen-dependent regulation of a heterodimeric hypoxia-inducible transcription factor (HIF), which enables upregulation of genes that assist in the hypoxic response. Posttranslational hydroxylation of the HIF-α subunit at either of two conserved prolyl residues enables binding to the von Hippel–Lindau protein elongin C/B complex that targets HIF-α for degradation via the ubiquitin proteasome pathway. Hydroxylation of an asparaginyl residue in the C-terminal transcriptional activation domain of HIF-α blocks its interaction with the transcriptional coactivator p300. The HIF prolyl and asparaginyl hydroxylases are oxygen dependent; therefore, their regulation of HIF directly links changes in oxygen concentration and the physiological response to hypoxia. The hypoxic response is implicated in a range of disease states, including cancer, ischemia, and heart disease. Manipulation of the HIF system for therapeutic advantage is therefore an area of medicinal interest.

In the late nineteenth century, François Viault observed that red blood cell production increased as humans moved from sea level to high altitude (1). The mechanism by which this response to an environment of limiting oxygen is induced has been a long-standing problem in physiology. A breakthrough in molecular understanding came when Semenza et al. (2) identified a transcription factor, hypoxia-inducible factor (HIF), which is responsible for the regulation of erythropoietin (EPO). EPO is a primary modulator of red blood cell production and therefore oxygen capacity in mammals, and consequently it is upregulated in conditions of hypoxia (3). An enhancer region upstream (5’) of the EPO gene was identified and was found to bind HIF in an oxygen-dependent fashion (2, 4). HRE sequences have subsequently been found upstream of numerous genes that are upregulated in response to hypoxia, including those involved in erythropoiesis, angiogenesis, metabolism, and cell growth (5). This review summarizes research on the HIF system with a focus on our current understanding of the chemistry that determines how the HIF system senses changes in oxygen availability.

Overview of the HIF System

HIF is an α,β-heterodimeric transcription factor whose transcriptional activity is regulated via its oxygen-dependent, posttranslational hydroxylation. In the presence of sufficient oxygen, HIF-α in humans is hydroxylated in both its oxygen-dependent degradation domain (ODDD) and its C-terminal transactivation domain (C-TAD). Prolyl-4-hydroxylation in the ODDD allows HIF-α to bind to the von Hippel–Lindau protein elongin C/B complex (VCB), which targets HIF-α for degradation via the ubiquitin proteasome pathway. Asparaginyl hydroxylation at the C-TAD prevents HIF-α interacting with the p300 protein, which is part of the transcriptional coactivator complex for HIF target genes. All four identified human HIF hydroxylases are Fe(II)– and 2-oxoglutarate (2OG)-dependent oxygenases with an absolute requirement for oxygen as a cosubstrate. When oxygen availability is limited, HIF-α hydroxylation is incomplete or absent. HIF-α can then dimerize with HIF-β, interact with the transcriptional coactivator complex via p300 binding, and bind to conserved pentanucleotide hypoxia response elements (HREs: TAGTC) to promote transcription of genes that enable a response to the challenge of hypoxic conditions. An overview of the HIF regulatory system is shown in Fig. 1.
Hypoxic Response and Associated Diseases

Figure 1 Overview of the HIF dual regulatory system. In the presence of oxygen (O2), the HIF-α subunits are downregulated and inactivated by active HIF hydroxylases—the PHDs and FIH, respectively. The PHDs catalyze hydroxylation of HIF-α prolyl residues, which enhances binding of HIF-α to the von Hippel-Lindau tumor suppressor complex (VCB) ∼1000-fold and enables subsequent ubiquitin-mediated proteasomal degradation of HIF-α subunits. FIH catalyzes hydroxylation of an HIF-α asparaginyl residue; this substance blocks p300 coactivator recruitment and results in transcriptional inactivation. Limiting oxygen levels (hypoxia) reduce HIF hydroxylase activity. This limitation permits heterodimerization of HIF-α with HIF-β subunits, translocation into the nucleus, and recruitment of transcriptional coactivators including p300. Subsequent binding to HREs on HIF target genes upregulates their transcription, enabling the body’s hypoxic response.

HIF-α/β and Its Interaction with DNA

Both α and β subunits of HIF are basic helix-loop-helix PAS proteins (HIF-1α/β) (6, 7), as observed in other transcription-factor proteins (Fig. 2). The short, basic region of HIF-1α/β proteins is directly responsible for DNA binding, and the adjacent helix-loop-helix domain allows dimerization of the two HIF subunits after translocation of HIF-α to the nucleus. The PAS domain (PAS-A/PAS-B/PAS-C, PAS: PER (periodic circadian protein), aryl hydrocarbon receptor nuclear translocator (ARNT), single-minded protein (SIM) (7); PAC: motif C-terminal to PAS motifs) is involved in more specific interactions between the two subunits (for review see Reference 8). It is proposed that the specific interaction between the HIF-α and HIF-β subunits promotes folding and dimerization of the HIF-α region that enables subsequent DNA binding (8). HIF-α is a Class I bHLH-PAS protein, which means that it can homodimerize and heterodimerize; HIF-β (also known as ARNT, which is a ubiquitously expressed protein) is a Class I bHLH-PAS protein, which heterodimerizes with Class I bHLH-PAS proteins (6, 9). The general PAS domain structure comprises a central β-sheet region flanked by several α-helices; NMR and mutation studies on HIF-1α have identified a hydrophobic region on the β-sheet as being responsible for the specific interaction with HIF-β, as well as homologous interactions between other PAS domains (10).

High-resolution structural data are not yet available on the interaction between the HIF heterodimer and the HRE. Interactions between bHLH proteins and DNA have shown that a conserved His-Glu-Arg triad in the bHLH binds in the DNA major groove (11). Modeling studies predict that the HIF-β domain binds to HREs in a similar fashion. HIF-α, however, only retains the Arg residue of this conserved triad, and the interaction of HIF with DNA is predicted to be mediated by other residues, specifically a serine and two alanines (12). These proposals are supported by mutagenesis studies (13).

HIF-α contains a large, central regulatory region. The oxygen-dependent degradation domain of HIF-α contains two sites of...
HIF-1α forms (31, 32). HIF-1α transcription is under hypoxia (30), and two variants that include HIF-1α compared with the intensively investigated HIF-1/2 isoforms. Three isoforms of HIF-1β (ARNT, ARNT2 (21), and ARNT3 (22)) exist; HIF-1α lacks the C-TAD that exists in both HIF-1α isoforms as it apparently has no subunit (23). mRNAs levels of HIF-1α are not affected by oxygen levels (7, 24, 25), but levels of HIF-1α mRNA increased after two hours of hypoxia (26, 27).

Additional interactions mediated by p300/CBP, as well as other coactivator mechanisms (15), are then responsible for optimal target gene expression.

### Types and Roles of Different HIF-α Isoforms

Three isoforms of HIF-α [HIF-1α, HIF-2α (endothelial PAS domain protein), and HIF-3α] have been defined in humans and are encoded by distinct loci (19, 20). HIF-1α and HIF-2α are more important than HIF-3α in the hypoxic response. Three isoforms of HIF-α (HIF-1α [ARNT], ARNT2 (21), and ARNT3 (22)) exist; HIF-1α (ARNT) is the most common form. Each HIF-α isoform can interact with any HIF-β subunit (23). mRNAs levels of HIF-1α, HIF-2α, and HIF-3α are reportedly not affected by oxygen levels (7, 24, 25), but levels of HIF-1α mRNA increased after two hours of hypoxia (26, 27).

HIF-1α and HIF-2α have the same domain architecture and common mechanisms for DNA binding, dimerization, oxygen-dependent degradation (NODD and CDDD domains), and transcriptional activation. The sequences of these domains are related closely but contain differences that may have functional consequences (e.g., with respect to HIF hydroxylation selectivity). The interdomain sequences are related less closely, and the differences may enable differential binding of regulatory proteins to HIF-1α and HIF-2α. HIF-3α is more markedly different from the other two HIF-α isoforms and apparently has no functional NODD domain, and only the CDDD domain within the N-terminal TAD has been identified. Additionally, HIF-3α lacks the C-TAD that exists in both HIF-1α and HIF-2α (28). Relative little is known about the importance of HIF-3α as compared with the intensively investigated HIF-1α isoforms.

Five HIF-1α splice variants have been reported (29). These variants include HIF-1α56, which lacks the C-TAD but not the C-terminal nuclear localization signal (NLS) and is transcriptionally active under hypoxia (30), and two variants that lack the C-terminal ODD and act as dominant negative isoforms (31, 32). HIF-3α has at least six different splice variants (HIF-3α 1–6), of which only HIF-3α 1–3 contain the ODD domain (33). HIF-3α negatively regulates HIF-mediated gene expression in murine comea by dimerizing with HIF-1α in the cytosol, maintaining an avascular phenotype under hypoxia (34), but it is upregulated by HIF-1α, which represents a negative feedback regulatory circuit (35).

Both HIF-1α and HIF-2α genes are expressed widely in human tissues but with different patterns of expression both in terms of organs and individual cell types within organs. For example, both HIF-1α and HIF-2α are expressed abundantly in the kidney but in breast cancer cell lines, HIF-1α is the major isoform that is hypersponically induced (36). Importantly, HIF-1α and HIF-2α seem to have different profiles with respect to the genes that they regulate. Expression of carbonic anhydrase IX seems to be predominantly regulated by HIF-1α, whereas erythropoietin is predominantly regulated by HIF-2α (37, 38). HIF-1α and HIF-2α null mice also give different phenotypes that support proposals of different roles for the two isoforms (39–42). An important objective is to develop a molecular understanding of how the selectivity, which is apparent in terms of the different expression profiles for HIF-1α and HIF-2α, is achieved.

### The Role of Hydroxylation

The direct interface between HIF-α regulation and oxygen is its posttranslational hydroxylation. In humans, in addition to constituting the crucial step in mammalian oxygen homeostasis (43), hydroxylation seems to be special (to date) amongst post-translational modifications (PTMs) in that only a single atom is incorporated, as compared with other additive PTMs that entail sterically more demanding groups such as phosphorylation, acetylation, glycosylation, and ubiquitylation (44). The dependence of HIF-1α stability on oxygen was demonstrated by the hydroxylation-dependent von Hippel-Lindau protein (pVHL) capture of HIF-1α peptides, which gradually increased in vivo from ambient to hypoxic oxygen levels (45). In normoxia, HIF-1α protein is hydroxylated at Pro402 (in the N-terminal ODD) and/or Pro564 in the C-terminal ODD. Both prolyl hydroxylation sites from part of a conserved motif, LXXLAP (14). Prolyl hydroxylation results in binding of pVHL and rapid degradation by the proteasome (46, 47). Under normoxia, HIF-α is thus essentially not detectable (7). The affinity of HIF-α ODD peptides for VCB increases about 1000-fold by trans-4-prolyl hydroxylation (48). Crystallographic analyses have revealed that the optimal arrangement of the new hydroxyproline alcohol group stabilizes a hydrogen bonding network in VCB between Ser111 and His115 of pVHL (Fig. 3) (48, 49). Notably, the precise positioning of the hydroxyl group may be induced by the stereoelectronic gauche effect within the vicinal N-C=C-OH arrangement, as the pyrrolidine ring of hydroxyproplyl 364 was observed in the C=C exo conformation. This conformation is the same as in collagen, in which hydroxyproplyl residues serve to stabilize the triple helix (50); a deficiency of collagen prolyl-hydroxylation leads to the disease scurvy (51). The α-domain (residues 155–213) of pVHL binds to elongin C, which nucleates a complex that contains elon- gig B, cullin-2 and Rbx1, forming the VCB-Cul2 RING-type E3 ubiquitin ligase (52). Provided oxygen is not limiting, the
that Asn803 is part of an activator complex CBP/p300 (17, 28, 57). NMR studies indicate that the hydrophobic parts of the K349 and R350 side chains. Additionally, a deeply buried in the protein–protein interface, packed against I353 and at N803 abrogates its binding to the p300 CH1/TAZ1 domain. N803 is deeply buried in the protein–protein interface, packed against I353 and the hydrophobic parts of the K349 and R350 side chains. Additionally, a network of N803 side-chain hydrogen-bonding interactions including the medicinally important 2-histidine-1-carboxylate motif, which is characteristic of this enzyme family.

### The HIF Hydroxylases

The enzymatic, posttranslational prolyl and asparaginyl hydroxylation of HIF-α was found to be dependent on oxygen and Fe(II), enhanced by ascorbate, and inhibited by 2OG analogs (28, 62–64), all suggestive of the enzymes responsible being part of the large family of Fe(II)- and 2OG-dependent oxygenases. This finding led to the identification of three human HIF prolyl hydroxylases (PHDs or EGLNs 1-3) and FIH (factor inhibiting HIF) as the enzyme that catalyzed asparaginyl hydroxylation (45, 65). These substances all contained the iron-coordinating 2-histidine-1-carboxylate motif, which is characteristic of this enzyme family.

The Fe(II) and 2OG-dependent oxygenase family is a subset of the nonheme iron (II)-dependent oxygenases, and they catalyze a wide range of oxidative reactions, often involving cleavage of an otherwise unreactive C–H bond. Examples of reaction types catalyzed by this class of enzymes include oxidative cyclization, epimerization, desaturation, C–C bond cleavage, ring fragmentation, and hydroxylation (for reviews see References 66 and 67). These reactions are often not reproducible in nonenzyme catalyzed reactions. The enzymes and the substrates on which they act are found throughout nature.

In plants and bacteria, they are known to act on small molecule substrates, for example, TauD in Escherichia coli can desulfonate taurine to provide a source of sulfur for growth (68) and anthocyanidin synthase in plants catalyzes a desaturation step in flavonoid biosynthesis (69). The enzymes also play a role in the biosynthesis pathways of antibiotic production, including the medicinally important β-lactam antibiotics (70), and can catalyze chlorination of C–H bonds (71). In higher organisms, Fe(II)/2OG-dependent oxygenases have been found to be involved in important biological processes such as DNA repair (e.g., ABH2/3 (72)) and collagen biosynthesis. Recently, a protein linked to increased fat mass in humans (FTO) has been found to be a Fe(II)/2OG-dependent oxygenate, catalyzing DNA demethylation (73).

#### Catalytic activity of the HIF hydroxylases

A consensus mechanism for the Fe(II)/2OG-dependent oxygenases has been proposed (Fig. 4). In most cases, iron is coordinated octahedrally at the active site by two histidines and the carboxylate group of a glutamate or an aspartate residue. The other three coordination positions are occupied by ligated water molecules. 2OG binds to the active site by ligating in a bidentate fashion to two coordination positions on the Fe(II) (replacing two water molecules) via its 1-carboxylate and 2-oxo...
groups. Binding of substrate (close to but not in contact with the Fe) causes a conformational change that results in alteration of the Fe(II) geometry to five-coordinate square pyramidal, and the last bound water molecule is released. This binding leaves a position free for an oxygen molecule to ligate to the Fe(II), possibly occupying the position whereby it is directed toward the target substrate. The reaction then seems likely to proceed via an Fe(III)-superoxo species, which attacks the 2-oxo group of the 2OG (susceptible to such an attack as it has been activated by the Lewis acidity of the Fe) to form a peroxide. This compound can then collapse to form carbon dioxide, iron coordinated succinate, and a highly reactive Fe(IV)-oxo species. This ferryl-oxo species is positioned adjacent to the target C–H bond such that either direct insertion of oxygen into the C–H bond occurs or the hydrogen atom is abstracted. Then, rapid rebound of the hydroxyl group is generated onto the carbon radical. Hydroxylated substrate and succinate are released from the enzyme active site, and the vacant coordination positions are once again occupied by water molecules.

Evidence to support this proposed mechanism comes from a variety of sources. The crystal structure of a cephalosporin synthase (74) showed the coordination of Fe(II) and 2OG at the active site before addition of oxygen or substrate. These complexes have subsequently been observed in multiple other structural studies (reviewed in Reference 66). EPR, circular dichroism, and UV-Vis spectroscopy studies on TfdA and CAS2 also confirmed the six-coordinate nature of the active site in solution, and they demonstrated the loss of water molecules and change in coordination environment promoted by substrate binding (75–77). Generation of data to support the later stages of the proposed reaction mechanism were more challenging, as the species involved are highly reactive and consequently unstable. Breakthrough experiments carried out on TauD however managed to demonstrate the existence of the postulated Fe(IV)-oxo species using rapid stopped-flow UV-Vis and freeze-quench Mossbauer and EPR techniques (78). The same species was later observed in a viral prolyl hydroxylase (79), which supports the hypothesis that this species is indeed the key reactive intermediate in a conserved mechanism of these enzymes.

The mechanism of the HIF hydroxylases is likely to be similar to the consensus mechanism of the Fe(II)/2OG-dependent oxygenases described above. However, the role of these particular enzymes as oxygen sensors means that their mechanism of action is of interest, particularly in terms of their dependence on oxygen. Although the intermediates in the reaction cycles of FIH and the PHDs have not been characterized to date, studies on these enzymes provide clues as to how they conform and/or differ from other enzymes in this family. Steady-state kinetic work has monitored the K_m of these enzymes with respect to their substrates and also to oxygen. These enzymes have found that in terms of oxygen sensitivity, the HIF hydroxylases (at least in an in vitro oxygen consumption assay) do not differ from other Fe(II)/2OG-dependent oxygenases (specifically TauD and PAHX) (80). Different studies are not in agreement as to whether the PHDs are more sensitive to oxygen than FIH, but all studies have reported that these enzymes are suited to their roles as oxygen sensors (80–82). Differences in the range of reactions catalyzed by the PHDs are recognized (e.g., hydroxylating NODD and/or CODD in HIF-1α-3); specific interactions are more or less likely to occur in different cell types and environments (14, 81, 83–91). For example, PHD3 has been shown to be more active on HIF-2α, which is predominantly expressed in the lung, endothelium, and carotid body (19), than on HIF-1α.

Other cellular factors are also likely to influence the catalytic activity of the HIF hydroxylases. Many Fe(II)/2OG-dependent oxygenases are known to require ascorbate for optimal activity, and ascorbate has been shown to affect cellular HIF levels accordingly (92). The role of ascorbate in the reaction cycle is not clearly understood, although one possibility is that it reduces Fe(III) to Fe(II) when 2OG has been decarboxylated in

Figure 4  Proposed outline reaction mechanism for the Fe(II)/2OG-dependent oxygenases.
the absence of prime substrate (uncoupled turnover), as is the case with collagen prolyl hydroxylase (93). Reactive oxygen species have also been proposed to reduce HIF hydroxylase activity, by oxidizing Fe(II) to Fe(III) via the Fenton reaction, decreasing its cellular availability (94). A scorbate may “repair” this situation, thus playing a general role of ensuring maximum availability of active enzyme (94, 95). In the absence of 2OG, ascorbate and Fe(II)/Fe(III) can lead to oxidative damage to the active site of some 2OG oxygenases/realted enzymes (96).

Evidence suggests that elevated levels of the tricarboxylic acid cycle intermediates fumarate and succinate (the latter a product of hydroxylase catalysis) in some tumors may lead to activation of the HIF system via hydroxylase inhibition (97, 98). Both succinate and fumarate are PHD2 inhibitors competing with 2OG for binding to Fe(II) (99, 100).

The interface between NO and the HIF system is complex and requires additional investigation (for review see Reference 101). Under hypoxic conditions, NO inhibits HIF-1α stabilization. However, under normoxic conditions, NO upregulation causes HIF-1α to accumulate. Because NO can act as an oxygen analog binding to 2OG oxygenases (102), it is tempting to speculate that the latter effect is caused by direct inhibition of the HIF hydroxylases. However, the far reaching effects of both NO and HIF on the cell biology mean that additional work is required to dissect the opposing effects of NO under normoxic and hypoxic conditions.

The situation is even more complex in the case of the interaction between reactive oxidizing species (ROS) and the HIF system (for review see References 103, 104). Quantifying the role of ROS, such as superoxide, is difficult because of their reactivity and ability to affect many cellular processes. Good evidence suggests that under stress conditions ROS can regulate HIF, possibly via interaction with the HIF hydroxylases (see e.g., References 94, 105, and 106), but the molecular mechanisms are unclear, as is the relevance of ROS regulation of HIF under normal physiological conditions.

The role of ascorbate, NO, and ROS, as well as other cellular factors, in oxygen sensing is not yet fully clear and requires more investigation. Additionally, other signaling pathways could likely be integrated with HIF regulation as indicated by studies showing that in HIF-1α, Thr-796 phosphorylation blocks A-803 hydroxylation by FIH (107).

Structural studies on the HIF hydroxylases

Since the first crystal structure of IPNS (108), which is an enzyme closely related to the Fe(II)/2OG-dependent oxygenases, structures have been determined for many Fe(II)/2OG-dependent oxygenases, all revealing a consensus core double-stranded β-helix (DSBH) fold or jelly-roll motif (46, 109). Typically this core consists of two four-stranded antiparallel β-sheets, with the major sheet supported by closely packed α-helices. The DSBH core is stabilized even more by internal hydrophobic interactions. The active site and HX/D/E, H iron binding motif reside within this core structure. A additional or varying structural features from the DSBH domain define different structural subfamilies of this class of enzyme (46). Crystal structures of FIH and PHD2 (catalytic domain) have been solved (Fig. 5C 109–112). These structures have shown that both enzymes contain the predicted core DSBH helix, but they belong to two different subfamilies of the Fe(II)/2OG-dependent oxygenases. The C-termini of FIH was shown to be involved in the formation of a functionally important dimer (107), whereas this finding was not the case for PHD2. Although in the absence of substrate, PHD2 crystallizes in a trimeric form, it does not reflect the situation in solution where PHD2 is monomeric (112). The entrance to the active site of PHD2 was also much narrower than that of FIH and other man Fe(II)-2OG-dependent oxygenases. This finding may have functional implications for this particular enzyme (e.g., in terms of the kinetics of substrate binding) and also rationalizes the observed unusually strong interaction with 2OG and iron observed for PHD2 (113). The active sites of the two enzymes show that the iron binding triad is in a similar position; however, the 2OG 1-carboxylate group coordinates to the iron in a different position in the two enzymes. The residues that stabilize the 2OG 5-carboxylate group also differ, which is characteristic of their belonging to different subfamilies of the Fe(II)/2OG-dependent oxygenases. The structure of FIH in complex with its HIF-α CAD peptide substrate shows that during substrate binding, the enzyme undergoes an induced-fit conformational change involving local conformation of specific amino acids that make contact with the substrate (110). Although structural data have not yet been published for PHD2 in complex with its substrate, it is likely that a flexible loop in PHD2 will be involved in NODD and CODD binding, perhaps changing conformation to act as a “lid” over the active site to enclose the substrate within it (112). In both PHD2 and FIH, it is necessary for substrate binding to place the target amino acid close to the active site iron, enabling efficient hydroxylation on oxygen binding.

Structures for the other human HIF hydroxylases, PHDs 1 and 3, are not yet reported, but sequence comparison and modeling studies suggest that they will be very similar to PHD2. The structure and functional implications of the N-terminal region of PHD2, which contains a MADII-like zinc finger (114), which reportedly can inhibit the hydroxylation activity of the catalytic domain, are unknown.

Alternative substrates for the HIF hydroxylases

Alternative substrates may exist for the PHDs; proposed examples include RNA polymerase II and ICB kinase (which is negatively regulated by PHD1) (115, 116). However, unequivocal evidence (e.g., demonstration of hydroxylation by mass spectrometry) has not yet been demonstrated for these proteins. In contrast, FIH has been shown to catalyze hydroxylation of asparaginyl repeat domain (ARD) proteins from the NF-κB (nuclear factor κB) and Notch family at highly conserved asparaginyl residues (117, 118). The ARD is a common protein motif, with over 200 human members of the ARD protein family being predicted. Evidence that ARD hydroxylation occurs frequently in human cells supports the assertion (117, 118) that posttranslational hydroxylation of cytoplasmic proteins in
Hypoxic Response and Associated Diseases

Eukaryotes is actually more common than previously thought (119). The significance of ARD hydroxylation is unclear, and it may serve to stabilize the ARD fold in a similar way that prolyl-4-hydroxylation stabilizes the collagen triple helix (50). ARD proteins were found to compete with HIF-α for FIH binding, and it was thus proposed that the hydroxylation status of the pool of ARD that can interact with FIH regulates the amount of FIH that is "free" to hydroxylate HIF-α (120). Recently, FIH-mediated ARD hydroxylation has been observed in ASB4 (ankyrin repeat and SOCS box protein 4), allowing it to target substrate proteins for ubiquitin-mediated degradation (121) in an oxygen-dependent fashion. This observation indicates that hypoxic regulation by the HIF hydroxylases may extend beyond the HIF system.

Diseases Associated with the HIF System

The hypoxic response plays an important role in mammalian biology (122), and alterations in tissue oxygenation are hallmarks of many human diseases including stroke, heart disease, vascular disease, and cancer (123). Across a range of cancers, observed associations between aggressive phenotypes and micro-environmental hypoxia have focused attention on the molecular dissection of hypoxia pathways (124).

The most direct link (125) between genetic events predisposing to cancer and the HIF system occurs in von Hippel–Lindau (VHL) disease, which is an autosomal dominant inherited cancer syndrome affecting 1/35,000 humans (126). VHL disease is often regarded as a model syndrome among the oxygen-dependent diseases and is characterized by renal cell carcinomas (RCC), as well as tumors in the central nervous system, retina, adrenal glands, and pancreatic islet cells (127); malignant tumors also occur in the inner ear, pancreas, and epididymis (128). VHL syndrome is caused by mutations in the VHL gene that can result either in prematurely truncated protein transcription or in point mutations of specific amino acids. RCC are a common form of kidney cancer, and mutations in both alleles of the VHL gene cause ~80% of sporadic RCC (129). Mutations in the VHL gene affect the function of the pVHL protein, and the resultant disease phenotype has been partly rationalized by crystal structures of the pVHL E3 ligase complex with HIF-α (48, 49). VHL disease is associated with significant stabilization of HIF-α (130), as mutation of pVHL in either its α or β domain results in an inability to recruit other E3 ligase complex subunits or to recognize hydroxylated HIF-α, respectively. HIF overproduction and subsequent overexpression of HIF targets stimulates excessive blood vessel formation and nutrient uptake, which leads to formation of highly vascular tumors (131).

Another disease that develops from VHL mutations is the increase in red blood cell production, or polycythemia, endemic in the Chuvash population (132). The Chuvash disease does not manifest in tumors but is associated with venous abnormalities and a tendency toward arterial thrombosis (133). It is an autosomal recessive disorder in which mutations that affect pVHL (e.g., R200W) result in modulation of the HIF pathway (134–136).

Figure 5

Comparison of the overall structures of FIH and PHD2 (upper panel) and differences in the active site of these enzymes (lower panel). Note the different orientation of the 2OG cosubstrate or cosubstrate analog, relative to the triad of Fe binding residues.
RCC cells show a bias toward relative overexpression of HIF-2α rather than HIF-1α (137, 138), and subsequent inhibition of HIF-2α was shown to be sufficient to suppress tumor growth (129). Consequently, it was proposed that their different transcriptional selectivity of HIF-1α and HIF-2α affects retardation or promotion of RCC tumor growth, respectively; the former promotes expression of proangiogenic factors, whereas the latter activates expression of protumorigenic genes such as cyclin D1, TGF-α, and VEGF (140). Recently, ablation of HIF-2α in mice after birth was reported to result in tumor suppression, suggesting that HIF-2α may be the critical isoform that regulates EPO under physiological and stress conditions in adults (141).

Recent clinical findings underline the proposed role for PHD2 as the key oxygen sensor involved in HIF regulation in normoxic tissues (86). An inherited heterozygous PHD2 P317R was shown to be associated with familial erythrocytosis (142). PHD2 Pro317 is located only two residues from the iron-binding Asp315 in a p-helix and its proximity to P317 R was shown to be associated with familial erythrocytosis (86). An inherited heterozygous PHD2 mutation, R371H, which caused decreased HIF binding and catalytic activity (143).

Changes in the levels of HIF are also associated with a range of other disease states. HIF-1α levels have been reported to increase as an early response to myocardial ischemia in patients undergoing coronary bypass surgery (144). Ischemic tissues induce the production of proangiogenic cytokines that stimulate blood vessel remodeling and growth (145). In elderly patients with atherosclerotic narrowing of blood vessels, these proangiogenic responses are impaired, which can lead to critical limb ischemia (146). Impaired recovery from this ischemia was attributed to age-related impairment of HIF-1α protein expression, as observed in mice: HIF-1α gene therapy was reported to counteract these pathogenic effects of aging (147).

Recently, complex interactions between HIF and proinflammatory signaling cascades have emerged. Oxygen delivery to sites of tissue damage, for example from trauma, inflammation, or infection, may be compromised or insufficient for metabolic demands. Recent work has provided direct evidence for the regulation of the innate immune response by the HIF/p300 pathway (148), and the key pro-inflammatory transcription regulator NF-κB has been linked to HIF-1α-dependent and independent proinflammatory macrophage cytokine release (149). Intriguingly, both ankyrin repeats within the NF-κB repressor IκBα, and the p105 subunit of NF-κB itself have been shown to be posttranslationally hydroxylated by the asparaginyl hydroxylase FH (117) that also hydroxylates HIF-α, blocking interaction with its transcriptional coactivators.

**Therapeutic Possibilities**

Therapeutic possibilities that involve the HIF system may be divided into those that upregulate or those that downregulate HIF target genes or proteins (for reviews see References 150–152 and others). Because strong evidence suggests that HIF-α upregulation occurs in many human tumors (reviewed in Reference 153), downregulation of HIF-α is an attractive anticancer strategy. The HIF target gene VEGF induces angiogenesis and is of particular interest with respect to tumor therapies; it has already been targeted successfully (154).

Several screens looking to identify small molecules that downregulate HIF-α have been conducted (155–157). For example, the mammalian target of rapamycin is involved in protein synthesis, and although it seems to be a generic target, specific inhibitors of this protein have been found to result in lower cellular levels of HIF and are in clinical development as anticancer agents (158). Curcumin, a constituent of the spice and colorant turmeric that has been linked to cancer treatment, has been found to minimize the availability of HIF by promoting degradation of HIF-α (159). Heat-shock protein 90 (HSP90) is known to stabilize HIF-α by binding to the PAS B domain, preventing nonspecific degradation in hypoxia before nuclear translocation (160); HSP90 inhibitors (e.g., geldanamycin) have also been shown to reduce HIF-1α levels (161). Of course, promoting the activity of the HIF hydroxylases would also reduce the availability of HIF: to this end, treatment with 2OG has shown decreased VEGF production and angiogenesis in vitro (162), although it is unclear whether 2OG is likely to be limiting in vivo.

Anticancer drugs can also be designed to prevent HIF from performing its functions, for example targeting the dimerization of the α and β subunits, DNA binding, or preventing HIF from binding to its cotranscriptional activators, specifically p300. Some progress has been made taking advantage of the latter opportunity. For example, the interesting natural product chetomin, which is a disulphide that contains diketopiperazine, blocks binding of HIF-α to p300, resulting in abrogation of the transcription of hypoxically regulated genes both in vitro and in vivo (163). DNA-binding small molecules can also prevent HIF binding to the HRE, which results in decreased VEGF expression in cell cultures (164).

The hypoxic response can also be inhibited using gene therapy, by downregulating either HIF itself or, more specifically, the HIF-target genes such as erythropoietin and VEGF. However, the effects of HIF-silencing are not straightforward, and they do not necessarily correspond with a decrease in tumor size or vasculature (reviewed in Reference 165), as hypoxia-mediated cell death helps impede tumor growth.

EPO is very widely used for the treatment of anemia, which demonstrates the potential for upregulation of HIF target genes in medicine. HIF itself could be directly upregulated; a recent study has successfully shown that adenosine transfer of constitutively active HIF-α was tolerated and shown to improve the condition of humans with peripheral limb ischemia (166). In aging mice, the same treatment resulted in improved blood flow in ischemic tissue (147). Alternatively, upregulation of the HIF-target genes can be therapeutically advantageous, and many studies have involved increasing VEGF levels with encouraging outcomes (reviewed in Reference 167).

Little evidence exists for most individuals that moderately limiting oxygen causes long-term health defects. Thus, the points at which oxygen availability is sensed by the HIF system (i.e., the HIF hydroxylases) are arguably preferred targets for therapeutic intervention. Very good precedent was set for...
targeting metalloenzymes by small-molecule inhibitors (e.g., cyclooxygenases), so significant efforts are being developed toward the generation of HIF hydroxylase inhibitors. The requirement of the HIF hydroxylases for Fe(II) also suggests that competition for binding at the metal site by other metals may be a therapeutic possibility. Indeed, Co(II) ions have long been known to induce the hypoxic response (168). However, this strategy is likely to be nonspecific. Instead, following the demonstration that the 2OG analog, N-oxalylglycine, acts as a PHD inhibitor (45, 169), efforts have focused on identifying other organic inhibitors (45, 169–176). To date, all of those reported likely act as 2OG competitors and chelate to the active site Fe(II). In some cases, inhibitors have been shown to target the HIF hydroxylases in primates, which resulted in reportedly well-tolerated VEGF production (177).

It will be clearly desirable to develop HIF hydroxylase inhibitors that are selective for their targets (e.g., over other 2OG oxygenases). Another important question with targeting the HIF system is to what extent selectivity in terms of modulating the level of specific combinations of HIF target genes is desirable. For some diseases, it may be optimal to mimic the "natural" hypoxic response as closely as possible, whereas for others, possibly anemia, it may be best to target specific genes. The availability of HIF hydroxylase inhibitors as small-molecule probes, together with knockout mouse studies, should help to address questions as to the therapeutic use of targets within the HIF system.

Acknowledgments

We thank the E.P.S.R.C., the Rhodes Trust, the B.B.S.R.C., the Wellcome Trust and the National Institutes of Health for funding. We also thank all members of our laboratory and our close collaborators, especially the Ratcliffe/Pugh group, for their contributions.

References

Hypoxic Response and Associated Diseases


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.


64. Duncan T, Trewick SC, Koivisto P, Horsfall LE, Schofield CJ. Identification of a region of HIF-1alpha that is unique to its prolyl hydroxylase domain. J. Biol. Chem. 2002;277:26343-26351.


74. Duncan T, Trewick SC, Koivisto P, Horsfall LE, Schofield CJ. Identification of a region of HIF-1alpha that is unique to its prolyl hydroxylase domain. J. Biol. Chem. 2002;277:26343-26351.


Hypoxia Response and Associated Diseases

87. Chan DA, Subotzky PK, Yen SE, Giaccia AJ. Coordination regu-
lation of the oxygen-dependent degradation domains of hypoxia-

88. Huang J, Zhao Q, Mooney SM, Lee FS. Sequence determi-
nates in hypoxia-inducible factor-1α for hydroxylation by the

89. Landauz MB, Varas-Vega A, Vilot M, Cuevas Y, del Peso
A. Analysis of the hypoxia-inducible factor-1α 3′-untranslated
region binding to substrate

90. Metzen E, Bercht-Feilenschmidt U, Slentz P, Manson JH,
Stark J, Kling M, Huang WQ, Witztum C, Hellewig-Burget
T, Jelkmann W, Acker H, Fandrey J. Intracellular localization
of human HIF-1α alpha-hydroxylase: implications for oxygen

91. Tuckerman JR, Zhao Y, Hewitson KS, Tian YM, Pugh CW,
Ratcliffe PJ, Miao DR. Determination and comparison of specific
activity of the HIF-proline hydroxylases. FEBS Lett. 2004;576:
145–150.

92. Knowles HJ, Rawal RR, Harris AH, Ratcliffe PJ. Effect of
acetaldehyde on the activity of hypoxia-inducible factor in cancer

KI. Ascorbate is consumed stoichiometrically in the uncoupled
reactions catalyzed by prolyl 4-hydroxylase and lysyl hydroxy-

94. Gerald D, Berra E, Frapart YM, Chan DA, Giaccia AJ, Mansuy
IM, Yen SE, Roussel MA, Pouyssegur J, Yaniv M, Mechta-Grigoriou F.
JunD reduces AP-1 activity by stabilizing hypoxia-inducible factor
alpha (HIF-1α) levels. J. Biol. Chem. 2002;277:39792–39800.

Oxidation and mutagenesis studies on the iron(II) binding site of

96. Roach PL, Cafferty ME, Lienard BM, McDonough MA, Aplin RT, He-
witson KS, Pugh CW, Ratcliffe PJ, Schofield CJ. Disruption of different
and substrate phosphorylation inhibits factor (HIF)-2α.

97. Chandel NS, McClintock DS, Feliciano CE, Wood TM, Me-
lendez JA, Rodriguez AM, Schumacker PT. Reactive oxygen
species generated at mitochondrial complex III stabilize hypoxia-inducible factor alpha (HIF-1α) as a mechanism of

98. Richard DE, Berra E, Pouyssegur J. Hypoxia pathway mediates
the induction of hypoxia-inducible factor (HIF) 1α in vascular

99. Roach PL, Cafferty ME, Lienard BM, McDonough MA, Aplin RT, He-
witson KS, Pugh CW, Ratcliffe PJ, Schofield CJ. Disruption of different
and substrate phosphorylation inhibits factor (HIF)-2α.

100. Koivunen P, Hirsila M, Remes AM, Hassinen IE, Kivirikko
KI. Inhibition of hypoxia-inducible factor-inhibiting hypoxia-
inducible factor (HIF) reveals mechanism of HIF-1α.

101. Cash TP, Pan Y, Simon MC. Reactive oxygen species and cellular

102. Getzen A, Kitabchi A, Suplervic and derived reactive oxygen
species in the regulation of hypoxia-inducible factors. Meth-

103. Quan J, Chen X, Ong CY, Limwongse H, Chan DA, Giaccia AJ,
Mansuy IM, Howard HH, Carbajal BJ, Chang S, Pouyssegur J,

104. Cummins EP, Berra E, Comerford KM, Ginouves A, Fitzgerald
A, Oldham NJ, Schofield CJ. Structure of factor-inhibiting hypoxia-inducible factor 1: an asparaginyl hydroxylase

105. Chandel NS, McLaughlin MA, Seibel JF, Schiem-
ingler I, Pugh CW, Ratcliffe PJ, Schofield CJ. Structure of factor-inhibiting hypoxia-inducible factor (HIF) reveals mech-
anism of oxidative modification of HIF-1α. J. Biol. Chem.
2003;278:1802–1806.

FIH-1 reveals a unique active site pocket and interac-
tion sites for HIF-1α and von Hippel-Lindau. J. Biol. Chem.
2003;278:7553–7563.

CJ. Hypoxia-inducible factor prolyl hydroxylase 2 has a high affinity for ferrous iron and 2-oxoglutarate. Mol. Biol. 2003;31:
321–324.

A, Oldham NJ, Schofield CJ. Structure of factor-inhibiting hypoxia-inducible factor 1α: a mechanism of postme-
chanistic studies on the inhibition of the hypoxia-inducible
factor-inhibiting hypoxia-inducible factor (HIF) reveals hydroxylase by citric acid cycle intermediates: possible links
between cell metabolism and stabilization of HIF. J. Biol.
Chem. 2007;282:4524–4532.


110. Schführung CJ, and Zhang ZH. Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes.
Hypoxic Response and Associated Diseases


122. Wiesener MS, Maxwell PH. HIF and oxygen sensing: as important to life as the air we breathe? Annu. Med. 2003;35:183–190.


139. Wiesener MS, Maxwell PH. HIF and oxygen sensing: as important to life as the air we breathe? Annu. Med. 2003;35:183–190.


Hypoxic Response and Associated Diseases

153. Melillo G. Inhibiting hypoxia-inducible factor 1 for cancer ther-

151. Semenza GL. Development of novel therapeutic strategies that

150. Hewitson KS, Schofield CJ. The HIF pathway as a therapeutic

149. Peyssounard C, Czejdo-Martin P, Doerdes A, Zinkernagel A S,
Johnson RS. Nitric oxide: cutting-edge essential role of "hypoxia in-
ducible factor-1" in development of lipopolysaccharide-induced sepsis.

148. Mielelli S, Schofield CJ. The HIF pathway as a therapeutic

147. Melillo G. Targeting hypoxia cell signaling for cancer therapy.

146. Melillo G. Inhibiting hypoxia-inducible factor 1 for cancer ther-

145. Herdt R S. Therapeutic options to target angiogenesis in human

144. Olenyuk BZ, Zhang GJ, Klco JM, Nickols NG, Kaelin WG Jr,
Mekel M, Evdokimov AG, East S. Design and synthesis of substituted

143. Warshakoon NC, Wu S, Boyer A, Kawamoto R, Renock S, Xu K,


Further Reading

See Also

Approaches to Enzyme Inhibition
Post-Translational Modifications, Chemical Biology of Diseases Related to Metallo-Enzymes and Metallo-Proteins, Chemistry of Oxygen-Activating Enzymes, Chemistry of Ubiquitin Proteasome Pathway
Lysosomal Disorders
Doug A. Brooks, Molecular Medicine Sector, Sansom Institute, University of South Australia, Adelaide, Australia
Maria Fuller, Lysosomal Diseases Research Unit, Department of Genetic Medicine, Women’s and Children’s Hospital, Adelaide, Australia
doi: 10.1002/9780470048672.wecb297

Lysosomal storage disorders are a group of over 50 different genetic diseases that result from lysosomal dysfunction. This disruption of lysosomal function can involve either a specific lysosomal hydrolase deficiency, a defect in lysosomal protein processing, or impaired lysosomal biogenesis. To appreciate the pathogenesis of lysosomal disorders fully, it is important to understand the dynamics of endosome–lysosome organelles, their capacity for the uptake and degradation of complex macromolecules, and how lysosomal biogenesis and hydrolisis are altered by substrate storage. The focus of this article will be on recognized lysosomal disorders and what is known about the composition and the function of endosome–lysosome organelles in these diseases. The lysosomal network will be discussed with a view to correlating the main site of endosome–lysosome degradation and the site of substrate accumulation in lysosomal disorders. A major unanswered question for lysosomal storage disorders is how an enzyme deficiency and the resulting storage of undegraded macromolecules impact on cells to cause organ dysfunction and disease.

A patient with possible Hurler syndrome was first described by Berkhan in 1907 (1), and it may be the first report of a lysosomal storage disorder. Nonetheless, the first detailed clinical description of a patient with a lysosomal storage disorder was by Charles Hunter, who described two brothers who are now recognized as having Hunter syndrome (2). Soon after this report, Meinhard von Pfaundler and Gertrud Hurler described Hurler-Pfaundler syndrome, which is now known as Hurler syndrome (3). A more detailed history of syndrome identification and the subsequent recognition of other lysosomal storage disorders have been documented by Whitley in 1993 (4).

In the 1960s, Hers and colleagues (5, 6) recognized that Pompe disease was caused by a deficiency of α-glucosidase and, using electron microscopy, evidence of storage vacuoles was reported. The identification of this enzyme deficiency, together with De Duve and colleagues’ description of lysosomes and their contents (7, 8), led to the concept of “lysosomal storage disorders.” This finding resulted in two seminal publications by Hers and Van Hoof that described “Lysosomes and Storage Diseases” (9, 10). It should be noted that much of the basic knowledge about the cell biology of lysosomes was contributed to strongly by these initial investigations on lysosomal storage disorders.

In 1999, Meikle and colleagues (11) reported the prevalence of lysosomal storage disorders as 1 in 7,700 live births for an Australian study that involved 27 different diseases. Since then, several new disorders have been recognized, and it is now accepted that, as a group, more than 50 different lysosomal storage disorders exist. Moreover, in some populations the prevalence of certain lysosomal storage disorders has been reported to be high, including 1 in 18,500 for aspartylglucosaminuria in the Finnish population (12) and 1 in 3,900 for Tay-Sachs in the Ashkenazi Jewish population (13). This finding led to past speculation that the combined incidence of lysosomal storage disorders may be as high as 1 in 1500 births (14), and more recent estimates suggest this prevalence is approximately 1 in 1000 births (http://www.science.org.au/sats2007/hopwood.htm). Lysosomal storage disorders are now recognized as a substantial group of genetic diseases that result in lysosomal dysfunction, leading to a failure to degrade specific substrates, which then accumulate in endosome–lysosome organelles.

Endosomes and Lysosomes
From early observations, De Duve defined lysosomes as cytoplasmic particles that were associated with a range of acid hydrolases (7). At the electron microscope level, membrane-bound vacuoles or compartments were recognized and shown to contain acid hydrolases that were detectable by cytochemical staining (15). The definitive description of a lysosome (16) includes a membrane-bound organelle compartment that is acidic and
Lysosomal Disorders

contains a range of mature acid hydrolases (e.g., proteases and glycosidases). This organelle represents the most distal compartment in the endocytic pathway (Fig. 1) and is distinct from the prelysosomal compartment and endosomal compartments based on the absence of mannose-6-phosphate receptors (the receptors that are responsible for the targeting and trafficking of soluble lysosomal enzymes; see below). Lysosomes are heterogeneous in size, shape, and composition, and they exhibit high density in organelle fractionation experiments. The lysosome is the most distal compartment for lysosomal enzymes trafficking from the biosynthetic pathway. As defined by Storrie in 1988 (16), the lysosome must be the principal domicile of a “lysosomal protein,” but the presence alone of a lysosomal protein in an organelle structure does not necessarily establish that organelle as a lysosome (for example, lysosomal proteins can also be detected in endosomes and phagosomes).

Lysosomes are now known to be essential organelles involved in the turnover and the reuse of cellular macromolecules such as proteins, lipids, glycoproteins, and glycosaminoglycans. Lysosomes have other diverse functional roles in immune function, pigmentation, signaling, cell adhesion/motility, and membrane repair, and essentially they are a dynamic interface with the extracellular environment. Numerous subcellular events are required for the synthesis and the delivery of functional degradative enzymes to lysosomes and for the assembly of functional lysosomal organelles (17). This process is referred to as lysosomal biogenesis, and the transport of newly synthesized lysosomal proteins proceeds through another set of organelle compartments called endosomes (Fig. 1).

Endosomes mediate the delivery of degradative enzymes from the biosynthetic compartments to both lysosomes and the extracellular milieu and function in the processing and transport of secretory products (Fig. 1). Endocytic organelles are also involved in the internalization and the delivery of material from the extracellular milieu and cell surface to compartments inside the cell (18, 19). These two main intracellular pathways from the biosynthetic compartment and cell surface are convergent (Fig. 1) (20). Other specialist compartments exist within the network of endocytic organelles, which include phagosomes that sequester cytoplasmic material, and organelles for turnover and recycling. This complexity of endosome and lysosome organelles necessitates strict control of targeting and trafficking events. Intracellular traffic between these compartments can occur through specific vesicle formation that involves budding and fusion of membrane to and from different organelles, but other types of transient interaction are possible (see below).

Components of Endosomes and Lysosomes

One crucial role of the membrane that encloses the endosomal and lysosomal compartments is the isolation of the potent acid

![Diagram of intracellular organelles](image)
hydrolyses that are key constituents of these compartments (21). The limiting membrane of lysosomes has proteins involved in membrane structure, compartment acidification (ATPase), ion transport, and vesicular traffic. Approximately 20-30 major polypeptides of molecular mass 15-200 kDa exist in lysosomal extracts, and most of these are highly glycosylated (16, 22). Typically, the mannose-6-phosphate receptors that target luminal lysosomal proteins are absent from the end-stage lysosomal compartment but are present in endosomes. The lysosome is composed of at least 20 known membrane proteins and over 50 luminal lysosomal proteins (22). The 50 or more known lysosomal storage disorders mostly relate to a dysfunction of one or more of the soluble acid hydrolases that are normally involved in macromolecular break-down, which includes proteases, glycosidases, sulfatases, phosphatases, and lipases. However, several lysosomal storage disorders involving membrane proteins, transporters/channels, and altered vesicular trafficking machinery have been recognized (23).

The major lysosomal associated membrane proteins—LAMP-I and LAMP-II—are type-I integral membrane proteins; they have a single membrane-spanning sequence, a highly glycosylated luminal domain, and a short cytoplasmic tail sequence involved in targeting/trafficking (24). Two other major lysosomal integral membrane proteins—LIMP-I/CD63 and LIMP-II—are both type II integral membrane proteins, with four and two membrane-spanning domains, respectively. Newly synthesized LAMP and LIMP molecules traffic via the trans-Golgi network (from the biosynthetic compartment) to endosomal/lysosomal compartments based on either tyrosine (LAMP-I, LAMP-II, LIMP-I/CD63) or di-leucine (LIMP-II)-sorting signals in the cytoplasmic domains of these molecules (25, 26). At steady-state, most LAMP-I and LAMP-II molecules are localized to the limiting membrane of endosomes and lysosomes. However, LAMP-I and LAMP-II are also detected in autophagic vacuoles. A major proportion of LIMP-II molecules are localized to endosomes and seem to have a role in the biogenesis of these organelles (27). LIMP-I/CD63 is primarily localized to the internal membranes of endosomal vesicles of multivesicular bodies and, to a lesser extent, is found at the limiting endosome-lysosome membrane and cell surface. Functionally, LAMP-II has been shown to have an important role in autophagy and seems to have the capacity to compensate for a LAMP-I deficiency (28). The exact function(s) of LIMP-I/CD63 is unclear, but it seems to have a role in immune cell activation, where it is subsequently expressed at the cell surface. LIMP-I/CD63 is a member of the tetrascaran family of proteins (29), which have also been implicated in the control of membrane and cell volume, as well as cell adhesion, cell motility, and antigen presentation.

Specific membrane and membrane-associated proteins also control the organelle traffic and fusion events that are associated with the movement of proteins between different intracellular compartments. Membrane proteins of the secretory and endocytic pathways depend on sorting signals that reside in their cytoplasmic domains. Many proteins exhibit multiple signals that determine their passage along these diverse pathways. Therefore, the steady-state distribution of any given membrane protein is dictated by the specific combination of sorting signals in the protein and the interaction of these signals with specific recognition molecules (26). This intracellular network of organelles and vesicles (Fig. 1) is in constant and dynamic flux.

Glycerophospholipids, cholesterol, and sphingolipids are the essential building blocks for all eukaryotic cell membranes (30). The lysosomal membrane, like other eukaryotic membranes, is composed not only of highly glycosylated proteins but also is enriched in amphiphilic lipids (30, 31). The lipid and protein composition of the lysosomal membrane is believed to be very complex, and this complexity ensures selective degradation such that the lysosomal membranes remain intact. This finding has led to the assumption that at least two distinct pools of lipid exist in the lysosomal membrane (32). Glycerophospholipids are important components of lysosomal membranes, which are involved directly in lipid rafts, and these are involved in both membrane transport and signaling (31). The early studies on the glycosphingolipidoses (a subgroup of lysosomal storage disorders) suggested that aggregates of lipids accumulated as multivesicular storage bodies in lysosomal compartments (33, 34). Notably, cholesterol is normally enriched in the membranes of early endocytic organelles but not in lysosomes.

Targeting of Protein Constituents to Lysosomes

Most lysosomal proteins have either high mannose or complex oligosaccharide side chains that are attached to asparagine residues in the polypeptide at consensus (NXS, NAT sites) where N is asparagine, S is serine, T is tyrosine, and X is any amino acid) glycosylation sites (16, 35, 36). In the case of soluble lysosomal proteins, N-linked glycosylation is attached to the growing polypeptide chain in the endoplasmic reticulum, and this forms the basic structure that is modified to generate the mannose-6-phosphate targeting signal that are involved in targeting these proteins to the lysosome (37). This latter processing event occurs in the cis-Golgi apparatus and involves the enzyme N-acetylglucosamine-1-phosphotransferase, which adds an N-acetylglucosamine-1-phosphate to certain mannose oligosaccharides. The removal of the N-acetylglucosamine residue in the trans-Golgi exposes the mannose-6-phosphate targeting signal that then allows soluble lysosomal hydrolases to interact with mannose-6-phosphate receptors for trafficking to endosomes and delivery to lysosomes (reviewed in References 37–39).

Thus, within the lumen of organelle compartments, integral membrane proteins act as cargo receptors that recruit soluble molecules for traffic within the network of organelles comprising the secretory, endocytic, and lysosomal pathways. For example, the cation-dependent mannose-6-phosphate receptor (CD-MPR; 46-kDa dimer, which requires calcium (40)) and cation-independent mannose-6-phosphate receptor (C1-MPR; 300-kDa, also called the IGF II receptor (38)) are involved in binding soluble lysosomal proteins in the trans-Golgi network for targeted delivery of these enzymes to the endosome.
lysosome system. Mannose-6-phosphate receptors are type I glycoproteins (have a single transmembrane sequence) and recycle between the trans-Golgi network, endosomal compartments, and the cell surface. Surprisingly, this recycling process does not seem to depend on whether the receptor is loaded with ligands (26). Mannose-6-phosphate receptors normally release their lysosomal enzyme cargo molecules in the prelysosomal compartment as a result of the low pH environment (pH 5.0–6.0). The free mannose-6-phosphate receptor is then returned to the trans-Golgi for subsequent cargo delivery. Therefore, lysosomes are defined as acid hydrolase-rich organelles that lack both the CD-MPR and CI-MPR, which distinguishes them from endosomes (42).

Sorting signals that are either lysosine-based (NPXY, YXX; where N is asparagine, P is proline, and X is any amino acid) or leucine-based (DE)XXXL(LI), DXXLL; where D is aspartic acid, E is glutamic acid, L is leucine, I is isoleucine, and X is any amino acid) are found within the cytosolic domains of transmembrane proteins (26). Sorting occurs through coated areas of membranes comprised of proteins such as the adaptor protein (AP) complexes and Golgi gamma ear adaptin (GGA 1–3) proteins, which act to bind targeting motifs (26). The NPXY motif, which is found within a subset of type I membrane proteins, mediates internalization from the plasma membrane alone. The YXX motif is found in the CI-MPR and the CD-MPR, LAMP-1, LAMP-2, and CD63, and it is involved in a wide variety of sorting processes. Positioning of the motif relative to the membrane is also critically important for the recognition of these targeting signals. The YXX–AP-2 interaction is facilitated by a conformational change induced by phosphorylation, which is part of targeting regulation. The dileucine signal (DE)XXXL(LI) found in many type I, type II, and multispansing proteins also binds to AP complexes. The LL and LI motifs exhibit a distinct preference for the complexes that are critical for the fine specificity of targeting. The DXXLL motif is found within several transmembrane receptors, CI-MPRs and CD-MPRs, which cycle between the trans-Golgi network and the endosomes. Mannose-6-phosphate receptors bind the VHS domain of GGA proteins only via this signal, which is regulated by phosphorylation of serine residues. GGA1 and GGA3 also have the DXXLL motif within their hinge regions. Binding the hinge region with the VHS domain invokes an auto-inhibitory effect. Phosphorylation of the GGA serine results in auto-inhibition and transfer of the mannose-6-phosphate receptors from GGA1 to AP-1. Ubiquitin has been shown to be involved in sorting at the cell surface, endosomes, and trans-Golgi network. Endocytic proteins, which include epin, Hrs, and STAM, have a ubiquitin-interacting motif that binds directly to ubiquitin, which suggests that they could act as adaptors for sorting ubiquitinated cargo at discrete intracellular sites (26).

Interaction of Endosomes and Lysosomes and the Site of Substrate Hydrolysis

Many ligands that enter the endocytic pathway via receptor interaction are either sorted for traffic along specific organelle pathways (Fig. 1) or transit to the late endosome where ligands are dissociated from the receptors in the acid pH environment. For lysosomal enzymes, this dissociation event from mannose-6-phosphate receptors allows additional traffic to the most distal element of the endocytic machinery, the lysosomal compartment. A constant flux of membrane proteins seems to occur between late endosomes and lysosomes (43). It has been postulated that lysosomes fuse with endosomes to form a transient compartment that seems to be the major organelle involved in macromolecule degradation [i.e., “the cell stomach” (44)]. This finding implies that lysosomes are effectively a reservoir for acid hydrolases that are then tapped when required, whereas the late endosome is the degradative compartment. Moreover, lysosomal constituents can be recovered from the prelysosomal compartment and reformed as a lysosomal organelle (43). This theory helps to explain the heterogeneity of endocytic organelles, but it is consistent with the initiation of proteolytic/hydrolysis of macromolecules in endosome and prelysosomal compartments. The molecular machinery that mediates these vesicular traffic, fusion, and budding processes is yet to be defined fully. However, studies on a FYVE finger protein localized to early endosomes called Hrs and the endosomal-sorting complex required for transport (ESCRT) have indicated that a group of proteins that recognize ubiquitin motifs are involved directly in endosomal sorting and recruitment of proteins into multivesicular endosomes (45–48).

The mechanism for the transfer of endocytosed material between endosomes and lysosomes has generated many theories, which include maturation of endosomes into lysosomes, vesicular transport between endosomal and lysosomal compartments, transient interaction via channel formation (“kiss and run”), and direct fusion (44). Depending on the cargo involved, all of these mechanisms seem to be used (perhaps in different degrees) to regulate the level of hydrolytic capacity in any given compartment. Thus, lysosomes can interact with not only endosomal compartments but also phagosomes, autophagosomes, and the plasma membrane to effect different functional roles. For a lysosomal storage disorder, this model of lysosomal function would predict that endocytic cargo, such as glycosaminoglycan that is destined for degradation after internalization from the cell surface, would be delivered to a late endosome for degradation. In an attempt to degrade the substrate, lysosomes would either infuse lysosomal hydrolases into the late endosome and/or fuse directly to the late endosome to generate a hybrid organelle. The latter would provide maximum delivery of hydrolases, and a mechanism by which the cell could try to compensate for the reduced catalytic capacity that develops from a hydrolase deficiency. However, a failure to degrade the substrate contents may result in the maintenance of these hybrid endosome-lysosome organelles.
Evidence that the latter model of endosome-lysosome fusion and lysosomal recovery is valid comes from the lysosomal storage disorder mucolipidosis type IV (MLIV), which involves a defect in a Ca\(^{++}\) channel that seems to prevent the retrieval of lysosomes from the hybrid organelle (49). Thus, intra-organellar Ca\(^{++}\) is presumed to be required for the condensation process that leads to lysosomal organelle reform. The trafficking of the glycosphingolipid lactosylceramide is inhibited in MLIV and, in common with other lipids, it may then accumulate inappropriately because of this blockage. In other lysosomal storage disorders that involve a hydrolase deficiency, the failure to degrade intraorganelar substrate(s) may signal an incomplete process of degradation and, in the same way, potentiate the hybrid organelle and lead to the same adverse effects on cell function. Similarly, a defect in the organelle trafficking machinery can also impact on endosome-lysosome degradation to generate similar compartments, which also then accumulate undegraded storage material. Thus, a recent report of a defect in the endosomal sorting complex required for transport (ESCRT-III) machinery (50) impacts directly on late endosomes (multivesicular bodies), which generates compartments with a morphology remarkably similar to that observed in other lysosomal storage disorders that are caused by hydrolase deficiencies (e.g., Sanfilippo syndrome). This finding raises questions about the similarity between the vesicular pathology in different lysosomal storage disorders and whether common points for pathogenesis lead to similar clinical outcome.

Commonalities for Lysosomal Storage Disorders in Terms of Clinical Phenotype and Vesicular Pathology

Most patients with lysosomal storage disorders are born with no clinically obvious signs of disease, but in the severe forms of the disease, onset and progression of symptoms is rapid. In most cases, the severe form of each disorder is devastating (see Table 1) (51, 52) and results in an early death. Lysosomal storage disorders have been classified based on the clinical presentation, substrates stored, or similarities in the defect (53). For the purposes of this discussion, we have grouped some representative examples of lysosomal storage disorders according to the type of defect: mucopolysaccharidosis (defects in glycosaminoglycan degradation), oligosaccharidoses (defects in glycoprotein and glycolgen degradation), sphingolipidoses and lipidoses (defects in glycolipid degradation), and finally protein processing and transport defects (Table 1). Short stature, skeletal dysplasia, coarse facial features, joint problems, visceromegaly, cardiac disease, CNS (central nervous system) pathology, and early death are all common disease manifestations, but these features are not present in all disorders (Table 2). This finding implies that a common mechanism for pathology may occur in some disorders, but some unique features presumably relate to the type of substrate being stored and/or the relative organ distribution and turnover rate of the substrate.

Potentially compartment specific differences exist between each lysosomal storage disorder. For example, in Pompe disease, the storage of glycogen in endosome-lysosome organelles results from sequestration of cytoplasmic material and involves a phagosomal compartment. In contrast, other lysosomal storage disorders involve material that has undergone traffic from different endosome/phagosome compartments. Thus, compartment-specific storage effects may occur, although it may still result in a common molecular mechanism by impact at a certain point in the endosome-lysosome pathway. Some commonalities observed in lysosomal storage disorder pathology may result from substrate being turned over in similar organs. Thus, the rate of substrate turnover will impact directly on the level of storage and thus cell and organ dysfunction.

Primary and Secondary Storage Materials

Most known lysosomal storage disorders result from a single gene defect that results in the reduction of a single catabolic event, and hence the accumulation of a specific substrate. Although this concept is apparently simple, the reality is that the storage of a primary compound can result in a complex cascade of dysfunction (51). Glycosphingolipid substrates such as GM2 and GM3, and unesterified cholesterol can, for example, accumulate after the deposition of undegraded glycosaminoglycans, and this accumulation is evident in the lysosomal storage disorders mucopolysaccharidosis types I, II, and III. Thus, progressive disease manifestations such as skeletal dysplasia, heart disease, and CNS dysfunction may be more the result of this secondary storage than from the primary storage material. In turn, these manifestations may reflect a general disruption to lysosomal dysfunction and account for some commonalities in clinical presentation for lysosomal storage disorders.

Impact of Storage on Vesicular Structure and Traffic

Secondary storage in response to a primary defect and its associated stored substrate suggests a common process of lysosomal dysfunction in some storage disorders. Thus, GM2 and GM3 gangliosides have been observed to not only accumulate in GM2 gangliosidosis, Tay Sachs, and Sandhoff patients, but also in a range of other storage disorders including mucopolysaccharidosis types I, II, and III, and Niemann-Pick type C (54). In these disorders, vesicular structures with characteristic multilamellar inclusions, membrane swirls, and internal vesicles have been observed, which resemble either autophagosome or multivesicular endosome structures (52). Moreover, in mucopolysaccharidosis type III, the storage compartments have been reported to contain different amounts of the primary storage material heparin sulphate, GM2, and GM3 gangliosides (54). In addition, spheron structures that contain ubiquitin have been reported (54). Chloroquine toxicity also results in vesicular structures that have similar morphology, which include whorled inclusions.
### Table 1

A representative sample of 37 known (50 or more) lysosomal storage disorders and the common clinical symptoms observed in these patients at the severe end of the clinical spectrum (adapted from References 51 and 52).

<table>
<thead>
<tr>
<th>Lysosomal Storage Disorder (Syndrome)</th>
<th>Enzyme deficiency</th>
<th>Substrate Stored (common in severe form)</th>
<th>Clinical Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mucopolysaccharidosis I (Hurler, Hurler-Scheie and Scheie syndromes)</td>
<td>α-L-Iduronidase</td>
<td>HS, DS</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac disease, corneal clouding, CNS involvement</td>
</tr>
<tr>
<td>Mucopolysaccharidosis II (Hunter Syndrome)</td>
<td>Iduronate 2-sulphatase</td>
<td>HS, DS</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac disease, corneal clouding, CNS involvement</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IIIA (Sanfilippo A syndrome)</td>
<td>α-N-Acetylglucosaminidase</td>
<td>HS</td>
<td>Coarse hair, CNS involvement, aggressive behaviour, dysmorphic features (+/−), skeletal dysplasia (+/−)</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IIIB (Sanfilippo B syndrome)</td>
<td>α-N-Acetylglucosaminidase</td>
<td>HS</td>
<td>Coarse hair, CNS involvement, aggressive behaviour, dysmorphic features (+/−), skeletal dysplasia (+/−)</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IIIC (Sanfilippo C syndrome)</td>
<td>Heparan N-sulphatase</td>
<td>HS</td>
<td>Coarse hair, CNS involvement, aggressive behaviour, dysmorphic features (+/−), skeletal dysplasia (+/−)</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IIID (Sanfilippo D syndrome)</td>
<td>Heparan N-sulphatase</td>
<td>HS</td>
<td>Coarse hair, CNS involvement, aggressive behaviour, dysmorphic features (+/−), skeletal dysplasia (+/−)</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IV A (Morquio A syndrome)</td>
<td>N-Acetylgalactosamine 6-sulphatase</td>
<td>KS</td>
<td>Short stature, skeletal dysplasia, cardiac disease, corneal clouding</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IVB (Morquio B syndrome)</td>
<td>N-Acetylgalactosamine 6-sulphatase</td>
<td>KS</td>
<td>Short stature, skeletal dysplasia, cardiac disease, corneal clouding</td>
</tr>
<tr>
<td>Mucopolysaccharidosis VI (Maroteaux-Lamy syndrome)</td>
<td>N-Acetylgalactosamine 4-sulphatase</td>
<td>DS, CS</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac disease, corneal clouding</td>
</tr>
<tr>
<td>Mucopolysaccharidosis VII (Sly syndrome)</td>
<td>β-D-Glucuronidase</td>
<td>HS, DS</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac disease, corneal clouding</td>
</tr>
<tr>
<td>Mucopolysaccharidosis IX</td>
<td>Hyaluronidase</td>
<td>HA</td>
<td>Short stature, joint stiffness</td>
</tr>
</tbody>
</table>

**Oligosaccharidoses: Defects in glycoprotein and glycogen degradation**

<table>
<thead>
<tr>
<th>Enzyme deficiency</th>
<th>Substrate Stored</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-Mannosidase</td>
<td>α-Mannosides</td>
</tr>
<tr>
<td>β-Mannosidosis</td>
<td>β-Mannosides</td>
</tr>
</tbody>
</table>

(continued overleaf)
<table>
<thead>
<tr>
<th>Lysosomal Storage Disorder (Syndrome)</th>
<th>Enzyme deficiency</th>
<th>Substrate Stored</th>
<th>Clinical Symptoms (common in severe form)</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-L-Fucosidosis</td>
<td>α-Fucosidase</td>
<td>α-Fucosides Glycolipids</td>
<td>Skeletal dysplasia, joint stiffness, visceromegaly, cardiac disease, angiokeratoma, CNS involvement</td>
</tr>
<tr>
<td>Sialidosis</td>
<td>α-Sialidase</td>
<td>Sialyloligosaccharides</td>
<td>Short stature, skeletal dysplasia, CNS involvement</td>
</tr>
<tr>
<td>Galectosidosis</td>
<td>α-Sialidase, α-galactosidase, peptidase β-protein</td>
<td>Oligosaccharides</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac and renal disease, CNS involvement</td>
</tr>
<tr>
<td>Aspartylglucosaminuria</td>
<td>Aspartylglucosaminidase</td>
<td>Aspartylglucosamine</td>
<td>Skeletal dysplasia, visceromegaly, CNS involvement</td>
</tr>
<tr>
<td>Pompe disease</td>
<td>α-D-Glucosidase</td>
<td>Glycogen</td>
<td>Muscle weakness, cardiac disease</td>
</tr>
<tr>
<td>GM1 gangliosidosis</td>
<td>β-Galactosidase</td>
<td>GM1-gangliosides, oligosaccharides</td>
<td>Cardiac disease, CNS involvement</td>
</tr>
<tr>
<td>GM2 gangliosidosis (Sandhoff and Tay-Sachs)</td>
<td>Hexosaminidase A and B</td>
<td>GM2-gangliosides, oligosaccharides, glycolipids</td>
<td>CNS involvement, visual impairment</td>
</tr>
<tr>
<td>Fabry</td>
<td>α-Galactosidase</td>
<td>α-Galactosyl sphingolipids, oligosaccharides</td>
<td>Bone disease, visceromegaly, CNS involvement (severe form)</td>
</tr>
<tr>
<td>Niemann-Pick types A and B</td>
<td>Acid lipase</td>
<td>Cholesterol esters</td>
<td>Heart and kidney disease, angiokeratoma, CNS involvement (recurrent stroke)</td>
</tr>
<tr>
<td>Farber</td>
<td>Sphingomyelinase</td>
<td>Sphingomyelin</td>
<td>Visceroemgaly, CNS involvement</td>
</tr>
<tr>
<td>Krabbe</td>
<td>Galactosylceramidase</td>
<td>Galactosylceramides</td>
<td>Nodular swelling around joints, hypotonia, visceromegaly, CNS involvement, angiokeratoma</td>
</tr>
</tbody>
</table>

**Sphingolipidoses and lipidoses: Defects in glycolipid degradation**

<table>
<thead>
<tr>
<th>Lysosomal Storage Disorder (Syndrome)</th>
<th>Enzyme deficiency</th>
<th>Substrate Stored</th>
<th>Clinical Symptoms (common in severe form)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schindler, Kawasaki</td>
<td>α-Galactosidase</td>
<td>Galactosaminidase Glycolipids</td>
<td>Cardiac and renal disease, angiokeratoma, CNS involvement</td>
</tr>
<tr>
<td>Pompe disease</td>
<td>α-D-Glucosidase</td>
<td>Glycogen</td>
<td>Muscle weakness, cardiac disease</td>
</tr>
</tbody>
</table>

*Sphenoid metabolic diseases: Defects in sphingolipid degradation*
<table>
<thead>
<tr>
<th>Disorder</th>
<th>Location</th>
<th>Enzyme</th>
<th>Substrate</th>
<th>Clinical Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metachromatic leukodystrophy</td>
<td>Arylsulphatase</td>
<td>Sulphatides</td>
<td>Weakness, hypotonia, psychoses, ataxia, CNS involvement, behaviour changes</td>
<td></td>
</tr>
<tr>
<td>Multiple sulphatase deficiency</td>
<td>Endoplasmic reticulum sulphatase</td>
<td>Sulphatide modifying factors</td>
<td>MPS, sulphatides, glycolipids</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, CNS involvement</td>
</tr>
<tr>
<td>Mucolipidosis type I</td>
<td>Goji</td>
<td>1-sulphatase modifying factors</td>
<td>Oligosaccharides and glycolipids</td>
<td>Short stature, skeletal dysplasia, coarse facial features, joint stiffness, visceromegaly, cardiac disease, corneal clouding, CNS involvement</td>
</tr>
<tr>
<td>Mucolipidosis type II/III</td>
<td>Mucolipid 1, G.1+1 cation channel</td>
<td>PC, PL, MPS, SL, gangliosides</td>
<td>Cornal clouding, CNS involvement</td>
<td></td>
</tr>
<tr>
<td>Sialic acid storage disorder</td>
<td>Sialic acid transporter</td>
<td>Cysteine transporter</td>
<td>Sialic acid</td>
<td>Visceromegaly, renal disease, CNS involvement</td>
</tr>
<tr>
<td>Mucolipidosis type IV</td>
<td>Cysteine transporter</td>
<td>Cysteine transporter</td>
<td>Cysteine</td>
<td>Renal disease, CNS involvement</td>
</tr>
<tr>
<td>Hermansky-Pudlak syndrome (disorders of lysosomal biogenesis)</td>
<td>AP3 adaptor protein</td>
<td>Cereolipid</td>
<td>Ceroid lipofuscin</td>
<td>Lung and kidney disease, oculocutaneous albinism, CNS involvement</td>
</tr>
<tr>
<td>Fronto-temporal dementia</td>
<td>Endosomal sorting complex</td>
<td>Required for transport (ESCRT)</td>
<td>CNS involvement</td>
<td></td>
</tr>
</tbody>
</table>

HS, heparin sulphates; DS, dermatan sulphates; KS, keratan sulphate; HA, hyaluronic acid; PC, phosphatidylcholine; PL, phospholipid; SL, sphingolipid.
of lipid, multivesicular endosome-like structures and zebra bodies. In turn, this toxicity has been reported to be almost identical to the vesicular pathology in Fabry disease (55). This commonality of vesicular structures and pathology could be explained by the general impact on lysosomal function created by the primary storage material. Although these examples involve an enzyme deficiency and could therefore tend to support this concept, other lysosomal storage disorders involve dysfunction in either membrane transporters or vesicular machinery, and they generate similar vesicular pathology. For example, a defect in the lysosomal transporter mucolipin-1 (the cation channel that is involved in calcium export and necessary for the recovery of lysosomes from late endosome–lysosome hybrid organelles) causes the storage of lipids and gangliosides in enlarged multivesicular lamellar structures. Moreover, a dysfunction of the ESCRT-III vesicular machinery, which results in neuropathology, has also been reported to cause the formation of similar multivesicular structures in a mouse model and these vesicles are almost identical to that reported in a mouse model of mucopolysaccharidosis IIIA (58). Possible explanations for these commonalities in vesicular pathology might therefore include the impact of storage on a critical event in lysosomal function, such as vesicle formation/recovery or vesicular traffic.

**Treatment Strategies and Ability to Correct Residual Pathology**

Treatment strategies are available for lysosomal storage disorders and include hematopoietic stem cell transplantation, enzyme replacement therapy, substrate reduction therapy, chemical chaperones, and gene therapy (recently reviewed in References [57–61]). In practice, however, the clinical spectrum observed in lysosomal storage disorders means that—at the present time—a single therapeutic strategy is unlikely to treat all sites of pathology effectively. Thus, hematopoietic stem cell marrow transplantation is used currently for many patients with neuropathology, but it is not optimal (not effective for some disorders) and has significant risks. Hematopoietic stem cell transplantation is currently recommended for patients at risk of cognitive impairment and has been evaluated in combination with enzyme replacement therapy (62). Enzyme replacement therapy by intravenous infusion is being used for patients at the attenuated end of the clinical spectrum, in some lysosomal disorders (e.g., MPS-I [mucopolysaccharide] (58, 62)), but it is of limited use for the treatment of those disorders with neuropathology. Recently, small molecule therapeutic strategies have been employed as potential alternatives or adjunct treatment strategies (reviewed in Reference 63). For example, substrate deprivation therapy uses small molecule inhibitors to reduce substrate synthesis, whereas enzyme enhancement therapies have been investigated using chemical chaperones (63) to improve the folding of mutant protein and thus enhance the level of residual enzyme activity in patient cells. Both new therapeutic strategies have been investigated in Gaucher disease (64), and substrate reduction therapy is in clinical practice. The complex pathology that can result from lysosomal storage and particularly to clear residual pathology. Thus, in general, the earlier the treatment is implemented the more likely an effective therapeutic outcome will be achieved. The optimum treatment strategy to cure lysosomal storage disorders may be gene replacement therapy, but it is still in the developmental stages and will likely be a long way from widespread clinical practice. Nonetheless, therapeutic strategies that result in significant improvement in the quality of life for lysosomal storage disorder patients are in clinical practice.

**References**

Lysosomal Disorders

34. Harzer K, Masarokeni G, Frolich E. Concurrent increase of cholesterol, sphingomyelin and glucosylceramide in the spleen from non-neurologic Niemann–Pick type C patients but also patients possibly affected with other lipid-traﬁcking disorders. FEBS Lett. 2003;537:177–181.
42. Bright NA, Reaves BJ, Mallick BM, Luzio JP. Dense core lysosomes can fuse with late endosomes and are re-formed from the resultant hybrid organelles. J. Cell Sci. 1997;110:2021–2040.
Further Reading


See Also

Glycosylation of Proteins in the Golgi Apparatus
Chaperones, Molecular and Chemical Membrane Trafficking
Lysosomal Trafficking
Endocytosis, Receptor-mediated
Glycolipids, Synthesis of Lysosome, Topics in Chemical Biology
Obesity, type 2 diabetes (T2D), lipid disorders, and hypertension are chronic and disabling diseases that afflict hundreds of millions of individuals worldwide. In this article, they are collectively referred to as "cardiometabolic diseases," because they have "common ground." They increase the risk of cardiovascular disease (CVD) morbidity and mortality. Considerable research has been performed to understand the etiology of cardiometabolic diseases and to translate this research into effective treatment paradigms. However, it has been challenging to understand the initiation and progression of cardiometabolic diseases. This difficulty is attributed to the complexities involved in metabolic regulation, a resultant "snowballing" effect, and a downward spiral as disease progresses. The first line of therapy for treating obesity, T2D, and/or dyslipidemia is lifestyle intervention, which can impact significantly on disease and decrease CV risk, but is usually not effective because of lack of patient compliance. Patients are then given a plethora of drugs to treat the individual risk factors. However, new therapies are critically needed, because most patients require multiple drugs, some drugs may treat one CVD risk factor and exacerbate another, and most patients do not meet treatment goals. A "magic bullet" that treats the underlying causes that contribute to cardiometabolic risk has been a longstanding goal but remains elusive. Ongoing research to uncover novel targets holds promise for future therapeutics that might treat multiple facets of cardiometabolic diseases. This introductory review focuses on the epidemiology and etiology of cardiometabolic diseases, current therapies, and future treatment strategies.

It is difficult to separate obesity, type 2 diabetes (T2D), dyslipidemia, and hypertension because they are a consequence of a dysregulation in metabolism, are interrelated, and each might drive disease progression of the other. These comorbidities are usually found as a cluster within patients, but each can occur in the absence of the others. Lifestyle (high caloric intake, low physical activity, and cigarette smoking) along with a genetic predisposition result in alterations in metabolism that lead to cardiometabolic diseases and increase the risk of CVD. The multiple factors that contribute to CVD are summarized in Fig. 1 (1). Recent studies have provided insight into the mechanisms by which metabolic abnormalities impair insulin-receptor signaling, trigger inflammation and endothelial dysfunction, and increase CVD risk. This finding might provide new opportunities to treat the multiple risk factors that result in insulin resistance and cardiovascular disease. (Author’s note: The author recognizes that numerous "landmark" studies have been published that have contributed to the understanding of cardiometabolic diseases, but she has chosen to reference newer publications and review articles in this introductory overview. Also, a list of abbreviations is provided at the end of this review.)

The Global Crisis in Obesity, Diabetes, and CV Risk

CVD is attributed to disorders of the heart and blood vessels, and it includes coronary heart disease (heart attacks), stroke, hypertension, peripheral artery disease, rheumatic heart disease, congenital heart disease, and heart failure. On a global perspective, 30% of all deaths are from CVD, which ranks it the number
Obesity

Overweight and obesity occurs when an energy imbalance exists within the body, in which energy intake exceeds energy used, resulting in an increase in body fat. The cause of an energy imbalance for each individual may be different, and it is the result of interplay between environmental and genetic factors, which makes it a complex disease to understand and treat. Although some societies might be more genetically prone to obesity, changes in environmental factors, particularly the quantity and quality of food consumed and a sedentary lifestyle, are the key drivers for the global obesity epidemic (3).

For adults, overweight and obesity are easily defined by the body mass index (BMI). This value is calculated from a person’s weight in kilograms divided by the square of the person’s height in meters (kg/m²); conversion tables are available for body weight in pounds and height in inches. An adult who has a BMI between 25 and 29.9 is considered overweight; an adult who has a BMI of 30 or higher is considered obese (4). Estimates of overweight and obesity for children and adolescents take age and gender into consideration.

Diabetes

Epidemiologic studies have established a strong relationship between obesity and the risk for T2D (8, 9). Therefore, because of the obesity epidemic, diabetes rates are soaring. WHO data indicate that over 240 million people have diabetes worldwide; the number is expected to reach 380 million by 2025 (5). The five countries with the largest numbers of people with diabetes are India, China, the United States, Russia, and Germany. For the U.S. population, 20.6 million people or 7% of the population have diabetes. Adults with diabetes have heart disease and risk of stroke about 2 to 4 times higher than adults without diabetes (10, 11). The CDC reports that diabetes is the sixth leading cause of death among adult Americans, and that two thirds of diabetics die from CVD and stroke (6). Diabetes is the major cause of renal disease, adult blindness, and lower-limb amputation. The economic burden of diabetes in the United States totals $174 billion annually; this figure has increased by 32% since 2002 (12).

Diabetes can manifest in several forms, and this review will focus on T2D, which affects 90–95% of all diabetics. (For an overview of diabetes diagnosis, classification, and standard of care, see References 12–15.) T2D is a metabolic disorder resulting from an interplay between environmental and genetic factors, which makes it a complex disease to understand and treat.
that is attributed to a defect in the secretion of insulin by the β-cells of the pancreas in response to metabolic signals, combined with the inability of cells to respond to insulin (i.e., insulin resistance), which results in impaired nutrient uptake and use and increased hepatic glucose output (Fig. 2). T2D is considered a “silent” disease, which progresses over decades, and it is usually diagnosed after a complication is evident or a cardiovascular event occurs.

A biopsy studies in humans have indicated that a curvilinear relationship is observed between β-cell mass and fasting blood glucose concentration, and a steep increase in blood glucose concentration is associated with β-cell deficiency (16). The data have provided research incentives focused on restoring β-cell mass to reverse diabetes or preventing diabetes by protection of β-cell mass. Abnormalities in β-cell function are therefore critical in defining the risk and development of T2D, particularly in obese subjects (17). The adipose tissue of obese individuals releases high levels of nonesterified fatty acids (NEFA), glycerol, hormones, proinflammatory cytokines, and other factors that increase insulin resistance (Fig. 2). When insulin resistance is accompanied by impaired insulin secretion by β-cells, blood glucose levels are not controlled, which results in overt diabetes.

Diabetes is diagnosed by a Fasting Plasma Glucose Test (FPG) or an Oral Glucose Tolerance Test (OGTT) (18). A FPG level of ≥ 126 mg/dL indicates diabetes. In the OGTT test, a blood glucose level 2 hours after drinking a glucose solution of ≥ 200 mg/dL indicates diabetes. A stable measure of long-term glucose status is to determine the percent of a glycosylated isoform of hemoglobin in blood (termed % HbA1c), which is a slowly turning over protein within the body that is modified by high blood glucose. Current treatment guidelines are to reduce HbA1c to <7% (13), because it reduces the patient’s risk of developing microvascular complications (nephropathy, retinopathy, and neuropathy) as well as macrovascular disease (cardiovascular disease and stroke) (19).

Dyslipidemia and hypertension

Abnormal circulating lipids (i.e., high LDL concentration (LDLc), low HDL concentration (HDLc), hypertriglyceridemia, Lp(a), and small dense LDL), and hypertension are strongly associated with CVD risk (2,20–22). These lipids are often found in obese and/or diabetic patients. Data from the Framingham Heart Study is used to estimate a person’s 10-year risk for “hard” CHD outcomes (myocardial infarction and coronary death), in which risk is calculated based on age, gender, total and HDL cholesterol, systolic blood pressure, and whether the individual is on blood-pressure-lowering medications or is a smoker. Numerous primary and secondary intervention studies have shown that improving lipid profile and lowering blood pressure significantly reduces disease morbidity and mortality (22, 23). Improved control of blood lipids in diabetics can reduce cardiovascular events by 20–50% (15, 24). In addition to lipids and blood pressure, inflammation and a procoagulant and prothrombotic state are additional CVD risk factors that need to be addressed, particularly in obese and insulin-resistant patients (25, 26). Results from a subgroup analysis of the Pravastatin or Atorvastatin Evaluation Trial (PROVE IT) (27) showed that intensive lipid-lowering intervention significantly reduced acute coronary events in diabetics, but most patients did not reach the dual goal of LDLc <70 mg/dL and high sensitivity C-reactive protein <2 mg/L. (CRP is an inflammatory marker associated with CV risk). The data highlight the need for additional strategies in this high-risk group, particularly those that target inflammation.

Figure 2 Model for the critical role of impaired insulin release in linking obesity with insulin resistance and T2D. Impaired insulin secretion results in decreased insulin levels and decreased signaling in the hypothalamus, which leads to increased food intake and weight gain, decreased inhibition of hepatic glucose production, reduced efficiency of glucose uptake in muscle, and increased lipolysis in the adipocyte. These results lead to increased plasma NEFA levels. The increase in body weight and NEFA contribute to insulin resistance, and the increased NEFA suppresses the β-cell’s adaptive response to insulin resistance. The increased glucose levels together with the elevated NEFA levels can synergize to affect β-cell health and insulin action adversely, which is often referred to as “glucolipotoxicity.” From Reference 17, with permission.
Metabolic syndrome versus additive cardiometabolic risk factors

Cardiometabolic diseases may present as separate diseases but more often cluster within patients. Because of this clustering, the concept of "The Metabolic Syndrome" (also termed insulin resistance syndrome or syndrome X) has been put forth. Several definitions of the Metabolic Syndrome are available, with the overall viewpoint that multiple interrelated risk factors increase the risk for atherosclerotic cardiovascular disease and increase the risk for T2D. Guidelines to define the metabolic syndrome have been developed by the WHO, National Cholesterol Education Program-Adult Treatment Panel (NCEP-ATP), and International Diabetes Federation (IDF), which are reviewed in Reference 28. Each definition includes traditional risk factors: obesity and an abdominal fat distribution, insulin resistance, diabetes, dyslipidemia (high TG, low HDL) and hypertension.

Shown in Table 1 are the new IDF definition (29) and the revised National Cholesterol Education Program-Aduit Treatment Panel III (NCEP-ATP IIII) definition (30), which are the most widely used. Each definition has its strengths and weaknesses, and both groups are working toward harmonizing criteria (31).

According to CDC statistics that use the NCEP criteria (32), the metabolic syndrome is found in 20% of the U.S. adult population. Its prevalence increases with aging, and it is highest in Hispanic women with 35% prevalence. According to the WHO, 25% of the world's population and at least two thirds of diabetics have the metabolic syndrome. Considerable evidence suggests that the metabolic syndrome is a significant predictor of CVD and T2D, and risk increases with increasing number of risk factors (33). However, a recent assessment (34) states that "the Metabolic Syndrome is a stronger predictor of T2D than CHD and it does not predict CHD as well as the Framingham Risk Score," but it serves as a simple clinical tool for identifying high-risk subjects predisposed to CVD and T2D.*

The clinical significance of the metabolic syndrome came into question in 2005, after "A Critical Appraisal of the Metabolic Syndrome" that was jointly written by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD) (35). These organizations questioned the clinical value of diagnosing the metabolic syndrome, stating that CV risk is not greater than the sum of the individual risk factors, and treatment of the syndrome is no different than the treatment for each of its components. Their recommendation was to continue efforts to understand the relationships of risk factors that contribute to CVD. Numerous rebuttals to the ADA/EASD statement have been given, particularly by the American Heart Association (AHA) (36). In 2007, the AHA and ADA jointly issued a publication that attempted to harmonize the recommendations of both organizations where possible, and recognized areas where they differ (24). Overall, it is feasible that common ground will be reached in the near future.

**Underlying Mechanisms that Lead to Insulin Resistance and CVD**

Chronic excessive nutrient intake leads to the deposition of fat, in not only its normal storage site, which is the adipose tissue, but also in liver and skeletal muscle. Nutrient excess also triggers an inflammatory response, with the release of inflammatory cytokines (tumor necrosis factor-alpha (TNF-α), interleukin-6 (IL-6), and CRP). These inflammatory mediators, along with the intracellular accumulation of lipid metabolites, lead to impaired insulin receptor signaling and defective metabolism in skeletal muscle and liver (37, 38). Nutrient excess also damages cells by generating reactive oxygen species, which results in not only its normal storage site, which is the adipose tissue, but also in liver and skeletal muscle. Nutrient excess also triggers an inflammatory response, with the release of inflammatory cytokines (tumor necrosis factor-alpha (TNF-α), interleukin-6 (IL-6), and CRP). These inflammatory mediators, along with the intracellular accumulation of lipid metabolites, lead to impaired insulin receptor signaling and defective metabolism in skeletal muscle and liver (37, 38). Nutrient excess also damages cells by generating reactive oxygen species, which results

### Table 1: The IDF and NCEP definitions of the metabolic syndrome

<table>
<thead>
<tr>
<th>International Diabetes Federation</th>
<th>National Cholesterol Education Program-ATP III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central obesity: waist circumference-ethnicity specific plus any two of the following:</td>
<td>Any three of the five criteria: Elevated waist circumference for US population: &gt;102 cm (40 in) in men, &gt;88 cm (&gt;35 in) in women lower cut points for insulin resistant individuals or ethnic groups</td>
</tr>
<tr>
<td>High TG: &gt;150 mg/dL (1.7 mmol/L)</td>
<td>High TG: &gt;150 mg/dL (1.7 mmol/L)</td>
</tr>
<tr>
<td>Low HDL cholesterol: &lt;40 mg/dL (1.03 mmol/L) in men, &lt;50 mg/dL (1.3 mmol/L) in women</td>
<td>Low HDL cholesterol: &lt;40 mg/dL (1.03 mmol/L) in men, &lt;50 mg/dL (1.3 mmol/L) in women</td>
</tr>
<tr>
<td>High blood pressure: &gt;130 mmHg systolic, &gt;85 mmHg diastolic</td>
<td>High blood pressure: &gt;130 mmHg systolic, &gt;85 mmHg diastolic</td>
</tr>
<tr>
<td>Elevated plasma glucose: Fasting plasma glucose &gt;100 mg/dL (5.6 mmol/L) or previously diagnosed T2D; if above 5.6 mmol/L, an oral glucose tolerance test is strongly recommended but not necessary to define the presence of the syndrome.</td>
<td>Elevated fasting glucose: &gt;100 mg/dL (5.6 mmol/L)</td>
</tr>
</tbody>
</table>

**Limitations:** Central obesity is required, criteria and cutoff values might need to be further defined

*Information was obtained from References 29 and 30.

* Central obesity not necessary if three of the other risk factors are present.
in protein modifications and in the depletion of nitric oxide that maintains vascular tone. These changes lead to cell dysfunction, alterations in blood lipids, elevated blood pressure, coagulation; fibrinolysis, and additional inflammation, which drive the downward spiral of insulin resistance, endothelial dysfunction, and atherosclerosis (39).

**Metabolic stress and impaired cellular function**

Nutrient excess leads to the intracellular accumulation of long chain acyl-CoA and diacylglycerol, and in the activation of several serine/threonine kinases, which include protein kinase C isoforms, inhibitor of kappa B kinase (IKK), and c-jun N-terminal kinase (JNK) (37). A key step in insulin receptor signaling is the tyrosine phosphorylation of IRS1 and 2, which regulates carbohydrate, lipid, and protein metabolism. The ac-
tivation of several serine/threonine kinases results in serine phosphorylation of IRS1 and 2, which inhibits insulin receptor signaling. The activation of the two principal inflammatory pathways, IKKα/IKKβ and JNK, by nutrient excess and inflammatory cytokines, also impairs insulin receptor signaling and propagates the stress response (37). In skeletal muscle, it results in impaired insulin action on cell metabolism, which includes decreased insulin-stimulated glucose transport and decreased glycogen synthesis, leading to impaired glucose use and hyperglycemia. In the liver, glycogen synthesis is decreased, and gluconeogenesis is stimulated, which results in an increase in hepatic glucose production and hyperglycemia (38, 40).

Recently, the sphingolipid ceramide, which is a product of fatty acyl-CoA, has been identified as the link between excess nutrients (i.e., saturated fatty acids) and inflammatory cytokines (i.e., TNFα), to the induction of insulin resistance. Moreover, ceramide has been shown to be toxic to pancreatic β-cells, cardiomyocytes, and endothelial cells, which contributes to diabetes, hypertension, cardiac failure, and atherosclerosis (41). However, the role of ceramide in mediating insulin-resistance humans is still unclear (42).

The endoplasmic reticulum (ER) is a network of membranes in which secreted and membrane proteins are assembled into their secondary and tertiary structures. The ER seems to be the site for sensing metabolic stress and to translate this stress into inflammatory signals (43). Under certain stress conditions, such as energy excess, lipids, and pathogens, the ER activates a complex response system known as the unfolded protein response to slow down protein synthetic pathways and to restore function integrity to the organelle. Data from experimental models have shown that obesity leads to ER stress, which activates both JNK and IKK, and initiates pathways that trigger inflammation and insulin resistance. In the pancreatic β-cell, ER stress impairs insulin secretion and contributes to progression of T2D.

**Metabolic Diseases, Biological Mechanisms of**

**Inflammation**

Obesity, and particularly the accumulation of abdominal fat, creates an inflammatory milieu that is the key driver of insulin resistance and CVD (44–46). In the normal state, adipose tissue coordinates regulates the synthesis and secretion of peptides that regulate numerous processes in the body (47, 48), which include fat mass, nutrient homestasis and energy expenditure, the immune response, blood pressure control, hemostasis, bone mass, and reproductive function. In obesity, adipose tissue inflammation results in the secretion of proinflammatory peptides and reduction of anti-inflammatory peptides, which lead to dele-
terious effect on the liver, muscle, and the vasculature (Fig. 3). A reduction in abdominal fat improves the atherogenic lipid pro-
file, reduces inflammation, and decreases blood pressure, which thereby decreases CV risk (44, 49). Current strategies are fo-
cused on identifying additional inflammatory markers that put the patient at cardiometabolic risk.

**Chronic hyperglycemia induces numerous alterations in the vasculature that accelerate the atherosclerotic process. Several major mechanisms contribute to the pathological alterations in blood vessels in diabetes, including:** 1) the nonenzymatic glycosyl-
lation of proteins and lipids, which form advanced glycation endproducts (AGEs) that can interfere with their normal func-
tion; and 2) the induction of oxidative and nitrosative stress, as well as exacerbation of proinflammatory responses (50). These abnormalities lead to impaired endothelial platelet inhibition and platelet activation, which could result in arterial thrombosis, and consequently myocardial infarction and stroke (51).

**Impaired endothelial function**

Because of metabolic stress, the endothelium loses its ability to balance vasodilating and vasoconstricting factors to maintain homestasis. Nitric oxide (NO) is the most important mediator of vasodilation, and loss of NO bioavailability contributes to the loss of vessel tone and damage to the endothelium. This dam-
age is particularly evident in the insulin-resistant state, in which insulin plays a key role in maintaining endothelial function and stimulating NO production. In diabetes, defective insulin signaling in the endothelial cell results in an imbalance be-
 tween the vasodilating agent NO and the vasoconstricting agent endothelin-1, which results in additional endothelial dysfunction and hypertension (52). Furthermore, markers of an activated endothelium appear prior to the presentation of overt diabetes, which suggests that the endothelium plays a primary role in the disease process.

A dysfunctional endothelium makes it susceptible to dam-
age by adhesion molecules, inflammatory cytokines, activated platelets, and lipids, which culminates in atherosclerosis (46, 53). In the kidney, endothelial dysfunction impairs glomerular filtration and leads to the progressive loss in renal function. En-
dothelial dysfunction can be addressed by treating the patient with angiotensin-converting enzyme inhibitors (ACEI) and an-
giotensin II receptor blockers (ARBs), which block the deleteri-
owns effects of an activated renin-angiotensin-aldosterone system (RAAS) on the endothelium (54). Reviews of endothelial dam-
age and current treatment strategies can be found in References 53 and 55. Novel therapies are needed to prevent progressive endothelial damage, whereas early intervention might improve endothelial function, offering an opportunity to protect against both CVD and organ damage. The consequences of metabolic dysregulation on endothelial damage and organ injury that culminate in CVD are summarized in Fig. 4.
Figure 3. Adipokine expression and secretion by adipose tissue in insulin-resistant, obese subjects. Obesity results in adipose tissue inflammation with macrophage infiltration. This result leads to 1) a decrease in adiponectin, which is an anti-inflammatory adipokine, that is positively correlated with insulin sensitivity and plays a protective role on the vasculature; and 2) an increase in inflammatory cytokines (TNFα, IL-6, and resistin) which causes insulin resistance, inflammation, and atherosclerosis. From Reference 47 with permission.

Figure 4. The contribution of impaired metabolism to cardiovascular risk. A dysregulation in nutrient metabolism contributes to cardiovascular risk by the following mechanisms: 1) impaired β-cell function, tissue damage, and insulin resistance; 2) an atherogenic lipid profile that is characterized by high TG, low HDL, and abnormal lipoproteins; 3) hypertension and vascular dysfunction; and 4) alterations in cytokines and adipokines that lead to a proinflammatory and procoagulant state.
Current and Next Generation Therapeutics

Numerous drugs are marketed to treat metabolic diseases, which include weight reducing agents, antidiabetics, lipid-lowering treatments, and antihypertensives. The pharmacology, positive attributes, efficacy, and limitations for each drug class are summarized in Tables (2-5) (56, 57).

Weight-loss agents

The three currently marketed weight-loss drugs are listed in Table 2. Orlistat is a pancreatic lipase inhibitor that acts at the level of the gastrointestinal (GI) tract to inhibit the absorption of dietary fat. Sibutramine and rimonabant act at different transporters or receptors to modulate central and peripheral metabolism. The paucity of weight loss agents that are available is attributed to the hurdles in developing weight-loss drugs. Large placebo-controlled trials are required; the placebo group is given a low-calorie diet and exhibits weight loss. A variation is observed in response and a significant number of nonresponders, and safety considerations often emerge. Most weight-loss treatments are associated with improvement in insulin sensitivity and lipids, but the drugs have side effects that generally limit use.

Antidiabetic agents

Current marketed antidiabetic treatments are shown in Table 3. These drugs may act to: improve insulin sensitivity, decrease hepatic glucose output, decrease the absorption of glucose in the GI tract, or stimulate the secretion of insulin. Numerous studies have shown that antidiabetic and weight-loss interventions, either by lifestyle or by drug treatment, prevent the development of diabetes in patients at risk for developing that disease (58). However, antidiabetic and weight-loss drugs are not prescribed as preventative agents. Because diabetes is a complicated disease, a treatment paradigm has been developed, and multiple treatments are generally required (57). However, a patient’s diabetes still progresses even when given several antidiabetic drugs in combination. This progression seems to be caused by the progressive loss of β-cell function despite multiple treatments (59).

Orlistat (oral) is a pancreatic lipase inhibitor that acts at the level of the gastrointestinal (GI) tract to inhibit the absorption of dietary fat. Sibutramine and rimonabant act at different transporters or receptors to modulate central and peripheral metabolism. The paucity of weight loss agents that are available is attributed to the hurdles in developing weight-loss drugs. Large placebo-controlled trials are required; the placebo group is given a low-calorie diet and exhibits weight loss. A variation is observed in response and a significant number of nonresponders, and safety considerations often emerge. Most weight-loss treatments are associated with improvement in insulin sensitivity and lipids, but the drugs have side effects that generally limit use.

Antidiabetic agents

Current marketed antidiabetic treatments are shown in Table 3. These drugs may act to: improve insulin sensitivity, decrease hepatic glucose output, decrease the absorption of glucose in the GI tract, or stimulate the secretion of insulin. Numerous studies have shown that antidiabetic and weight-loss interventions, either by lifestyle or by drug treatment, prevent the development of diabetes in patients at risk for developing that disease (58). However, antidiabetic and weight-loss drugs are not prescribed as preventative agents. Because diabetes is a complicated disease, a treatment paradigm has been developed, and multiple treatments are generally required (57). However, a patient’s diabetes still progresses even when given several antidiabetic drugs in combination. This progression seems to be caused by the progressive loss of β-cell function despite multiple treatments (59).

Orlistat (oral) is a pancreatic lipase inhibitor that acts at the level of the gastrointestinal (GI) tract to inhibit the absorption of dietary fat. Sibutramine and rimonabant act at different transporters or receptors to modulate central and peripheral metabolism. The paucity of weight loss agents that are available is attributed to the hurdles in developing weight-loss drugs. Large placebo-controlled trials are required; the placebo group is given a low-calorie diet and exhibits weight loss. A variation is observed in response and a significant number of nonresponders, and safety considerations often emerge. Most weight-loss treatments are associated with improvement in insulin sensitivity and lipids, but the drugs have side effects that generally limit use.
It is cosecreted with insulin by another regulatory hormone involved in glucose homeostasis. via sitagliptin, which is an oral DPP-IV inhibitor. Amylin is tentatively endogenous GLP-1 action by inhibiting its degradation to inject a GLP-1 protein mimetic (i.e., exenatide) or 2) to physiological treatments to promote GLP-1 action are as follows: 1) central effect to inhibit food intake. Two approved pharmacological treatments to promote GLP-1 action are as follows: 1) to inject a GLP-1 protein mimetic (i.e., exenatide) or 2) to potentiate endogenous GLP-1 action by inhibiting its degradation via sitagliptin, which is an oral DPP-IV inhibitor. Amylin is another regulatory hormone involved in glucose homeostasis. It is cosecreted with insulin by -cells and decreases PPG excursions by slowing gastric emptying and decreasing glucagon secretion. Pramlintide, which is an injectable analog of amylin, is approved to treat both T1 and T2D patients.

Lipid-lowering agents

Many lipid-lowering agents are used alone or in combination to achieve lipid goals (Table 4). Although numerous lipid-lowering agents are available, which are used alone and in combination, additional therapies are required to decrease other atherogenic components and to enhance reverse cholesterol transport. The "statins" inhibit HMGCoA reductase, which results in an inhibition of cholesterol synthesis and upregulation of LDL clearance. This result leads to a marked reduction in LDLc, with additional positive effects to decrease TG modestly and to increase HDLc. Bile acid sequestrants may elicit an antidiabetic effect (61); a newer seqeustrant colesevelam HCl received approval to treat both T2D and high LDL cholesterol.

Table 3 Currently prescribed anti-diabetic agents

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Pharmacology and positive attributes</th>
<th>Expected efficacy</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin (injectable or inhaled protein, multiple products)</td>
<td>Increases insulin levels, improves glucose tolerance, improves lipid profile.</td>
<td>HbA1c reductions of 1.5–3.5% (less reduction for injected vs injectable)</td>
<td>May cause hypoglycemia, weight gain</td>
</tr>
<tr>
<td>Metformin (oral)</td>
<td>Decreases hepatic glucose output plus additional mechanisms, improves glycemic control, ↓TG and HDLc</td>
<td>HbA1c reductions of 1.0–2.0%</td>
<td>May cause GI problems, lactic acidosis (rare)</td>
</tr>
<tr>
<td>Sulfonylureas and non-SU secretagogues (oral, multiple agents)</td>
<td>Increases insulin secretion by pancreatic beta cells, improves glycemic control</td>
<td>HbA1c reductions of 1.0–2.0% for sulfonylureas</td>
<td>May cause hypoglycemia, weight gain</td>
</tr>
<tr>
<td>α-Glucosidase inhibitors (oral, multiple agents)</td>
<td>Delays GI absorption of carbohydrates, improves glycemic control</td>
<td>HbA1c reductions of 0.5–0.8%</td>
<td>May cause GI problems</td>
</tr>
<tr>
<td>Thiazolidine-diones (oral, rosiglitazone, pioglitazone)</td>
<td>PPARγ agonist, increases insulin sensitivity and improves glycemic control, ↓TG (rosiglitazone may ↑TG), ↑HDLc and ↓LDLc (pioglitazone has greater beneficial effects on blood lipids)</td>
<td>HbA1c reductions of 0.5–1.4%</td>
<td>Associated with weight gain, edema, not recommended for patients with CHF</td>
</tr>
<tr>
<td>Exenatide (injectable protein)</td>
<td>Long acting GLP-1 analog, improves glycemic control, decreases TG and increases HDLc, causes weight loss.</td>
<td>HbA1c reductions of 0.5–1.0%</td>
<td>Associated with GI problems</td>
</tr>
<tr>
<td>Sitagliptin (oral)</td>
<td>Inhibits DPP-IV thereby potentiliating the action of GLP-1; improves glycemic control, weight neutral.</td>
<td>HbA1c reductions of 0.5–0.8%</td>
<td>Recently approved, little experience</td>
</tr>
<tr>
<td>Pramlintide (injectable protein)</td>
<td>Amylin analog, improves glycemic control, causes weight loss.</td>
<td>HbA1c reductions of 0.6%</td>
<td>A associated with GI problems</td>
</tr>
</tbody>
</table>

Information was obtained from References 56 and 57.

New targets include potentiating the effects of glucagon-like peptide-1 (GLP-1) and amylin, which are reviewed in Reference 60. GLP-1 is a key regulatory hormone that is secreted by the L-cells of the intestine. This hormone stimulates intestinal glucose absorption, inhibits gastric emptying and glucagon secretion, and has a central effect to inhibit food intake. Two approved pharmacological treatments to promote GLP-1 action are as follows: 1) to inject a GLP-1 protein mimetic (i.e., exenatide) or 2) to potentiate endogenous GLP-1 action by inhibiting its degradation via sitagliptin, which is an oral DPP-IV inhibitor. Amylin is another regulatory hormone involved in glucose homeostasis. It is cosecreted with insulin by β-cells and decreases PPG excursions by slowing gastric emptying and decreasing glucagon secretion. Pramlintide, which is an injectable analog of amylin, is approved to treat both T1 and T2D patients.
the "statin" dose to reduce side effects. However, the "Enhance" clinical trial results (62) have raised concerns about the lack of vascular benefit with ezetimide, and additional outcome trials are ongoing.

The fibrates act at the level of the liver to decrease VLDL-TG substantially and to increase HDLC secretion. Nicotinic acid also decreases TG and is considered to be the most potent agent to increase HDLC. The effects of fibrates and nicotinic acid on TG and HDLC are generally greater than those observed with "statins," which are generally the most effective LDLc-lowering agents.

Blood pressure agents

Several classes of blood-pressure-lowering agents are available and are often used in combination to achieve goals (Table 5). Results of large clinical trials have indicated that the benefits of antihypertensive treatment is caused by the lowering in blood pressure and largely independent of the drugs employed (22). A challenge has been to understand why differential responses to a specific class or combination are observed.

Three antihypertensives (ACEI, ARB, and renin blockers) block different steps the RAAS pathway, which results in a downstream blockade of AT1 receptors that result in vasodilation, decreased secretion of vasopressin, and decreased secretion of aldosterone, contributing to a blood pressure-lowering effect. The decrease in aldosterone decreases sodium and water resorption in the kidney and decreases potassium excretion, which leads to a lowering in blood pressure. These drugs are often used to treat hypertension, diabetics nephropathy, and congestive heart failure (63). ACEI also blocks the bradykinin pathway, which induces nitric oxide and vasodilation, but it is often associated with the persistant dry cough and/or angioedema that may limit ACEI therapy. This side effect is rarely observed with ARBs. Although ACEI and renin antagonism decrease circulating angiotensin II, ARBs block its activity at the AT1 receptor. ARBs increase angiotensin II levels by uncoupling the negative-feedback loop, and increase its stimulation of AT2 receptors, which is associated with beneficial and negative effects. A direct renin inhibitor has recently entered the market, which may result in more complete inhibition of the RAAS system inhibition than with ACEI or ARBs, but more clinical experience and outcome trial results are necessary to assess its potential adequately (64).

Beta adrenergic receptor antagonists reduce cardiac output (caused by negative chronotropic and inotropic effects), decrease renin release from the kidneys, and cause smooth muscle relaxation. However, blockage may also decrease secretion of insulin from pancreatic β-cells, which limits its use in T2D. Calcium channel antagonists act on L-type voltage gated channels in the heart and blood vessels to reduce vascular resistance and arterial pressure. Diuretics are also widely used to decrease blood pressure, particularly in the elderly and hypertensive black populations.

Future treatment strategies

Future treatments are focused on correcting the metabolic dysregulation, which contributes to cell damage and tissue dysfunction. A list of selected novel targets for treating cardiometabolic diseases is shown in Table 6. The focus for new antibiostasis approaches has been to target the gut-brain axis to regulate feeding behavior and energy expenditure, as well as modulating peripheral metabolism (65). These targets should also be effective in treating obese T2D. However, it has been difficult to target feeding behavior, because considerable redundancies...
<table>
<thead>
<tr>
<th>Treatment</th>
<th>Pharmacology and positive attributes</th>
<th>Conditions favoring use</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACEi (oral, multiple agents)</td>
<td>Prevents conversion of Ang I to Ang II, which prevents action of Ang II at its receptor to cause vasoconstriction and cardiac stimulation</td>
<td>HF, post-MI, nephropathy, LV hypertrophy, carotid atherosclerosis, atrial fibrillation, MetS</td>
<td>May cause cough, elevated potassium levels, low blood pressure, dizziness, headache</td>
</tr>
<tr>
<td>ARBs (oral multiple agents)</td>
<td>Blocks the action of Ang II at its receptor</td>
<td>HF, post-MI, nephropathy, LV hypertrophy, atrial fibrillation, MetS</td>
<td>May cause cough, elevated potassium levels, low blood pressure, dizziness, headache</td>
</tr>
<tr>
<td>Aliskiren (oral)</td>
<td>Direct renin inhibitor, inhibits the RAAS pathway.</td>
<td>HF, post-MI, diabetic nephropathies, hypertension, kidney disorders</td>
<td>Side effects include angioedema, hyperkalemia, hypotension, GI symptoms</td>
</tr>
<tr>
<td>Beta blockers (oral, multiple agents)</td>
<td>Blocks the βAR, decreases the chronotropic, inotropic and vasodilator responses to βAR stimulation</td>
<td>HF, post-MI, angina pectoris, tachyarrhythmias, glaucoma, pregnancy</td>
<td>May cause weight gain, decrease insulin sensitivity and adversely affect plasma lipids</td>
</tr>
<tr>
<td>Thiazide diuretics (oral, multiple agents)</td>
<td>Inhibits Na^+/-Cl^- reabsorption from the distal convoluted tubules in the kidneys by blocking the thiazide-sensitive Na^+/-Cl^- symporter</td>
<td>Isolated systolic hypertension (elderly), HF, hypertension in blacks</td>
<td>May cause hypokalemia and increased serum cholesterol; long-term use may increase homocysteine (associated with atherosclerosis)</td>
</tr>
<tr>
<td>Calcium antagonists: dihydropyridines (oral, multiple agents)</td>
<td>Block L-type voltage-gated calcium channels in muscle cells of the heart and blood vessels; often used to reduce systemic vascular resistance and arterial pressure</td>
<td>Isolated systolic hypertension, angina pectoris, LV hypertrophy, carotid/coronary atherosclerosis, pregnancy, hypertension in blacks</td>
<td>Side effects include peripheral edema, dizziness, not used to treat angina (with the exception of amlodipine); contra-indicated in certain patient populations.</td>
</tr>
<tr>
<td>Calcium antagonists: (oral, verapamil diltiazem)</td>
<td>L-type calcium channel blocker, decreases impulse conduction through the AV node, protects ventricles from atrial tachyarrhythmias; causes smooth muscle relaxation and vasodilation.</td>
<td>Angina pectoris, carotid atherosclerosis, supraventricular tachycardia</td>
<td>Side effects include headache, constipation. Side effects include dizziness, flushing, peripheral edema. (Diltiazem is contra-indicated in certain patient populations.)</td>
</tr>
</tbody>
</table>

Information was obtained from References 22 and 56.
Table 6 Future treatments for metabolic diseases

<table>
<thead>
<tr>
<th>Novel pharmacological targets</th>
<th>Predicted Positive Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bwt Loss</td>
</tr>
<tr>
<td>Improve Energy Homeostasis</td>
<td></td>
</tr>
<tr>
<td>Central regulation: NPY, ↑MSH-R, Gut regulation: ↑Ghrelin, ↑1CCK, ↑1PPY</td>
<td></td>
</tr>
<tr>
<td>Peripheral regulation: AMPK, ↑PYY, ↓11-HSD</td>
<td></td>
</tr>
<tr>
<td>Adipokine and cytokine regulation: adiponectin, ↑inflammation</td>
<td></td>
</tr>
<tr>
<td>Improve Glucose Homeostasis</td>
<td></td>
</tr>
<tr>
<td>Insulin signaling stimulation: ↓PTEN, Glycogen regulation: ↓2GSK, ↓3G5, Glucose transporter inhibitors, β-cell restoration</td>
<td></td>
</tr>
<tr>
<td>Improve Lipid Profile</td>
<td></td>
</tr>
<tr>
<td>Reverse cholesterol transport stimulation: ↑1HDL, ↑apo A1, ↑1ABC transporter, ↓1CETP LDL reduction: Aβ42-40 B, ↓MTP, ↓PCSK9</td>
<td></td>
</tr>
<tr>
<td>Lower Blood Pressure</td>
<td></td>
</tr>
<tr>
<td>Neutral endopeptidase inhibitors, Nitrin oxide donors, Endothelin receptor antagonists</td>
<td></td>
</tr>
</tbody>
</table>

Summary of predicted positive effects on body weight reduction, improvement in diabetes control, improvement in lipid profile, blood pressure decrease, and improvement in the vasculature.

exist in the regulation of food intake. A nother focus to treat the metabolic dysregulation in insulin-resistant and diabetic patients is to augment insulin receptor signaling and also to treat β-cell impairment. Future approaches to treating vessel wall damage include activating reverse cholesterol transport and novel anti-hypertensive targets. Several challenges are faced in developing novel therapeutics to treat cardiometabolic disorders. These challenges include identifying and validating novel targets, using predictive animal model of human efficacy, identifying responsive patient populations, obtaining an acceptable safety and tolerability profile for chronic treatment, and positioning a product in a huge, competitive marketplace. However, there is a high unmet medical need and room for new, differentiated products. Recent advances hold promise for novel therapies for treating multiple risk factors for T2D and CVD.

Appendix: List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>acetyl CoA carboxylase</td>
</tr>
<tr>
<td>ACEi</td>
<td>angiotensin converting enzyme inhibitor</td>
</tr>
<tr>
<td>AGE</td>
<td>advanced glycation endproducts</td>
</tr>
<tr>
<td>AGI</td>
<td>alpha glucosidase inhibitor</td>
</tr>
<tr>
<td>AMPK</td>
<td>AMP-activated protein kinase</td>
</tr>
<tr>
<td>ARB</td>
<td>angiotensin receptor blockers</td>
</tr>
<tr>
<td>pAR</td>
<td>beta adrenergic receptor</td>
</tr>
<tr>
<td>CCK</td>
<td>cholecystokinin</td>
</tr>
<tr>
<td>CETP</td>
<td>cholesteryl ester transfer protein</td>
</tr>
<tr>
<td>CHD</td>
<td>coronary heart disease</td>
</tr>
<tr>
<td>CVD</td>
<td>cardiovascular disease</td>
</tr>
<tr>
<td>CRP</td>
<td>C-reactive protein</td>
</tr>
<tr>
<td>FFA</td>
<td>free fatty acid</td>
</tr>
<tr>
<td>GH</td>
<td>growth hormone</td>
</tr>
<tr>
<td>GLP-1</td>
<td>glucagon-like protein 1</td>
</tr>
<tr>
<td>GS2</td>
<td>glycogen synthase 2</td>
</tr>
<tr>
<td>HDL</td>
<td>high density lipoprotein</td>
</tr>
<tr>
<td>HF</td>
<td>heart failure</td>
</tr>
<tr>
<td>I KK</td>
<td>inhibitor of kappa B kinase</td>
</tr>
<tr>
<td>J N K</td>
<td>c-jun N-terminal kinase</td>
</tr>
<tr>
<td>LDL</td>
<td>low density lipoprotein</td>
</tr>
<tr>
<td>Lp(a)</td>
<td>lipoprotein “little” a</td>
</tr>
<tr>
<td>LV</td>
<td>left ventricular</td>
</tr>
<tr>
<td>MDS</td>
<td>metabolic syndrome</td>
</tr>
<tr>
<td>MI</td>
<td>myocardial infarction</td>
</tr>
<tr>
<td>MSH-R</td>
<td>melanocyte stimulating hormone receptor</td>
</tr>
<tr>
<td>MTP</td>
<td>microsomal triglyceride transfer protein</td>
</tr>
<tr>
<td>NEFA</td>
<td>non-esterified fatty acids</td>
</tr>
<tr>
<td>NO</td>
<td>nitric oxide</td>
</tr>
<tr>
<td>NPY</td>
<td>neuropeptide Y</td>
</tr>
<tr>
<td>PCSK9</td>
<td>proprotein convertase subtilisin/kexin type 9</td>
</tr>
<tr>
<td>PRAR-γ</td>
<td>peroxisome-proliferator-activated receptor-γ</td>
</tr>
<tr>
<td>PGG</td>
<td>post-prandial</td>
</tr>
<tr>
<td>PTEN</td>
<td>phosphatase and tensin homolog</td>
</tr>
<tr>
<td>PTPase</td>
<td>phosphotyrosine phosphatase</td>
</tr>
<tr>
<td>PYY</td>
<td>peptide YY</td>
</tr>
<tr>
<td>RAAS</td>
<td>renin, angiotensin, aldosterone system</td>
</tr>
<tr>
<td>SCD-1</td>
<td>stearoyl-CoA desaturase 1</td>
</tr>
<tr>
<td>Sirt1</td>
<td>silent information regulator 1</td>
</tr>
<tr>
<td>TG</td>
<td>triglyceride</td>
</tr>
<tr>
<td>TH</td>
<td>thyroid hormone</td>
</tr>
<tr>
<td>T2D</td>
<td>type 2 diabetes</td>
</tr>
<tr>
<td>VLDL</td>
<td>very low density lipoprotein</td>
</tr>
</tbody>
</table>
11β-HSD1, 11 beta-hydroxysteroid dehydrogenase type 1

References

5. WHO website-diabetes-actionononidietabesies/

Further Reading

See Also
LDL and HDL Receptors
Lipid Homostasis
Lipoproteins, Chemistry of
Metabolic Diseases, Biological Mechanisms of

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Mitochondrial Medicine, Biochemical Evaluation of Mitochondrial Function

Richard J.T. Rodenburg and Jan A.M. Smeitink, Nijmegen Center for Mitochondrial Disorders, Radboud University Nijmegen Medical Centre, The Netherlands
doi: 10.1002/9780470048672.wecb343

Mitochondrial medicine is an emerging field in medicine that focuses on diseases in which the mitochondrial energy generating system plays a central role. These diseases include the “classical” mitochondrial disorders, in which a (genetic) defect in the mitochondrial energy generating system is the primary cause of the disease, but also more common disorders such as Parkinson and cancer, in which mitochondrial energy metabolism plays an important role in the pathogenesis. In this review, we present an overview of the mitochondrial energy generating system, starting at the conversion of pyruvate into acetyl-CoA by pyruvate dehydrogenase, via the TCA cycle in which reduction equivalents are formed, to the oxidative phosphorylation system, where the reduction equivalents are used to convert ADP into ATP. The mitochondrial energy generating system can be examined by global assays that measure the rate of pyruvate conversion, the rate of oxygen consumption, or the rate of ATP production. In addition, an overview is given of the spectrophotometric assays that are available to measure PDHc, several TCA cycle enzymes, and the enzymes of the oxidative phosphorylation. In the diagnostic analysis of patients suspected to suffer from a mitochondrial disorder, these assays are applied to evaluate the functioning of the mitochondrial energy generating system in muscle biopsies and other types of patient samples. In addition, the results provide clues for further investigations at the molecular genetic level. Thus, the biochemical analysis of patient material is an important step in establishing the diagnosis of a mitochondrial disorder.

The importance of mitochondria in health and disease has given rise to a new area in medicine, called mitochondrial medicine. The foundation for this medical discipline lies in the recognition of diseases in which disturbances in one of the many steps of mitochondrial energy production are present. The role of mitochondria in energy metabolism disorders is well recognized and has been the subject of study for many decades. The increasing awareness of the relationship between mitochondria and several more common disorders, like Parkinson disease and cancer, makes a thorough understanding of the chemistry of the mitochondrial energy generating system, and the analytical methods to examine the functioning of this system, necessary for an increasing variety of medical and biochemical specialists. Here, we review a) how the cell’s energy currency, ATP, is produced, and b) assays to determine the overall capacity of the system as well as single enzymes in relation to genetic disorders of energy production.

Biological Background

Mitochondrial disorders

Mitochondrial disorders can be defined as disorders that are caused by a defect in the mitochondrial energy generating system (MEGS). The clinical spectrum is very broad, but in almost all cases involve one or more tissues that have a high energy demand, in particular skeletal muscle and neuronal tissue. In addition, heart, liver, kidney, and other tissues can be involved as well. The severity and the course of the disease are
The production of ATP by the MEGS is a complex process involving many different transporters and enzymes (4). Mitochondrial ATP is the end product of the oxidation of pyruvate (alpha-keto propionic acid). Pyruvate is the final product of the glycolysis, the anaerobic catabolism of glucose. Other substrates for the MEGS are fatty acids and several amino acids, in particular glutamine. Pyruvate is transported into the mitochondrion where it is metabolized into acetyl-CoA. The acetyl-CoA is oxidized in the tricarboxylic acid (TCA) cycle, during which both NADH and FADH$_2$ are produced. These reducing equivalents are oxidized by the respiratory chain, which leads to the translocation of protons out of the mitochondrial matrix. The mitochondria use the resulting proton-motive force to generate ATP from ADP and phosphate. The ATP is released in the mitochondrial matrix, and can be exported to the cytosol. The MEGS is described in more detail in the next section.

The Chemistry of the OXPHOS System

The conversion of pyruvate to acetyl-CoA involves two key players. Pyruvate is imported into the mitochondrion by a specific transporter. To date, only one case has been described in which a pyruvate carrier deficiency could be shown by functional assays (5). However, no genetic defect responsible for pyruvate carrier deficiency has been identified yet. After pyruvate has entered the mitochondrial matrix, it is converted into acetyl-CoA by the pyruvate dehydrogenase complex, PDHc.

This large enzyme complex has a molecular mass of approximately 9 MDa and contains multiple copies of three enzyme subcomplexes: 20-30 copies of alpha-ketoacid dehydrogenase (E3; EC 1.2.4.1), 60 copies of dihydrolipoamide acyltransferase (E2; EC 2.3.1.2), and 6 copies of dihydrolipoamide dehydrogenase (E1; EC 1.8.1.4), as well as 12 copies of the structural building block E3 binding protein (6). The E1 subcomplex consists of a tetramer of two E1α and two E1β subunits. PDHc de-carboxylates pyruvate and esterifies the resulting acetyl-group to CoA. During this reaction, NAD$^+$ is reduced to NADH. PDHc requires 5 different cofactors, namely NAD$^+$, CoA, thiamine pyrophosphate, FAD, and lipoic acid. The activity of PDHc is tightly regulated. First of all, the activity is controlled in an allosteric manner by the reaction products acetyl-CoA and NADH. In addition, two specific enzymes regulate PDHc activity via a phosphorylation site in the E1 subunit. PDH kinase (EC 2.7.11.2) is ATP dependent and inactivates PDHc by phosphorylating the E1 component when the ATP/ADP ratio is low (7). This enzyme is also activated by high NAD$^+$ and acetyl-CoA levels, and inactivated by high pyruvate levels. Four isoforms of PDH kinase have been identified, each having a different tissue distribution. By contrast, when the ATP/ADP ratio is high and the pyruvate levels are high, PDH kinase is not active and PDHc is dephosphorylated by PDH phosphatase (EC 3.1.3.43), of which two isoforms are known (8). Due to this tight regulation, the oxidation rate of pyruvate by PDHc is directly coupled to the mitochondrial ATP production rate.

The generation of reduction equivalents in the TCA cycle

The acetyl-CoA generated by PDHc fuels the TCA cycle. Three dehydrogenases of the TCA cycle are responsible for the reduction of NAD$^+$ into NADH. These are isocitrate dehydrogenase (EC 1.1.1.42), a-ketoglutarate dehydrogenase (consisting of 3 enzymatic subunits E1 (EC 1.2.4.2), E2 (2.3.1.61), and E3 (EC 1.8.1.4), and malate dehydrogenase (EC 1.1.1.37). In addition, the TCA cycle enzyme succinate dehydrogenase (EC 1.3.9.1) converts FADH$_2$ into FAD. Furthermore, succinate-CoA ligase generates GTP or ATP, depending on the isotype of this enzyme complex (EC 6.2.1.4 and EC 6.2.1.5, respectively). The formation of these high-energy molecules is accompanied by the release of CO$_2$ in two of the reactions of the TCA cycle. In addition to acetyl-CoA, other metabolites can fuel the TCA cycle as well. This includes glutamine, which enters the TCA cycle at the site of 2-oxoglutarate. In tissues that catabolize fatty acids, the end product of the beta-oxidation is acetyl-CoA, which is converted into the ketone bodies 3-hydroxybutyrate and acetacetate (in the liver) or enters the TCA cycle (in most other tissues) and in this way contributes to the synthesis of ATP.

The oxidative phosphorylation system

NADH and FADH$_2$ are oxidized by the OXPHOS system to generate ATP. This is a coordinated multistep process that involves 5 large enzyme complexes: the respiratory chain complexes I, II, III, and IV, and ATP synthase (complex V). Complex I (NADH ubiquinone oxidoreductase; EC 1.6.5.3) is by far the largest respiratory chain enzyme complex. It consists of 45 different subunits and has a molecular weight of approximately 1 MDa. Complex I oxidizes NADH and the electrons that are released from NADH are transferred to a flavin mononucleotide present in complex I and subsequently via a channel of 8 iron sulfur clusters within the peripheral arm of the complex towards CoQ$_2$ that is present in the inner mitochondrial membrane (9). The redox reaction of complex I is directly coupled to the pumping of protons by complex I from the mitochondrial matrix across the mitochondrial inner membrane to the mitochondrial intermembrane space, which is in direct connection with the cytosol for small ions. Complex II (succinate dehydrogenase; EC 1.3.5.1) oxidizes FADH$_2$ and transfers electrons towards the respiratory chain.
CoQ₁₀ Complex II is the only respiratory chain complex that does not contribute to the mitochondrial proton gradient. Complex III (ubiquinol:cytochrome c oxidoreductase; EC 1.10.2.2) translocates the electrons from CoQ₁₀ to cytochrome c, a small, haeme-containing protein that acts as an electron carrier between complexes III and IV of the respiratory chain. The last step of the respiratory chain is complex IV (cytochrome c oxidase; EC 1.9.3.1), that oxidizes cytochrome c and transfers the electrons to molecular oxygen, which leads to the generation of water. The pumping of protons by the respiratory chain complexes I, III, and IV across the inner mitochondrial membrane increases the pH of the mitochondrial matrix and generates a potential difference between matrix and intermembrane space. Complex V (adenosine triphosphatase; EC 3.6.1.3) utilizes the potential energy of the proton gradient to convert ADP and phosphate into ATP.

Defects in the mitochondrial energy generating system

In theory, a defect in any of the transport proteins and enzymes mentioned above could result in a reduced mitochondrial energy generating capacity. To date, primary defects at the protein and DNA level have been found in PDHc, fumarase (EC 4.2.1.2) (10), α-ketoglutarate dehydrogenase (2-oxoglutarate dehydrogenase) (12), succinate-CoA ligase (12, 13), complexes I (9), II (14), III (15), IV (16), and V (17), the phosphate carrier (18), and α-NT. A functional defect in the pyruvate carrier has been identified as well (5). In addition to these structural components of the mitochondrial energy generating system, there are many additional proteins involved in the production of these structural components. These include the chaperones required to assemble the enzyme complexes of the respiratory chain. The structural building blocks of the OXPHOS system are encoded by multiple genes. Except for complex II, these genes are located in both the nuclear and the mitochondrial genomes. Most nuclear genetic defects in these structural genes result in an isolated enzyme deficiency, although several examples have been described of patients with a mutation in a complex I gene that also result in reduced enzyme activities of other enzyme complexes, e.g. complex III and PDHc, indicating that these complexes have a higher order of organisation that can be disturbed by defects in one of the complexes. The existence of the so-called supercomplexes has become more apparent by functional and structural studies in the last few years. Many different pathogenic point mutations and rearrangements in the mtDNA have been found in the last two decades. Depending on the type of mutation, these can cause either isolated or combined deficiencies of complexes I, III, IV, and V (18). The mitochondrial genome (or mtDNA) is replicated by a dedicated polymerase, POLG. Defects in the POLG gene leads to depletion of, and/or deletions in, the mtDNA. This causes isolated or combined deficiencies of the mtDNA encoded OXPHOS complexes I, III, IV, and V. Furthermore, defects in proteins involved in nucleotide metabolism can lead to mtDNA depletion and OXPHOS deficiencies as well. Depletion has also been observed in patients suffering from a defect in the TCA cycle enzyme succinate-CoA ligase, although the underlying mechanism is not yet fully understood (12, 20).

The transcription and translation of mtDNA encoded proteins involves several mitochondria-specific proteins, in which defects also lead to combined OXPHOS enzyme deficiencies (21, 22). A special class of defects are those leading to a CoQ₁₀ deficiency. The biosynthesis of CoQ₁₀ is a multistep process that has been completely elucidated in yeast. In humans, several steps of this process have been shown to exist as well, and three different genetic defects in this pathway have been identified to date (23).

Tools and Techniques to Study the OXPHOS System

There are several approaches to perform biochemical analyses of the OXPHOS system. This chapter will focus on assays to perform structural analyses, enzyme activity assays, and ATP production, oxygen consumption and substrate oxidation assays.

Structural analysis of the OXPHOS system

Blue-Native PAGE is a technique that is very suitable to study the relative amounts and the assembly status of OXPHOS complexes (24, 25). It can be performed as either a 1D or a 2D assay. In the 1-dimensional approach, the complexes are separated on a non-denaturing acrylamide gel containing the Commassie dye Serva Blue G. All 5 complexes can be visualised in this way. In the 2-dimensional approach, the second dimension is a denaturing SDS-PAGE, resulting in the separation of the OXPHOS complexes into their individual protein building blocks. After blotting, more specific staining methods can be performed, e.g. using anti-OXPHOS complex antibodies. This can provide detailed information on the assembly status of the OXPHOS complexes. A rather powerful tool to perform structural analysis of OXPHOS complexes is by immunopurification followed by mass spectrometric analysis of the isolated complexes. Using this approach, it has been shown that bovine complex I (and presumably also human complex I) consists of 45 different subunits (26). Recently, this technique has lead to the identification of Ecst as a protein involved in complex I assembly, as it was found to be associated with complex I (27). This is quite an unexpected finding, as Ecst was previously known as a cytosolic protein involved in a pro-inflammatory signal-transduction pathway from a Toll-like receptor and in embryonic development (28). By immunopurification and mass spectrometry, it could be shown that Ecst associates with complex I, and that the N-terminal part of Ecst appears to be required for mitochondrial import (27).

Enzyme activity assays

The traditional way to determine OXPHOS enzyme activities is by spectrophotometry. Several assays have been described for all 5 OXPHOS complexes. The assays are performed in homogenates of tissue samples or cultured cells, in crude mitochondria-enriched 600 g supernatants of tissue/cell homogenates, or in mitochondrial preparations from 14000 g pellets derived from 600 g supernatants. Obviously, the higher the
purity of the mitochondrial preparation, the higher the specific activity of the enzymes measured. Therefore, very high specific activities can be achieved by using immunopurified complexes (29), although this approach is not widely applied yet. In addition to enzyme activity assays in solution, BN-PAGE can be used to estimate OXPHOS enzyme activities by in-gel activity assays (30). In-gel activities are particularly suitable to monitor relative activities, e.g., to evaluate changes in enzyme activities under different experimental conditions. For quantitative enzyme activity measurements, spectrophotometric analysis is the method of choice. Table 1 contains a summary of the most commonly used spectrophotometric enzyme assays for measurement of the OXPHOS enzymes. Below, a brief description of these assays is given.

**Complex I**

Spectrophotometric assays for measuring the activity of complex I (or NADH:ubiquinone oxidoreductase) are usually based on measurements of NADH, which is oxidized by complex I to NAD+. In addition, the assay requires CoQ as a cosubstrate for complex I. Usually CoQ$_1$ or decylubiquinone are used because these have better solubility than CoQ$_{10}$. Assays usually contain bovine serum albumine, which probably is required to stabilize the protein sample and to aid the solubilization of CoQ analogues. The specific activity of complex I can be determined by measuring the rate of NADH conversion at 340 nm in the absence and presence of the specific complex I inhibitor rotenone (31). Recently, our lab described a new assay that measures complex I by including 2,6-dichlorophenolindophenol (DCIP) as a terminal acceptor of electrons that are derived from the oxidation of NADH and the subsequent reduction of the CoQ-analogue decylubiquinone (32). DCIP reduction can be followed spectrophotometrically at 600 nm. As the molar absorption coefficient of DCIP at 600 nm is more than 3 times higher than that of NADH at 340 nm, this new assay has a much higher sensitivity than the assay that measures NADH, with similar specificity. Complex I can also be measured as NADH:cytochrome c oxidoreductase, in which the combined activity of complex I + CoQ$_1$ + III is measured by addition of NADH and oxidized cytochrome c as substrates. The assay measures the rotenone-sensitive reduction of cytochrome c, which can be followed spectrophotometrically at 550 nm.

**Complex II**

Complex II is usually measured in two ways: either as succinate:ubiquinone oxidoreductase or as succinylate:cytochrome c oxidoreductase. The most commonly used assay for succinate:ubiquinone oxidoreductase (or isolated complex II) uses DCIP in the same way as described above for the new complex I assay, only in this case succinate is added as a substrate (instead of NADH). The specificity of DCIP reduction can be determined by measuring in the presence or absence of malonate, a specific inhibitor of complex II. The assay for succinate:cytochrome c oxidoreductase (or complexes II + III) uses succinate and oxidized cytochrome c as substrates and measures the reduction of cytochrome c, which can be followed spectrophotometrically at 550 nm. The assay is also suitable to screen for coenzyme Q deficiencies, as it is dependent on the endogenously present CoQ$_1$. In case of a CoQ$_1$ deficiency, a reduced succinate:cytochrome c oxidoreductase activity will be observed that can be normalized by addition of exogenous CoQ to the reaction mixture (33). Complex III (ubiquinol:cytochrome c oxidoreductase) reduces cytochrome c and oxidizes reduced CoQ. In addition to these substrates, the assay contains a strong complex IV inhibitor (e.g., potassium cyanide) to prevent re-oxidation of reduced cytochrome c. The complex III activity can be derived from the rate of cytochrome c reduction, which can be followed at 550 nm. The specificity of the assay is determined by measuring in the absence or presence of antimycin A, a specific inhibitor of complex III activity.

**Complex IV**

Complex IV (cytochrome c oxidase) is measured by addition of reduced cytochrome c to the reaction mixture. The oxidation of cytochrome c can be followed at 550 nm. This assay has a low background activity, and measurement in the absence and presence of a selective inhibitor is not necessary.

**Complex V**

The activity of the ATP-forming enzyme complex V is usually assessed by determining the reverse reaction ATP → ADP + Pi. The reaction is coupled to reactions catalyzed by pyruvate kinase (ADP + phosphoenolpyruvate → pyruvate + ATP) and lactate dehydrogenase (pyruvate + NADH → lactate + NAD$^+$). This final reaction can be followed spectrophotometrically by measuring NADH at 340 nm. The activity of complex V (ATPase) can be derived from the rate of NADH conversion in the presence and absence of the specific complex V inhibitor oligomycin.

**Other enzymes**

In addition to assays for the OXPHOS enzymes, assays have been described for many other enzymes involved in the MEGS. For PDHc, various types of diagnostic assays have been described and are widely used (33). Also for the TCA cycle enzymes various assays have been described, although to date, pathogenic defects have only been found in α-ketoglutarate dehydrogenase (38), succinicate-CoA ligase (12), and lumanase (39).

**The mitochondrial energy generating system**

As outlined above, the generation of ATP from pyruvate by the mitochondrion involves many different transporters and enzymes. A subset of the individual transporters can be assayed individually (e.g., the respiratory chain enzymes). In addition, the process of ATP generation can be studied by using assays that provide information of the mitochondrial energy generating system (MEGS) in toto, and moreover, on the functioning of individual enzymes in the context of the intact mitochondrion. Several types of assays have been developed and are described below. For all these assays, a crucial factor is the integrity of the mitochondria, since the generation of ATP is dependent on an inner mitochondrial membrane potential. Therefore, it is important to include control experiments that test the coupling of the ATP synthesis to the respiratory chain, as this will provide information on the integrity of the inner mitochondrial membrane.
Monitor different metabolic routes that lead to mitochondrial dehydrogenase. As the ATP production rate is under the control where it is coupled to oxalate that is the product of malate acetyl-CoA that is further metabolized in the citric acid cycle is subsequently metabolized by pyruvate dehydrogenase to pyruvate enters the mitochondria via a pyruvate carrier and ATP synthesis. For example, by addition of pyruvate and malate, tissue homogenates (41). Different substrates can be added to by digitonin (40), or in mitochondria-enriched fractions from The first type is performed in cultured cells permeabilized Two types of ATP production assays can be discriminated.

**ATP production assays**

Two types of ATP production assays can be discriminated. The first type is performed in cultured cells permeabilized by digitonin (40), or in mitochondria-enriched fractions from tissue homogenates (41). Different substrates can be added to monitor different metabolic routes that lead to mitochondrial ATP synthesis. For example, by addition of pyruvate and malate, pyruvate enters the mitochondria via a pyruvate carrier and is subsequently metabolized by pyruvate dehydrogenase to acetyl-CoA that is further metabolized in the citric acid cycle where it is coupled to oxalate that is the product of malate dehydrogenase. As the ATP production rate is under the control of the ADP/ATP ratio, an excess of ADP should be present in the pathway from the pyruvate carrier to ATP can be monitored. Any primary defect in this pathway will result in a decreased ATP production rate. Therefore, this type of assay is very suitable as a diagnostic tool to screen for a defect in the MEGS. The second type of ATP production assay is restricted to cultured cells. In this assay, luciferase is used as an ATP sensor in vivo. Luciferase, an enzyme derived from the firefly, converts its substrate luciferin under the emission of light. This reaction requires ATP, and thus, the amount of light produced is a measure for the amount of ATP. Cells can be stably transfected with an expression vector encoding luciferase (42), infected with a virus that encodes luciferase (43), or microinjected with plasmid DNA encoding luciferase (44). By addition of an appropriate targeting sequence, luciferase can be expressed specifically in the mitochondria in order to monitor intramitochondrial ATP. The ATP levels can be monitored by luminometry. A more sophisticated approach is to determine subcellular (e.g. intramitochondrial) ATP levels by a microscope coupled to a CCD camera. When cells are placed in a flow cell under the microscope/CCD camera system, they can be exposed to different stimuli that regulate ATP production, in particular hormones that lead to intracellular Ca²⁺ fluxes, which stimulate mitochondrial ATP production (43). This can be monitored real-time by means of the camera. In contrast to the first type of ATP assay, this real-time ATP assay is not only dependent on the integrity of the mitochondrial ATP generating machinery, but also on the intracellular mechanisms that regulate ATP production. This set-up is very suitable to perform functional studies of the mitochondrial energy generating system in vivo and the factors that affect the activity of this system, but is less suitable as a diagnostic tool.

**Substrate oxidation rate measurements**

Substrate oxidation rate measurements provide detailed information on the functioning of the MEGS (41). The MEGS performs three decarboxylation reactions, at the level of PDHc, isocitrate dehydrogenase, and 2-oxoglutarate dehydrogenase. By using radiolabeled substrates in which the radiolabel is present at a carboxyl residue, the activity of the MEGS can be followed by measuring the amount of released radiolabeled CO₂. Usually, combinations of radiolabeled substrates and unlabeled

### Table 1

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Activity measured</th>
<th>Substrates</th>
<th>Specific inhibitor</th>
<th>Detection</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>complex I</td>
<td>NADH:ubiquinone oxidoreductase</td>
<td>NADH, UQ₁</td>
<td>rotenone</td>
<td>NADH (340 nm)</td>
<td>[31, 33, 35]</td>
</tr>
<tr>
<td>complex I + III</td>
<td>NADH:cytochrome c oxidoreductase</td>
<td>NADH, DQ</td>
<td>rotenone</td>
<td>DCIP (600 nm)</td>
<td>[32]</td>
</tr>
<tr>
<td>complex II</td>
<td>NADH:ubiquinone oxidoreductase</td>
<td>NADH, DQ</td>
<td>rotenone</td>
<td>DCIP (600 nm)</td>
<td>[33]</td>
</tr>
<tr>
<td>complex II + III</td>
<td>NADH:cytochrome c oxidoreductase</td>
<td>NADH, DQ</td>
<td>rotenone</td>
<td>DCIP (600 nm)</td>
<td>[33, 35]</td>
</tr>
<tr>
<td>complex III</td>
<td>NADH:cytochrome c oxidoreductase</td>
<td>NADH, DQ</td>
<td>rotenone</td>
<td>DCIP (600 nm)</td>
<td>[33, 35]</td>
</tr>
<tr>
<td>complex IV</td>
<td>NADH:cytochrome c oxidase</td>
<td>NADH, DQ</td>
<td>rotenone</td>
<td>DCIP (600 nm)</td>
<td>[33, 35]</td>
</tr>
<tr>
<td>complex V</td>
<td>F₁-ATPase</td>
<td>ATP</td>
<td>oligomycin</td>
<td>NADH (via PK/LDH at 340 nm)</td>
<td>[33, 35]</td>
</tr>
<tr>
<td>PDHc</td>
<td>pyruvate dehydrogenase</td>
<td>pyruvate, NADH, CoA</td>
<td>none</td>
<td>CO₂ (radiochemically)</td>
<td>[35, 37]</td>
</tr>
</tbody>
</table>

Non-standard abbreviations: UQ₁: ubiquinone-Q₁; DQ: decylubiquinone; DCIP: 2,6-dichlorophenolindophenol, PK: pyruvate kinase; LDH: lactate dehydrogenase; AABS: p-(p-aminophenyl)azo benzene sulfonic acid; ArAt: arylamine acetyltransferase
co-substrates are used, either in the presence or absence of specific inhibitors. An example is the use of radiolabeled pyruvate, which is decarboxylated by PDHc. To force the reaction to proceed at maximum rate, the reaction contains an excess of ADP, in order to maintain a high ADP/ATP ratio. In addition, the acetyl-CoA formed by PDHc has to be removed by addition of an appropriate co-substrate. When carnitine is used as co-substrate, it will be coupled to the acetyl group of acetyl-CoA by carnitine-acetyl transferase. When malate is used as a co-substrate, this will be converted to oxalacetate in the TCA cycle which is subsequently coupled to the acetylgroup of acetyl-CoA by citrate synthase. When these assays are performed in a control sample, e.g., a muscle sample from a healthy individual, the ratio of the pyruvate oxidation in the presence of carnitine or malate will be approximately 1. Also in a sample from a PDHc deficient patient, this ratio will be near 1, however, in that case both reaction rates will be equally reduced due to the PDHc defect. Interestingly, in case of a respiratory chain defect, the ratio between these two reactions will be around 2 in favour of the reaction with carnitine (43). A third reaction that does not directly involve a TCA cycle enzyme is the conversion of pyruvate + malate to oxaloacetate + ATP. The advantage and disadvantage of this reaction is that the latter is the reverse reaction of the reaction catalyzed by PDHc. The reaction with malate as a co-substrate allows the evaluation of this conversion with malate as a co-substrate. As a final example, in case of a complex V defect, the reaction of pyruvate + malate will have a lower rate than in a control sample. Addition of the uncoupler CCCP will result in a normalization of the reaction rate, a phenomenon that is also observed in case of a defect in the phosphate carrier or the ATP:ADP antiporter ANT.

Oxygen consumption assays

In principle, oxygen consumption rate assays can be used for the same purposes as substrate oxidation rate assays. Also in this case, the pathway from the substrate of choice down to molecular oxygen conversion by complex IV can be evaluated, and even the steps beyond complex IV that are of influence on the activity of complex IV, such as complex V (45). The assay can be performed on frozen or tissue extracts or in whole cells in vivo or ex vivo. The classical way to detect molecular oxygen consumption by using a Clark-type oxygen electrode (46). More recently, molecular probes have been developed that make it possible to perform oxygen consumption rate measurements by fluorimetry (46, 47). This latter type of assay has the advantage that small volumes can be tested in 96-well plates using relatively simple laboratory equipment. By using appropriate combinations of substrates, the oxygen consumption rate measurements are suitable to locate primary defects in the MEGS, in a similar manner as with radiolabelled substrate oxidation rate assays.

Diagnostic Biochemical Analysis of the Mitochondrial Energy Generating System

The mitochondrial energy generating system requires efficient interplay between a large number of different proteins and protein complexes, which by themselves can be made up of large numbers of individual subunits that have to be assembled in an ordered manner. This complexity of the mitochondrial energy generating system is probably one of the main reasons for the heterogeneity of mitochondrial disorders. Due to this clinical diversity, establishing the diagnosis “mitochondrial disorder” usually requires a combination of clinical, biochemical, and genetic examination of the patient suspected for a mitochondrial disorder (48–50). The biochemical analysis of a muscle biopsy is the corner stone of the diagnostic examination for mitochondrial disorders. The results will show whether or not the MEGS functions properly in this tissue. Unfortunately, the number of (potentially) mitochondrial disease causing candidate genes is very large, and molecular genetic techniques to rapidly sequence hundreds of candidate genes in a diagnostic setting are not yet available. The biochemical results are not only diagnostic in their own right, but, in combination with the clinical features of the patient, also provide important clues that are used to select candidate genes for molecular genetic analysis. Nevertheless, the diagnosis mitochondrial disorder can not always be made at the molecular genetic level, in particular in those cases in which a comprehensive biochemical analysis has not been performed.

The biochemical diagnostic analysis of a patient suspected for a mitochondrial disorder is usually performed on a muscle biopsy, as this tissue has a very high energy demand and is more likely to exhibit signs of mitochondrial dysfunction than tissues with a relatively low energy conversion rate. Depending on the clinical features, it could be considered to examine other types of tissue, such as liver or heart. It has been shown that in mitochondrial depletion syndromes with liver involvement, e.g., due to mutations in DGUOK or MPV17, muscle tissue may not always show biochemical aberrations while liver shows clear signs of enzyme deficiencies (51, 52). Obvious drawbacks of organ or muscle biopsies are that an invasive procedure is required to obtain the tissue sample. Technically, it is possible to measure in fibroblasts or even blood samples, but these do not always express the mitochondrial defect. Nevertheless, these types of samples do have a very important added value to the biochemical diagnosis. First of all, a positive or a negative fibroblast result in combination with a positive muscle result has consequences for the selection of candidate genes for subsequent molecular genetic analysis. For example, in case of a mitochondrial depletion syndrome due to mutations in the POLG gene, very severe enzyme deficiencies can be observed in muscle and/or liver, whereas fibroblasts often show normal enzyme activities (53). By contrast, in case of a mitochondrial translation defect, respiratory chain enzyme deficiencies are observed both in muscle and in fibroblasts. A second important aspect of fibroblast (or lymphocyte) analysis is that a systemic expression of a biochemical defect may allow for prenatal diagnosis on the basis of biochemical analysis of chorionic tissue or amniocytes in families of mitochondrial patients in which the underlying molecular genetic defect has not (yet) been identified (54).
References


36. Lopez LC, Schuelke M, Quinzi CM, Korkt T, Rodenburg RJ, Nairn A, Dimaro S, Hirano M. Leigh syndrome with


The amyloidoses are a class of conformational diseases that arise from the conversion of normally unfolded or globular proteins into fibrillar aggregates that are either pathogenic or non-functional. At present there are more than 20 proteins that are associated with human amyloid diseases. This review focuses on three natively unfolded proteins that form fibrillar aggregates; amyloid-\(\beta\) (\(A\beta\)), islet amyloid polypeptide (‘amylin’), and \(\alpha\)-synuclein, the diseases they contribute to and chemical and biophysical approaches that are used to investigate these proteins’ aggregation.

Introduction

Conversion of native conformational folded proteins or peptides into highly-ordered insoluble fibrillar aggregates termed ‘amyloid’ has been demonstrated to be clinically-relevant in several diseases. Misfolded proteins are deposited in a variety of tissues, leading to serious and even fatal human diseases. To date there are over 20 proteins and peptides that are known to induce pathological conditions associated with protein misfolding (Table 1). These proteins include the amyloid-\(\beta\) peptides, prion proteins, \(\alpha\)-synuclein, transthyretin and polyglutamine-containing peptides (1). The hypothesis that these diseases are in fact caused by misfolded protein and the misfolded protein does not arise as a side effect of the disease is supported by familial disease, where proteins containing mutations that facilitate aggregation result in more severe and early onset disease variants.

Diseases arising as a result of protein misfolding are classified based on the location of the protein deposition. Thus, the amyloidoses are generally classified as either neuronal or systemic. The majority of conditions arise sporadically, with only 10% of the cases being linked to hereditary causes. The familial forms of the diseases are caused by mutations that make a protein more prone to aggregation, or by mutations in processing proteins like proteases. Under rare circumstances (5% of protein misfolding diseases), protein aggregation can be initiated by infection and transmission. All diseases caused by misfolded protein share the problem of definitive diagnosis as they are difficult to identify in vivo, especially during early stages of symptoms.

Precursors of amyloid are quite diverse, ranging from small unstructured peptides like amyloid-\(\beta\) peptides to large oligomeric multi-domain proteins like p53. Despite the differences in structure of the monomers, the structures of the misfolded \(\beta\)-sheet rich forms share similar features. In recent years it has become apparent that the proteins that form amyloid share common features, and it has been suggested that many if not all proteins share the intrinsic predisposition to form amyloid under the right conditions (2, 3). This hypothesis is affirmed by the absence of a correlation between the propensity of a peptide or protein to misfold and its sequence. Precursor proteins share neither homology in sequence nor structure. However, certain factors such as hydrophobic regions and/or unstable globular protein conformations contribute to a predisposition to form amyloid.

In this review, we will discuss the aggregation of three intrinsically unstructured peptides and proteins and the diseases to which they contribute. Specifically we will concentrate on amyloid-\(\beta\) (\(A\beta\)), islet amyloid polypeptide (IAPP), and \(\alpha\)-synuclein.

Biological Background of Protein Misfolding Diseases

Amyloid-\(\beta\) (\(A\beta\)) and Alzheimer’s disease

Alzheimer’s disease (AD) is the most common neurodegenerative disease in humans. AD is a mostly sporadic, late-onset
Table 1 List of Protein Misfolding Diseases and the Implicated Proteins and Peptides

<table>
<thead>
<tr>
<th>Clinical Syndrome</th>
<th>Fibril Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s Disease</td>
<td>Aβ peptide, 1-40, 1-42</td>
</tr>
<tr>
<td>Spongiform encephalopathies</td>
<td>Full-length prion and fragments</td>
</tr>
<tr>
<td>Primary systemic amyloidosis</td>
<td>Intact light chain and fragments</td>
</tr>
<tr>
<td>Secondary systemic amyloidosis</td>
<td>76-residue fragment of amyloid A protein</td>
</tr>
<tr>
<td>Familial amyloidotic polyneuropathy I</td>
<td>Transthyretin variants and fragments</td>
</tr>
<tr>
<td>Senile systemic amyloidosis</td>
<td>Wild-type transthyretin and fragments</td>
</tr>
<tr>
<td>Hereditary cerebral amyloid angiopathy</td>
<td>Fragments of cystatin-C</td>
</tr>
<tr>
<td>Haemodialysis-related amyloidosis</td>
<td>β₂-macroglobulin</td>
</tr>
<tr>
<td>Familial amyloidotic polyneuropathy II</td>
<td>Fragments of apolipoprotein A-1</td>
</tr>
<tr>
<td>Finish hereditary amyloidosis</td>
<td>71-residue fragment of gelsolin</td>
</tr>
<tr>
<td>Type II diabetes</td>
<td>Inlet-associated polypeptide</td>
</tr>
<tr>
<td>Medullary carcinoma of the thyroid</td>
<td>Fragments of calcitonin</td>
</tr>
<tr>
<td>Atrial amyloidosis</td>
<td>Atrial natriuretic factor</td>
</tr>
<tr>
<td>Lysozyme amyloidosis</td>
<td>Full-length lysozyme variants</td>
</tr>
<tr>
<td>Insulin-related amyloid</td>
<td>Full-length insulin</td>
</tr>
<tr>
<td>Fibrinogen α-chain amyloidosis</td>
<td>Fibrinogen α-chain variants</td>
</tr>
<tr>
<td>Synucleinopathies</td>
<td>α-synuclein</td>
</tr>
</tbody>
</table>

Figure 1  Sequences of Aβ, IAPP and α-synuclein.

A key hallmark of AD is the misfolding of two proteins, Aβ and tau, whose extracellular and intracellular aggregates in the brain are coincident with and most likely causative of the disease. The risk of developing AD increases exponentially with age. In the early stage, AD is often overlooked as it manifests in mildly impaired short-term memory. The long-term memory is unaffected. With progression of the disease, the impairment of memory both short and long-term becomes apparent. Patients undergo a change in personality from having paranoid delusions to hallucinations and anxiety. At the late stages of AD, the affected person loses his/her ability to participate in daily life and fails to perform the most basic actions such as eating or walking.

A key hallmark of AD is the misfolding of two proteins, Aβ and tau, whose extracellular and intracellular aggregates in the brain are coincident with and most likely causative of the disease.
that expresses the Swedish mutant APP, a human tau mutant. Recently, a triple transgenic mouse model has been developed in mice develop memory loss but are devoid of neurodegeneration. Other important characteristics of AD, such as NFTs. These (the Swedish mutation) and develop amyloid plaques, but lack murine models, such as Tg2576, express mutated human APP due to the multifactorial nature of the syndrome. Existing sis of the disease makes such a theory unlikely to be pos-

To date, there is no generally accepted theory that accounts for the development of AD pathology. The multifactorial ba-
sis of the disease makes such a theory unlikely to be po-
sible in the foreseeable future. In addition to the NFTs and amyloid-hypothesis of the disease, oxidative stress, systemic levels of redox active metal ions, cardiovascular disease, the apoE4 allele and type 2 diabetes are all clear risk factors for development of AD. However, recent research supports the notion that the Ap build up may be a key event in AD and that other manifestations of the disease, like NFT forma-
tion, result from an imbalance between Ap production and Ap clearance (17).

γ-secretases. Cleavage of APP by α- and γ-secretases releases the non-amyloidogenic fragment p3. Upon cleavage of APP by β- and γ-secretases, Aβ is released. The peptide is characterized by a hydrophobic N-terminus and a hydrophobic C-terminus, corresponding to parts of the extracellular and intramembrane domains of APP, respectively. Aβ exists mainly as a 40 or 42 residue long peptide (Aβ1-40 and Aβ1-42) (Fig. 1). The C-terminally extended, more hydrophobic sequences like Aβ1-42 aggregate faster in vitro than their shorter partners, which might explain why they are more abundant in senile plaques.

Native Aβ is unstructured and its function is still uncer-
tain. Several possible roles have been suggested, including cholesterol transport, antioxidative protection and TGF-β activ-
ity (5-7). Similarly, a definite function for APP is still uniden-
tified. APP is a putative Notch-like receptor for an unknown ligand (8) and has been suggested to play a role in intercellular adhesion (9).

The presence of extracellular plaques consisting of misfolded Aβ in the brain of AD patients led to the widely accepted hypothesis that insoluble aggregates of Aβ are the toxic con-
formation of the native peptide. However, there is a poor cor-
relation between the number of mature fibrils found in brain and neuronal death (10). Studies have now shown that soluble Aβ oligomers are able to damage and kill neurons in central nervous system cultures (11), providing a direct link between Aβ and AD. More recently, dodecamers of Aβ have been iden-
tified as the causative agent in impairment of spatial memory in mice (12), although the relevance of this result in the context of other memory changes is unknown (13). These oligomeric species are most likely precursors or folding intermediates and do not emerge by a distinct pathway (14). It has already been shown that oligomers in the absence of fibrils and monomers can inhibit hippocampal long-term potentiation in rats in vivo (15).

Developing animal models for AD is an ongoing challenge due to the multifactorial nature of the syndrome. Existing marine models, such as Tg2576, express mutated human APP (the Swedish mutation) and develop amyloid plaques, but lack other important characteristics of AD, such as NFTs. These mice develop memory loss but are devoid of neurodegeneration. Recently, a triple transgenic mouse model has been developed that expresses the Swedish mutant APP, a human tau mutant, as well as a human presenilin-1 (PS1) mutant (16). These triple transgenic mice build up misfolded Aβ and tau and more accurately mimic human AD pathology.

Islet amyloid polypeptide (IAPP) and type 2 diabetes

Deposition of amyloid in pancreatic islets is a feature of type 2 diabetes in man, but the factors that contribute to this phe-
nomenon are unknown. Progressive amyloidosis results in loss of insulin-producing cells and increased disease state. The pres-
ence of amyloid depositions in pancreatic tissue was first de-
scribed over a century ago (18). However, it took much effort and time to identify the peptide component of these accumu-
lations as islet amyloid polypeptide (IAPP, also called amylin) (19, 20). IAPP is a 37 residue long peptide (Fig. 1) which is cosecreted with insulin. It originates from a longer precursor, pro-IAPP (67 amino acids). The peptide is co-stored with insulin in β-cell secretory granules and secreted together with insulin in response to β-cell stimulation, primarily glucose intake. Its na-

tive structure both intra- and extracellular prior to aggregation in vivo is still unclear, under physiological concentrations in vitro it adopts a stable random coil structure. During maturation of β-cell secretory granules, the N- and C-termini of pro-IAPP are enzymatically cleaved by pro-hormone convertases 1/3 and 2. The IAPP peptide is naturally present in humans and other species and is usually excused via the kidney. It is unclear if IAPP has any function in vivo although multiple possible roles have been suggested.

IAPP forms insoluble aggregates, confined within the islets of Langerhans, which are usually incidental with the non-insulin-
dependent type 2 diabetes. In contrast to the autoimmune disease type 1 diabetes, which is characterized by the destruction of insulin secreting cells, type 2 diabetes is a late-onset disease with decreased insulin secretion and reduced sensitivity towards insulin of peripheral tissue. Misfolded IAPP is present in over 90% of type 2 diabetes patients (as diagnosed post-mortem), but also in 15% of non-diabetics, indicating that other factors that are not diabetes related may play a role in the aggregation of the 37mer. Unlike Aβ, IAPP is present in concentrations above the critical concentration even in healthy individuals, but aggregation is prevented. It has been shown that insulin but not pro-insulin forms a complex with IAPP that stabilizes the peptide in vitro and possibly in vivo.

The role of amyloid in type 2 diabetes appears to be complex: severe islet dysfunction correlates well with extensive amylo-
odosis; however, the degree of amyloidosis can vary extremely in long term patients. The factors that trigger the onset of the disease are both genetic and environmental. On the other hand, a common qualification of misfolding diseases is that they occur mostly sporadic, without the necessity of mutations.

Mature fibrils formed from IAPP demonstrate β-cell toxicity in vitro (21). Analogous to the toxic-conformation-dissociation with Aβ, it has also been proposed that smaller, soluble oligomers of IAPP have cytotoxic effects (22) and show membrane-disrupting activity.

A link between AD and diabetes?

It has long been suggested that diabetes is a risk factor for AD, and a cohort study on people older that 55 years revealed that patients suffering from diabetes have a 65% increase in the risk
of developing AD compared to the group that did not suffer from diabetes (23). Moreover, the two diseases share similar physiological processes, most notably degeneration and age dependence. Interestingly, insulin receptors are not only found in the peripheral system, but also in neurons in the CNS (24). Therefore, the role of these receptors is not limited to the regulation of blood sugar levels but they are also involved in neuronal differentiation and cell proliferation (25, 26), implicating an intricate connection between insulin and AD. This unexpected link of AD and diabetes has led to the postulation of a new, brain specific “type 3 diabetes” (27), which was further elucidated recently when it was shown that soluble oligomers of Ab disrupt insulin signaling by binding to the insulin receptors on the neuronal surface (28). Given that synaptic failure and memory dysfunction are characteristic for AD, this link is another plausible explanation for the observed symptoms of the disease.

α-Synuclein and synucleopathies

α-Synuclein (Fig. 1) is the protein component associated with synucleinopathies, protein misfolding diseases that are characterized by the presence of α-synuclein deposits. Synucleinopathies include Parkinson disease, Lewy body (LB) dementia, diffuse LB dementia, and also the LB variant of AD.

α-Synuclein has been found to be the major constituent of Lewy bodies (LBs). Intriguingly, these deposits are not extracellular as with Aβ and IAPP, but intracellular inclusion bodies found in dopaminergic and non-dopaminergic neurons and glia. The protein morphology within the inclusions can be distinct; at least five different conformations have been described: LBs, Lewy neurites, glial cytoplasmic inclusions, neuronal cytoplasmic inclusions, and axonal spheroids.

α-Synuclein exits in three isoforms of varying length in humans. The best characterized is the full length, 140 amino acid long protein; the other isoforms are shorter and derived by alternative splicing. The N-terminus of the 140mer is characterized by the presence of a seven eleven amino acid long repeat containing amino acid long protein; the other isoforms are shorter and derived by alternative splicing. The N-terminus of the 140mer is characterized by the presence of a seven eleven amino acid long repeat containing highly conserved hexameric motifs (KTKKEGV). These repeats are thought to be implicated in the interaction with lipids and adopt a helical structure upon contact with lipid vesicles. The central region of the protein is highly prone to aggregation.

α-Synuclein is natively unstructured and a definitive function remains elusive. Possible roles include the regulation of vesicular release and other synaptic tasks as α-synuclein has been found to exist in equilibrium of free protein with vesicle-bound protein. α-Synuclein interacts in vitro with synthetic or purified phospholipid membranes, altering their integrity and suggesting participation in membrane lipid component organization. Moreover, it may also act as inhibitor of lipid organization. An intriguing function of α-synuclein is that it may operate as a molecular chaperone.

α-Synuclein is known to bind many proteins; over 500 proteins interacted in a proteomic analysis of a neuroblastoma cell line. Contact with some proteins or protein complexes promotes aggregation, e.g. fusion proteins and the tau protein, while others, particularly the shorter synucleins, γ1 and γ2, inhibit misfolding. Due to the high plasticity of the protein, it can adopt many different conformations. This conformational diversity of α-synuclein has had the protein labeled as a “protein chameleon”. Under physiological conditions in vitro the protein is unstructured. In the presence of lipid vesicles, its N-terminus can acquire α-helical structure in vitro. It can form non-covalent dimers and higher order oligomers leading to insoluble aggregates. Interestingly, α-synuclein can also form covalent dimers that crosslink via a dityrosine bond, under oxidative conditions.

Evidence points towards the oligomeric state of α-synuclein as the toxic conformation causing neuronal death. It has been suggested that fibril formation is merely an escape route or protection mechanism in the brain to prevent further neuronal damage.

Biophysical Characterization of Fibrils and Molecular Mechanism of their Formation

Although precursor proteins of amyloid are extremely diverse in sequence, the end-product fibrils in general share several structural features such as cross-β sheet assemblies, non-branched fibrils of similar length and structural organization.

Key intermediates in fibril formation

In general, there is a simple picture of the events that occur during oligomerization of natively unstructured protein and peptides: unfolded monomers undergo organized self-assembly via several intermediate assemblies on the pathway to fibrils. Many of these intermediates have been characterized and named based on in vitro studies, but there is no defined nomenclature as of yet. The following intermediates have been identified so far in the fibril polymerization process:

Monomers of the native peptide or protein can either be unfolded or folded. However, even natively unstructured peptides, e.g. Aβ, can adopt some secondary structure depending on the physiological environment, and some proteins exist in a native oligomerized state, like transthyretin. Monomers can arrange into oligomers, which are small globular, possibly spherical, micelle-like assemblies that contain some secondary structure (29). It is still unclear whether these oligomers are direct intermediates of the fibril formation pathway, or whether they occur as part of an “off-pathway”. For the misfolding of Aβ, evidence points towards oligomers being precursors of fibrils (29). Moreover, these oligomers are thought to be the neurotoxic conformation in a variety of misfolding diseases, most importantly AD (11). Structural instabilities of folding intermediates make it difficult to determine the structure of these intermediates, but not impossible (24).

Two unfortunate names that can easily lead to confusion have been given to the “in-pathway” conformations protofibrils and protofibrils. Protofibrils are similar to oligomers, mostly unstructured but linear (31). However, in most cases the term “protofibrils” is used interchangeably with the term oligomer. The name protofibrils refers to linear aggregates with β-sheet structure that have a diameter of 3nm and are usually 50-100nm in length. Mature, elongated protofibrils, also called filaments, interweave to form protein fibrils. Fibrils...
are generally unbranched, approx. 10 nm in diameter and can be several μm long.

**Nucleated polymerization and other models of fibril formation**

The actual mechanism of self-association of monomeric protein into amyloid is complex and three mechanisms of structure conversion have been proposed (32). In **templated assembly**, a monomeric or native state peptide binds to an existing nucleus. Upon binding, there is a change in the secondary structure of the monomer as it is added to the growing chain. **Monomer directed conversion** involves the presence of a misfolded monomer that templates the structure conversion of a native monomer, followed by disassociation and chain formation. The third model is **nucleated polymerization**, which is the most widely accepted model for the fibril growth.

Nucleated polymerization is a crucial process in cells, involved in providing a rigid cytoskeleton by organization of tubulin into microtubules and actin into thin filaments. These processes are under tight control; however, nucleated polymerization of proteins can also be “accidental” and occur to proteins that, in healthy individuals, remain in their native structure. The kinetics of misfolding are characterized by the presence of a lag phase, a rapid growth or elongation phase, and a plateau phase. The overall rate of amyloid aggregation is restricted by the formation of a nucleus, a process that is thought to have either slow kinetics or is driven by an unfavorable equilibrium. A nucleus is defined as the least stable conformer in the aggregation process that is in equilibrium with monomeric protein (33). The formation of a nucleus, a “stochastic” event, is concentration dependent. Nuclei only occur above a critical protein monomer concentration (for Aβ this concentration is approx. 15μM). However, in the presence of preformed seeds the lag phase is completely abolished. Once a nucleus or seed is present, it acts as a template for the conversion of monomers into β-sheet rich conformers that assemble rapidly into protofibrils and fibrils. Upon consumption of monomers there is a plateau phase where no more fibrils are formed. This general model of fibril formation is able to explain the spontaneous onset of most amyloidoses as well as their rapid progression. However, the kinetics of this mechanism can be influenced to a great extent by quite a few factors both in vitro and in vivo: pH, temperature of the sample, presence of preformed seeds, presence of metal ions, and oxidative damage.

Mathematical descriptions of the kinetics of protein misfolding have largely been attempted using non-complex simulations and algorithms. It has been suggested that a complete mathematical analysis of these processes is currently beyond our capability (34).

**Structure of fibrils**

No high resolution structure of fibrils has been published yet, as fibrils are insoluble, non-crystalline material, and therefore the classic structure determination methods like NMR and X-ray crystallography fail.

First experiments to elucidate the structure of the proteins in the fibrillar organization were done by X-ray diffraction. Amyloid fibrils show a characteristic diffraction pattern, the so called β-cross pattern (35), which is indicative of β-sheets parallel to the fibril axis with the protein strands perpendicular to the fibril’s long axis (36, 37). The patterns of amyloid is characterized by reflections at 4.75 Å (along the fibril axis) and 10 Å (perpendicular to the fibril axis) which occur from regular repeats and stacking of monomers.

Further characterization of Aβ fibrils involved solid-state NMR to determine the structure of the peptide subunits. Depending on the experimental constraints applied, several models for full length (e.g., (38)) and fragments of Aβ (e.g., (39, 40)) have been developed, most of them proposing an antiparallel, in-register β-sheet organization. However, it has since been proposed that the monomers composing a fibril associate in parallel, not antiparallel β-sheets (41) with full length Aβ and a model based on these constraints has been determined (42) (Fig. 2).

Within the IAPP polypeptide, three key regions have been identified that are responsible for its misfolding: residues 8-20 (43), 20-29 (44), and 30-37 (45). Residues 20-29 were initially found to be the amyloidogenic chore of IAPP, as peptide fragments corresponding to that sequence was found to form fibrils in vitro. Taken together, these three regions might participate in the formation of an intramolecular β-sheet with both parallel and antiparallel organization (43). Electron paramagnetic resonance spectroscopy of spin-labeled IAPP fibrils suggested a structure similar to that of Aβ, in-register, parallel β-sheet structure with units showing a disordered N-terminal segment (46).

Solid-state NMR studies of α-synuclein fibrils have shown that the β-sheet rich domain is located between at least residues 38-95, with the N-terminal residues being unstructured and...
a rigid backbone starting at residue 22 (47). These results are consistent with data obtained from electron paramagnetic resonance (EPR) of spin-labeled protein, showing an ordered core around residues 31-109, with parallel, in-register β-sheets (42, 43).

**Lessons for drug development**

To date, there is no successful treatment for misfolding diseases. The development of structural models for fibrils and the elucidation of intermediates of fibril formation give insights into possible drug targets. Several approaches are possible using medicinal chemistry in order to improve upon their properties.

**Small molecule inhibitors of Ab assembly**

Several small molecule inhibitors that prevent assembly of Aβ have been reported, however, the literature is far from systematic. Most of the studies listed in Table 2 (references for studies found in (53)) test only a few compounds and do not attempt to derive a structure-activity profile. Moreover, the chemical structures do not lend themselves to good starting points for medicinal chemistry in order to improve upon their properties. To date, the known compounds that inhibit Aβ assembly are weakly potent and are poorly bioavailable, particularly to the brain. Furthermore, progress towards development of aggregation inhibitors in the clinical setting remains elusive due to a clear disconnect between the in vitro and in vivo fibril-formation process. These points are further demonstrated by the fact that the only amyloid-aggregation inhibitors that have advanced to late-stage clinical trials, AlzheWhen™, Crebell™ and Kucta™, do not fit the classical drug-like profile and their in vivo disposition and mechanism of action is not well understood.

**Methods Used for Fibril Structure Determination and Aggregation Assays**

Research on amyloid fibrils was facilitated by the fact that fibrils similar to the ones found in vivo can be formed in vitro from peptides and denatured proteins. However, structure determination of amyloid fibrils has proven to be difficult due to the lack of methods and the complexity of the aggregates. Standard methods for structure determination, primarily solution-state NMR and X-ray diffraction, fail for amyloid fibrils due to the lack of soluble material or the inability to form classic protein crystals. Only recently we have acquired an in-depth understanding of the detailed conformation of fibrils.

The techniques to follow aggregation of amyloidogenic proteins and to determine their molecular conformation range from spectrometric assays (thioflavine T, Congo red) over spectroscopic assays (FTIR and CD) and visualization techniques (AFM and TEM) to methods that provide detailed insight into the atomic coordinates (solid-state NMR and X-ray diffraction). The principal difficulty in studying amyloid formation in vivo is doubtlessly sample preparation and protein quality, as impurities, aggregation conditions (buffer, pH, temperature, etc.) and the chosen deseeding method can have a major impact on the fibril formation kinetics as well as the structure of the fibrils.

**Spectrophotometric assays**

Congo red is a diazo dye with the name-giving characteristic red color. Amyloid stained with Congo red depicts green birefringence under polarized light and this method has been used to diagnose amyloid diseases in histological samples. To date, it is still used to classify proteins aggregates as amyloid. However, due to subjective interpretation and many incidents of “false negatives”, this method is to be used carefully and should be accompanied by a parallel verification method. Congo red can also be utilized in an absorbance assay to quantify the amount of misfolded material (54).

A straightforward method used frequently to follow the misfolding kinetics of previously desired protein is the thioflavin T assay (55). Thioflavin T is a molecule that binds rapidly to protein aggregates but not to monomeric or oligomeric intermediates. Upon binding, a red shift of the excitation and emission wavelength is monitored, allowing observing the kinetics of protein misfolding with a fluorescence microscope or by fluorescence spectroscopy. As this assay detects only β-sheet rich conformers, it is not suitable to identify unstructured intermediates of the misfolding process.

**Spectroscopic assays**

To determine the presence of β-sheet in amyloid samples, spectroscopic techniques such as circular dichroism (CD) and Fourier transform infrared (FTIR) spectroscopy are routinely used. In far U.V. CD spectroscopy, the sample is examined before and after aggregation to record any change in secondary structure. Samples of soluble protein are subjected to left and right handed circular polarized light and the difference in its absorbance at far-U.V. wavelengths is recorded. Obtained spectra are then compared to a reference library to allow an assessment of secondary structural elements. Aβ is unstructured under physiological conditions and exhibits a classic random coil far U.V. CD spectra. However, as aggregation of Aβ progresses, and the monomers fold into protofibrils and fibrils, the CD spectra change to reflect β-sheet formation.
<table>
<thead>
<tr>
<th>Compound</th>
<th>Mode of action</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aminoindanes</td>
<td>Fibrils</td>
<td><img src="image1" alt="Aminoindanes" /></td>
</tr>
<tr>
<td>Aminopyrazoles</td>
<td>Fibrils</td>
<td><img src="image2" alt="Aminopyrazoles" /></td>
</tr>
<tr>
<td>Anti-parkinsonian agents</td>
<td>Fibrils</td>
<td><img src="image3" alt="Anti-parkinsonian agents" /></td>
</tr>
<tr>
<td>Beta-sheet breakers peptides</td>
<td>Fibrils</td>
<td><img src="image4" alt="Beta-sheet breakers peptides" /></td>
</tr>
<tr>
<td>Coenzyme Q10</td>
<td>Fibrils</td>
<td><img src="image5" alt="Coenzyme Q10" /></td>
</tr>
<tr>
<td>Ferulic Acid</td>
<td>Fibrils</td>
<td><img src="image6" alt="Ferulic Acid" /></td>
</tr>
<tr>
<td>FK506 complexes</td>
<td>Fibrils</td>
<td><img src="image7" alt="FK506 complexes" /></td>
</tr>
<tr>
<td>Compound</td>
<td>Mode of action</td>
<td>Example</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td>Hydroxyindoles</td>
<td>Fibrils</td>
<td><img src="https://example.com/image1.png" alt="Image" /></td>
</tr>
<tr>
<td>Hydroxyphenyl pyridazine diamine</td>
<td>Fibrils</td>
<td><img src="https://example.com/image2.png" alt="Image" /></td>
</tr>
<tr>
<td>Melatonin</td>
<td>Fibrils</td>
<td><img src="https://example.com/image3.png" alt="Image" /></td>
</tr>
<tr>
<td>Nicotine</td>
<td>Fibrils</td>
<td><img src="https://example.com/image4.png" alt="Image" /></td>
</tr>
<tr>
<td>Nordhydroguaiaretic acid</td>
<td>Fibrils</td>
<td><img src="https://example.com/image5.png" alt="Image" /></td>
</tr>
<tr>
<td>NSAIDs</td>
<td>Fibrils</td>
<td><img src="https://example.com/image6.png" alt="Image" /></td>
</tr>
<tr>
<td>Organothiourine indol-3-yl derivatives</td>
<td>Fibrils</td>
<td><img src="https://example.com/image7.png" alt="Image" /></td>
</tr>
<tr>
<td>Compound</td>
<td>Mode of action</td>
<td>Example</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Oxazalones</td>
<td>Fibrils</td>
<td><img src="image1" alt="Oxazalones" /></td>
</tr>
<tr>
<td>Phenothiazines</td>
<td>Fibrils</td>
<td><img src="image2" alt="Phenothiazines" /></td>
</tr>
<tr>
<td>Phenoxazines</td>
<td>Fibrils</td>
<td><img src="image3" alt="Phenoxazines" /></td>
</tr>
<tr>
<td>Phenylazobenzene sulfonamides</td>
<td>Fibrils</td>
<td><img src="image4" alt="Phenylazobenzene sulfonamides" /></td>
</tr>
<tr>
<td>Porphyrins</td>
<td>Fibrils</td>
<td><img src="image5" alt="Porphyrins" /></td>
</tr>
<tr>
<td>Compound</td>
<td>Mode of action</td>
<td>Example</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td>Rifampicin and tetracycline</td>
<td>Fibrils</td>
<td><img src="image1.png" alt="Rifampicin and tetracycline" /></td>
</tr>
<tr>
<td>Rhodanines</td>
<td>Fibrils</td>
<td><img src="image2.png" alt="Rhodanines" /></td>
</tr>
<tr>
<td>Small molecule anionic sulfonates of sulfates</td>
<td>Fibrils</td>
<td><img src="image3.png" alt="Small molecule anionic sulfonates of sulfates" /></td>
</tr>
<tr>
<td>Tannic acid</td>
<td>Fibrils</td>
<td><img src="image4.png" alt="Tannic acid" /></td>
</tr>
<tr>
<td>Vitamin A</td>
<td>Fibrils</td>
<td><img src="image5.png" alt="Vitamin A" /></td>
</tr>
</tbody>
</table>
Table 2 continued

<table>
<thead>
<tr>
<th>Compound</th>
<th>Mode of action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine-related polyphenols</td>
<td>Fibrils</td>
</tr>
<tr>
<td>Curcumin and rosmarinic acid</td>
<td>Fibrils, Oligomers</td>
</tr>
<tr>
<td>Cyclohexanethiol</td>
<td>Fibrils, Oligomers</td>
</tr>
<tr>
<td>Hydroxyphenyl triazine diamine</td>
<td>Oligomers</td>
</tr>
<tr>
<td>Phenols</td>
<td>Oligomers</td>
</tr>
</tbody>
</table>

![Chemical Structures](attachment:image.png)
The presence of β-sheet in a protein solution can also be determined using FTIR by inspecting the amide I band, which occurs in the region between 1600 cm⁻¹ and 1700 cm⁻¹, and taking into account possible contributions from side chains. Amyloid fibrils typically have β-sheet peaks below 1620 cm⁻¹. FTIR is not suitable to determine atomic coordinates, but it has given detailed insights into protein structure, and its easy sample preparation and the applicability to most molecules have led to the development of many experimental techniques (reviewed in (56)).

Visualization techniques (AFM and TEM)

Interesting information about the organization of fibrils has been obtained from atomic force microscopy (AFM) and transmission electron microscopy (TEM). Both methods give high resolution images of fibrils and show that fibrils are unbranched, twisted, and several µm long. These techniques are usually used to support data obtained from Thioflavin T assays, as this method gives no information about the presence of fibrils (it only confirms presence of β-sheets).

Molecular structure determination (solid-state NMR and x-ray diffraction)

Initial X-ray experiments on amyloid fibrils defined the classic β-sheet diffraction pattern with reflections at 4.75 Å and 10 Å, indicative of β-sheets parallel to the fibril axis, and the protein strand perpendicular to the fibril’s long axis (29–31). A more detailed view into the 3D organization of a fibril can be obtained by solid-state NMR. Solid-state NMR has the advantage that it can define the spatial arrangement of both the intrachain and intermolecular configurations. Using fibrillar, lyophilized protein, the molecular packing of amyloid can be defined (37, 58). Solid-state NMR spectroscopy revealed a more detailed structure of several β-sheet analogs and truncated sequences as well as the conformation of other peptides like α-synuclein, truncated peptides of transthyretin and the prion protein (Helmus J et al Molecular conformation and dynamics of the Y145Stop variant of human prion protein in amyloid fibrils. Proc Natl Acad Sci USA. 2008 Apr 29;105(17):6284-9, in (59)). The data obtained from β-sheet fibrils allowed the development of a structural model of the fibrillar form.

References

Many diseases are caused by defects in protein trafficking. Protein trafficking diseases occur when a mutant protein is recognized by the endoplasmic reticulum (ER) quality control system (ERQC), retained in the ER, and degraded in the cytosol by the proteasome rather than being trafficked to its correct site of action. Among these diseases are cystic fibrosis, lysosomal storage diseases (Fabry, Gaucher, and Tay-Sachs), nephrogenic diabetes insipidus, oculocutaneous albinism, protein C deficiency, and many others. A characteristic of many of these diseases is that the mutant protein remains functional, but it cannot escape the stringent ER quality-control machinery, and it is retained in the ER. This characteristic suggests that pharmacological interventions that promote the correct folding of the mutant protein would enable its escape from the ER and ameliorate the symptoms of the disease. In this review, we focus on specific examples of protein trafficking diseases in pharmacological or chemical chaperones have been shown to rescue trafficking of the mutant protein.

The etiology of several diseases can be traced to defects in protein trafficking. Recent studies on several different disease mutants have shown that chemicals and small molecules can correct trafficking of these mutants. As these compounds promote correct folding in a fashion analogous to the action of molecular chaperone proteins, these compounds have been termed “correctors” or “chaperones.” Several different classes of chaperones can be designated based on their mechanisms of action. Chemical chaperones are the least specific and the least potent, and which often require millimolar concentrations to function. Pharmacological chaperones are small-molecule correctors that can be subdivided into two classes based on their specificities. Specialized pharmacological chaperones are protein- or mutation-specific small molecules that interact directly with the mutant protein to provide a folding template, such as enzyme active-site inhibitors. Generalized correctors are less specific and are likely to function on the endoplasmic reticulum (ER) retention machinery, ER-associated degradation (ERAD), or other signaling pathways involved in trafficking rather than through a direct interaction with mutant proteins. We will begin with a brief overview of ER quality control, followed by specific examples of protein trafficking diseases and then discuss the different classes of correctors below.

Quality Control in the ER

Proteins that travel along the secretory pathway are subject to many quality-control checkpoints. Inside the ER, proteins must fold into their proper conformation before being sorted into vesicles destined for the Golgi apparatus. Once proteins have translocated into the ER through the translocon (Fig. 1), molecular chaperones, which include the heat shock chaperone family (e.g., Hsp70, BiP), the lectins calnexin and calreticulin, and the oxidoreductases (e.g., PDI), act on the nascent chain to promote the correctly folded conformation. Together with the oxido- reductase ERP57, the lectin chaperones act on the nascent protein in cycles of binding and release through the recognition of a monoglycosylated N-glycan. These cycles are controlled by N-glycan-modifying proteins including UDP-glucose:glycoprotein glucosyltransferase (UGGT) and glucosidases I and II (Fig. 1). Once the protein is correctly folded, the N-glycan is further modified by glucosidase II and proceeds along the secretory pathway. If several cycles of lectin binding do not result in a correctly folded protein, then the α(1→2)-ER mannosidase I (ManI) modifies the N-glycan, which...
Protein Trafficking Diseases. Small Molecule Approaches to

Figure 1  Protein folding in the ER is summarized. Folding pathways that lead to secretion are depicted in the upper half of the ER. Nonproductive folding and its consequences are depicted in the lower half of the ER. For simplicity, some pathways have been omitted, and some proteins are indicated by text alone. Once a protein enters the ER through the translocon, it undergoes N-glycosylation and is acted on by several chaperones including BiP. Glucosidases I and II (GI/GII) remove the terminal glucose on the N-glycan moiety which enables recognition of the protein by calnexin (CNX) or calreticulin (not shown). The oxidoreductase ERp57 is bound to CNX and also acts on the protein. If the resulting protein has a non-native fold, then UGGT recognizes it and adds a terminal glucose to enable rebinding with CNX. Once the protein has folded properly, the N-glycan is modified even more by GI and can be recognized by the Sec24 protein for COPII-dependent secretion. If the protein cannot be folded natively after several cycles of CNX binding, then ManI removes the terminal mannose on the N-glycan which prevents UGGT from reglycosylating it and targets it for destruction through ERAD in a process that involves the EDEM, Derlin, and the p97/valosin-containing protein (VCP) (1). In some cases, the misfolded protein aggregates and accumulates in the ER, which triggers the unfolded protein response.

Three different classes of ERAD-targeting components function depending on the location of the lesion: a cytosolic group, a lumenal group, and a transmembrane group. These proteins somehow sense the folding status of their substrates and triage those with defects for degradation. However, the mechanisms that regulate the decision to undergo ERAD instead of forward transport in the secretory pathway are not known. It is known that to be recruited to coat protomer II (COPII) vesicles...
leave the ER, proteins must be recognized by the Sec24 machin-ery, either directly through an ER exit code or through another receptor that can then interact with Sec24. Several different ER exit codes have been identified, including cytosolic diacidic codes such as the one found in the cystic fibrosis transmembrane conductance regulator (CFTR) protein, ditydrophobic motifs like that in ER-Ga1–intermediate compartment protein 53 (3), and various hydrophobic signals that are found in G-protein coupled receptors (GPCRs) (4). In yeast, simply inhibiting ERAD genetically does not permit ER-retained proteins to enter COPII vesicles, which suggests that some other level of recognition by the ERQC exists. However, even when folding of substrates has improved as judged by tryptic-sensitivity assays, inhibition of ERAD still does not permit entry into COPII vesicles (5). This discrepancy could be explained by the masking of the ER exit signal in the misfolded protein. In yeast, carboxypeptidase Y mutants whose ER exit signal is not obscured can escape ERAD even though other parts of the protein are misfolded (6). For many ER-retained disease mutant proteins, the ERQC performs its function too well and retains otherwise functional proteins. Small molecules that function in any process described above could potentially correct the trafficking of these ER-retained proteins and would be useful therapeutics.

Protein Trafficking Diseases

As many ER-retained mutant proteins are functional, rerouting them to their appropriate subcellular localization would restore the phenotype caused by their mislocalization. In some cases, it is estimated that only a small amount of functional protein (10–15%) is necessary to maintain health. Indeed, for some diseases a critical threshold seems to exist (7). Patients who express functional protein above this threshold have mild or no symptoms of the disease (7, 8), which suggests that even a modest increase in protein rerouting would improve the quality of life for many patients with these diseases. This finding provides the impetus for identifying correctors, even those with modest effects.

Properties of correctable protein trafficking disease mutants

Mutations that destroy protein activity such as those that abolish enzyme active sites are not amenable to rescue with correctors. However, mutations that alter the stability or folding of the protein, but that do not abrogate its activity, are good candidates for rescue by correctors. Most ER-retained missense mutations identified to date fall in this class. Frequently, these mutations are temperature sensitive and show some degree of correction at permissive temperatures (usually ~30°C). Another important feature of diseases that are amenable to rescue with correctors is the degradability of accumulated substrates or byproducts. For example, the lysosomal storage diseases Fabry, Gaucher, and Tay-Sachs all result in substrate accumulation in the lysosome; however, after treatment with pharmacological chaperones and rerouting of the enzymes, the corrected proteins reduce these substrate stores (9–12). In contrast, ER-retention diseases that result in the deposit of toxic or non-native proteins, such as fibrils associated with amyloidogenesis, likely would not be amenable to rescue with correctors because of the cell’s inability to degrade the non-native deposits, unless treatment with the corrector begins before irreversible damage is done. Several examples of trafficking diseases and their correctors are described below.

CFTR delta F508

One of the best-studied examples of a mutation that is amenable to rescue is the deltaF508 mutation in the chloride channel CFTR, which is the gene responsible for cystic fibrosis (13). Cystic fibrosis (CF) is the most common autosomal recessive genetic disease that affects the Caucasian population. It is a lethal disease characterized by severe dehydration of the cells lining the lung, intestine, and exocrine tissues (13). Most cases of CF can be attributed to a single mutation resulting in the deletion of a phenylalanine codon at position 508 in the protein, which is known as deltaF508. DeltaF508 CFTR can be rescued by incubation at permissive temperature (i.e., 27–30°C), by the addition of chemical chaperones such as sodium 4-phenylbutyrate, or other cosolutes such as glycerol, dimethyl sulfoxide, and trimethylamine N-oxide (Fig. 2a) (14–16). Modulation of ER calcium levels with curcumin and thapsigargin (Fig. 2b) has also been shown to correct deltaF508 CFTR trafficking; however, these results remain controversial (17, 18). Recent high-throughput screens have identified several classes of small-molecule correctors of protein trafficking, which include aminobenzothiazoles, aminobenzothiazoles, quinazolinylaminopyrimidinones, bisaminomethylbithiazoles, quinazolinones, khivorins and substituted 1-phenylsulfonpyridazines (19–22).

Three different assays were used to screen for correctors. One assay measured directly the amount of protein that trafficked to the cell membrane using immunofluorescence against an extracellular epitope tag in deltaF508 CFTR (19). The second assay monitored deltaF508 CFTR function by iodide influx with a halide-sensitive YFP construct as readout (20). The third assay also monitored function but through fluorescence energy transfer between a membrane-soluble voltage-sensitive dye bis-(1,2-dibutyryl)barbituric acid (tribromethine oxonol [DBS]AC(2)) and a plasma membrane localized fluorescent coumarin-linked phospholipid CC2-DMPE (22). The compounds identified in these screens are likely to function at different steps of the deltaF508 CFTR folding pathway. Corr4a, which is a bisaminomethylbithiazole, was shown to increase the folding efficiency of deltaF508 CFTR (20). Both Corr4a and the quinazolione VRT-325 were shown to delay ERAD of deltaF508 CFTR (20, 22). Compounds of the aminobenzothiazole class were found to act after ER folding as no increase in folding efficiency was detected, but the stability of deltaF508 was increased at the cell surface (20). VRT-325 also increased the stability of deltaF508 CFTR at the cell surface (22). This compound was shown to correct other CFTR mutants and even a mutant in the human ether-a-go-go-related gene (HERG) (see below). Derivatives of some of these compounds are now in clinical trials.
Protein Trafficking Diseases, Small Molecule Approaches to

O

Na

sodium 4-phenylbutyrate
glycerol
dimethylsulfoxide
trimethylamine N-oxide

(a)

thapsigargin
pyrimethamine
astemizole
cisapride
N-(n-nonyl)deoxynojirimycin

(b)
kifunensine
castanospermine
miglustat
mevastatin

Figure 2 Structures of selected chemical (a) and pharmacological (b) chaperones are shown.

Lysosomal storage diseases

The lysosomal storage diseases, such as Tay-Sachs, Sandhoff, Gaucher, and Fabry disease, are autosomal recessive loss-of-function disorders. The mutant enzymes fail to degrade their respective lysosomal substrates because of their retention in the ER and subsequent degradation. This result leads to an accumulation of substrates and a variety of phenotypes including enlargement of affected organs, skeletal lesions, neurological abnormalities, and premature death. The current therapies for these diseases include inhibition of substrate production and enzyme-replacement therapy. However, enzyme-replacement therapy is not suitable for neurological phenotypes associated with some types of these diseases because of the impermeability of the blood–brain barrier; hence, other treatments are necessary. Recently, pharmacological chaperones for these diseases have been identified. As these proteins are enzymes, the pharmacological chaperones identified tend to be competitive active-site inhibitors. This finding may seem counterintuitive, but inhibitors are frequently trafficking correctors at subinhibitory doses (see below). The rationalization is that inhibitors stabilize the fold of the mutant proteins at the neutral pH of the ER and allow them to evade the ERQC. When the proteins reach the lysosome, the high concentration of the substrates successfully compete away inhibitor binding thereby enabling degradation of the substrates to occur. Some examples of correctors for each of these diseases are discussed below.

Several active-site specific chaperones have been identified for beta-hexosaminidase A, which is the multisubunit enzyme responsible for Tay-Sachs and Sandhoff diseases (9, 11, 23). The screening assays monitor inhibition of purified enzyme activity fluorometrically using the fluorescent artificial substrate 4-methylumbelliferyl-β-N-acetylglucosamine. These assays are then followed by cell-based assays to determine the amount of functional protein that traffics to the lysosome. Several distinct structural classes of pharmacological chaperones have been identified for these diseases including aza-sugars, pyrimethamine (Fig. 2b), and substituted bicyclic and tricyclic nitrogen-containing heterocycles (9, 11, 23).

Gaucher disease is caused by a deficiency in lysosomal beta-glucosidase activity. Deoxynojirimycins are known inhibitors of several enzymes, which include beta-glucosidase. Several alkylated deoxynojirimycins were screened for correction of N370S Gaucher mutant lysosomal beta-glucosidase activity in an intact cell assay (10). N-(n-nonyl)deoxynojirimycin
Protein Trafficking Diseases: Small Molecule Approaches to hERG Mutants

hERG encodes the pore-forming subunit of the rapidly activating delayed rectifier potassium channel. Mutations or drug treatments that cause retention of hERG in the ER result in cardiac arrhythmias that can lead to sudden death. Hence, drugs are now routinely screened for their effects on hERG function before being pursued in clinical trials. Several ER-retained hERG mutants can be rescued by chemical and pharmacological chaperones. The G601S and N470D mutations can be rescued by growth at permissive temperature or incubation with known hERG channel blockers E-4031, cisapride, and astemizole (Fig. 2b) (38, 39). hERG G601S has recently been used in a small molecule screen to identify correctors of hERG trafficking (40). This assay measured the amount of hERG G601S that trafficked to the cell surface by chemiluminescent detection of an extracellular epitope tag. Several different hERG blockers were found to rescue the ER retained mutant, which is consistent with their action as pharmacological chaperones. Other compounds have also been shown to rescue the trafficking of certain hERG mutants selectively. For example, the sarcoplasmic/ER calcium ATPase (SERCA) inhibitor thapsigargin can rescue the trafficking of G601S and FB60C mutants, but not N470D, whereas E-4031 rescues G601S and N470D mutants but not FB60C mutants (41).

G protein-coupled receptors

Many diseases caused by mutations in GPCRs are also the result of ER retention. Mutations in the vasopressin 2 receptor result in nephrogenic diabetes insipidus, which is a disease characterized by the kidney’s inability to concentrate urine. Trafficking of several different ER-retained mutants can be corrected by small nonpeptide V2R and V1R antagonists (42–46). The V206D mutation can also be corrected with glycerol, DMSO, SERCA inhibitors thapsigargin and curcumin, and the calcium ionophore ionomycin (44). This mutation cannot be corrected by growth at permissive temperature or with the addition of 4-phenylbutyrate (44). In contrast, the A98P, L274P, and R113W mutations can be corrected with the osmolytes trimethylamine N-oxide and DMSO, as well as with growth at permissive temperature (47). However, glycerol treatment does not correct the trafficking of these mutations.

Mutations in another GPCR, which is the gonadotropin-releasing hormone receptor, result in hypogonadotropic hypogonadism. Seventeen mutations in this gene have been characterized as misfolding or misrouting mutants, and most of these can be rescued by incubation with peptidomimetic antagonists (48). These pharmacological chaperones are indoles, quinolones, and erythromycin macrolides. Another example of a GPCR with ER-retained disease mutations is rhodopsin, which when mutated causes retinitis pigmentosa (49). A mild rhodopsin mutant P23H associated with night blindness can be rescued by treatment with 11-cis-retinal, which is its covalently bound chromophore (49).

Alpha-1-antitrypsin Z variant

Alpha-1-antitrypsin mutations are associated with early-onset emphysema and liver disease that results in early death (50). The Z-variant of the disease is the most common and is present in over 95% of cases. This variant has been shown to be rescued in vitro by treatment with proteasome inhibitors (50). Several chemical and pharmacological chaperones have also been shown to correct the secretion of the mutant protein, which include the glucosidase inhibitor castanospermine, as well as the mannosidase inhibitors tifensulfone (Fig. 2b).
and 1,4-dideoxy-1,4-imino-D-mannitol hydrochloride (51). Incubation with the chemical chaperone 4-phenylbutyrate also results in increased trafficking. However, growth at permissive temperature does not result in trafficking correction, although it does decrease the amount of mutant protein that becomes degraded (52).

Correctors of Protein Trafficking

Chemical chaperones

Several different classes of chaperones can be designated based on their mechanisms of action. Chemical chaperones are those that act nonspecifically. Although these chaperones are nonspecific, their use as therapeutics may be limited because of the high concentrations required to achieve correction. The mechanisms of correction employed by chemical chaperones may include the induction of molecular chaperone transcription or the nonspecific stabilization of proteins through masking of hydrophobic domains. For example, 4-phenylbutyrate (4-PBA) is a histone deacetylase inhibitor that activates transcription of different genes including the heat shock proteins (53). Studies on deltaF508 CFTR suggest that 4-PBA reduces the protein levels of the constitutive Hsc70 chaperone (54). This in turn reduces the amount of deltaF508 CFTR that interacts with Hsc70. The decrease in Hsc70 protein levels induces an increase in Hsp70 levels, which has been proposed to be a more effective chaperone, thus helping deltaF508 CFTR to fold and enabling it to escape from the ERQC (55).

Another common mechanism of correction is ER glucosidase inhibition (Fig. 1, star 4), because several ER chaperones require calcium for function. Interestingly, for lysosomal storage diseases, trafficking correction occurs through changing the levels of calcium in the cytosol and not in the ER. These two different mechanisms of correction highlight the importance of calcium signaling in protein trafficking.

Generalized pharmacological chaperones

A further subclass of pharmacological chaperones includes compounds that act through other mechanisms. This subclass includes generalized chaperones—those that rescue the trafficking defects of many or all ER-retained proteins through a universal mechanism. Many different mechanisms of correction are possible. Examples of this subclass of chaperones are described below.

One might expect that simply inhibiting ERAD with proteasome inhibitors would be the most generalized form of correction (Fig. 1, star 3); however, this strategy gives inconsistent results with different ER-retained mutants. The NERG Y611H mutant is not corrected, deltaF508 CFTR is only weakly corrected and alpha-1-antitrypsin Z variant is corrected as visualized by confocal microscopy (50, 59, 60).

As discussed above, for many ER-retained proteins, altering the intracellular calcium levels with SERCA inhibitors or with calcium ionophores corrects trafficking of the mutant proteins, presumably by altering the ER’s capacity for folding (Fig. 1, star 4). However, because several ER chaperones require calcium for function. Interestingly, for lysosomal storage diseases, trafficking correction occurs through changing the levels of calcium in the cytosol and not in the ER. These two different mechanisms of correction highlight the importance of calcium signaling in protein trafficking.

Another common mechanism of correction is ER glucosidase inhibition (Fig. 1, star 2), which prevents the recognition of misfolded proteins by calnexin and calreticulin and presumably allows them to escape the ERQC. Inhibition of glucosidase with castanospermine and miglustat were shown to correct the trafficking of deltaF508 CFTR and alpha-1-antitrypsin (51, 61). Whether this mechanism also works on other ER-retained mutants remains to be determined.

For other small molecules, the mechanism of action is still unclear. For example, silfenal and structural analogs have recently been shown to correct trafficking of deltaF508 CFTR mutants (19, 21). Silidenal is an inhibitor of phosphodiesterase activity; however, the link between this function and CFTR trafficking correction remains to be elucidated.

Pharmacological chaperones may also show mutation-specific profiles. For example, two common mutations in Gaucher disease N370S and G202R are both localized to the catalytic domain of the protein and are both corrected by active-site substrate/cofactor L-Dopa (58). Some specialized pharmacological chaperones are not inhibitors, such as the VRT-325 and Corr4a correctors of deltaF508 CFTR (20). VRT-325 is an interesting compound as it has also been shown to correct trafficking of P-glycoprotein, a protein related to CFTR, and the structurally unrelated G6015 mutant of the NERG potassium channel. This compound has not yet been demonstrated to interact directly with NERG.
Conclusions

The etiology of many genetic diseases can be traced to defects in protein trafficking. Several mutants are functional but are retained in the ER because of the overly stringent ERQC. Corrector compounds that permit the escape of ER-retained mutant proteins from the ERQC have great potential as therapeutics. Several types of trafficking correctors have been identified. Chemical chaperones are nonspecific and require high concentrations to be effective, which thereby limits their potential as therapeutics. Pharmacological chaperones function at much lower concentrations and show greater promise for drug development. Pharmacological chaperones can be classified into specialized chaperones that are protein or mutation-specific or generalized chaperones that can correct the trafficking of many ER-retained mutants. The identification of a generalized pharmacological chaperone that can correct all ER-retained disease mutants without grossly affecting normal proteins would enable treatment of many different protein trafficking diseases. Defining the mechanisms of action for this class of pharmacological chaperones will be the next milestone in trafficking disease research.

Additive effects of correctors have recently been shown for the deltaF508 CFTR mutant (62), which suggests that treatments that combine two or more chaperones may be a good option if no single potent corrector can be identified for a particular trafficking disease. Indeed combinatorial studies are likely to be the next key area of screening for correctors of protein trafficking diseases.

Acknowledgments

Research on CFTR trafficking in the Thomas lab is supported by grants from the Canadian Cystic Fibrosis Foundation, Inc., and the Canadian Institutes of Health Research. H.M.S. is supported by a fellowship from the Canadian Cystic Fibrosis Foundation.

References


Protein Trafficking Diseases: Small Molecule Approaches to


Further Reading


Saliva in Health and Disease, Chemical Biology of

Tibor K. Fábián and Pál Fejérdy, Semmelweis University Budapest, Faculty of Dentistry, Clinic of Prosthetic Dentistry, Budapest, Hungary

Péter Csermely, Semmelweis University Budapest, Faculty of Medicine, Institute of Medical Chemistry Molecular Biology and Pathobioc hemistry, Budapest, Hungary
doi: 10.1002/9780470048672.wecb643

Saliva is a bodily fluid secreted by three pairs of major salivary glands (parotid submandibular and sublingual) and by many of minor salivary glands. Saliva is supplemented with several constituents that originate from blood serum, from intact or destroyed mucosal and immune cells, and from intact or destroyed oral microorganisms that result in a complex mixture of a variety of molecules. Saliva plays an important role in acquired pellicle formation on tooth surfaces, crystal growth homeostasis, bacterial adhesion, plaque formation, and—because of its lubricating effect—in maintaining mucosal integrity of the oral and upper gastrointestinal mucosal surfaces. It also plays an important role in physico-chemical defense, antimicrobial defense, and wound healing. Many saliva constituents including proteins, carbohydrates, lipids, and ions interact under fine regulation to fulfill these important tasks. Local and systemic disorders may disturb and interrupt these complex balanced functions, which can lead to mucosal and tooth damages. In other cases, systemic disorders induce salivary changes without any significant local effects. Many such changes are of high diagnostic interest because they can be rather specific to the causing conditions and can be used for screening and early diagnosis of several local and systemic disorders.

Saliva is a major determinant of the oral environment and serves as an easily available diagnostic tool of systemic conditions. Consequently, more intense saliva research can be observed in recent decades, which leads to a high amount of scientific data presented by numerous engaged researchers of this far-reaching field. With the rapid growth of knowledge, a need exists to summarize the obtained data of this interesting field. This article provides a brief introduction to the most important aspects of the chemical biology of saliva.

Origin of Salivary Chemical Components

Salivary components can originate from several sources, which leads to a rather complex collection of molecules. To understand the importance and meaning of a certain component, it is crucial to know the origin and the excretion mechanism of the component.

Constituents of salivary gland origin

Constituents of salivary gland origin (i.e., water, ions, proteins, carbohydrates, lipids) can be released from major salivary glands such as parotid, submandibular, and sublingual glands, and from minor salivary glands of the labial, buccal, lingual, palatoglossal, and palatal mucosa (1).

Based on the features of secreted primary saliva, secretory endpieces (acini) of salivary glands can be characterized as serous, seromucous, or mucous. The acini of parotid gland are mainly serous and seromucous, those of submandibular gland are mainly seromucous, and those of sublingual gland are mainly mucous. The acini of the minor salivary glands have various features that depend on their location (1).
Primary saliva is modified in the intercalated, striated, and excretory (collecting) ducts that lead from the acini to the mouth. Water and electrolyte transport into the saliva is believed to occur in the intercalated ducts. Striated ducts are responsible for electrolyte transport such as secretion of potassium and reabsorption of sodium ions. A transport of proteins like IgA, lysozyme, and kallikrein (and may be Hsp70 (2)) probably exists in the striated duct as well. An electrolyte transport in the excretory (collecting) ducts is also suspected (3).

**Constituents of other origin**

Each blood constituent may enter the oral cavity via intraoral bleeding. Serum exudates also reach saliva either from the gingival crevicular fluid or through the oral mucosa (mucosal transudate) and from the salivary glands via transcellular diffusion and ultrafiltration (via tight junctions) (3). Oral microbes and their fermentation products, enzymes, RNA, DNA, and structural elements are also usual constituents of the saliva (4). Fragments from the keratinized mucosal surfaces, mucosal cells with intact cell organelles from nonkeratinized surfaces, and some immune cells are also present. Cellular fragments, cytoplasmic products, enzymes, structural elements, membranes, RNA, and DNA of these cells are also usual salivary constituents. Certain amount of expectorated bronchial and nasal secretum, constituents of foods, administered drugs, smoke (from smoking), toothpastes, mouth rinses, and molecules released from dentures can also be found (2, 4).

**Molecular Participants of Saliva**

**Chemical Biology**

According to the above, many constituents exist in human saliva, and it is of scientific interest, only some have the focus of our attention. The most important constitutes are summarized below.

**Inorganic components**

Water is the most abundant constituent of saliva (~94%). The pH value of resting whole saliva is slightly acidic, which varies between pH 5.75 and 7.05, and it increases with increasing flow rate up to pH 8. Besides flow rate, the pH also depends on the concentration of salivary proteins, bicarbonate (HCO$_3^-$) and phosphate (PO$_4^{3-}$) ions that have considerable buffering capacity. Bicarbonate concentration is ~5-10 mM/L in resting conditions, and it may increase up to 40-60 mM/L with stimulation, whereas phosphate concentration is ~4-5 mM/L in saliva rather independently from the flow rate (5). Besides bicarbonate and phosphate, a significant amount of other ions are present to maintain the slightly hypotonic osmolarity of saliva. The most important ions are sodium (1-5 mM/L resting; 1–100 mM/L stimulated), chloride (5 mM/L resting; 70 mM/L stimulated), potassium (15 mM/L resting; 30–40 mM/L stimulated), calcium (1.0 mM/L resting; 3–4 mM/L stimulated), and many other ions such as ammonium (NH$_4^+$), bromide, copper, fluoride, iodide, lithium, magnesium, nitrate (NO$_3^-$), perchlorate (ClO$_4^-$), thiocyanate (SCN$^-$), and so on. They can be found in the saliva in lower concentrations (4, 5). (Data are summarized in Table 1).

**Proteins**

Human whole saliva has a protein content of about 0.5 to 3 mg/mL, and parotid saliva has a protein content of about 0.4 to 4 mg/mL, whereas submandibular and sublingual saliva of about 0.6 to 1.5 mg/mL. The protein concentration is more or less stable and independent from the flow rate (5). Besides maintaining osmolarity and buffer capacity, salivary proteins are also involved in several specific functions. The number of distinct salivary proteins is roughly between 100 and 140 (6, 7), from which roughly 30-40 % are produced by the salivary glands, whereas other proteins are originate from serum, from mucosal and/or immune cells, or from microorganisms (6). The most important proteins of glandular origin are alpha-amylase, glycoproteins with blood-group substances, cystatins, epidermal growth factor (EGF), gluten, histatins (HRPs), lactoferrine, lysozyme, mucins (MUC5B, MUC7), old terms: MGI, MGO), salivary peroxidase, proline-rich proteins (PRPs) and statherin. The most important serum derived proteins are albumin, alpha-antitrypsin, blood-clotting factors (VIII; IXa; XI) and members of the fibrinolytic system (proactivators, traces of plasminogen activator). Most important proteins that originate from immune cells are myeloperoxidase, calprotectin (Ca$^{2+}$ binding L1 leukocyte protein), cathepsin G, defensins, elastase, immunoglobulins (90% to 98% sIgA, 1% to 10% IgG, a few IgM, IgD, IgE). Finally, the most important protein constituents of microbial-unknown or mixed origin are alpha-2-macroglobulin, cystein peptides, DNases, RNases, kallikrein, secretory leukocyte protease inhibitor (SLPI), fibronectin, molecular chaperone (Hsp70), and streptococcal inhibitor. (Data are summarized in Table 2).

**Carbohydrates**

A significant amount of protein bound carbohydrates exists in the saliva. Some proteins may contain carbohydrates up to 80% of the molecule (i.e., MUC5B mucins) but 10-40% of carbohydrate moiety is rather usual in the case of any glycoproteins (4). The most important constituents are aminosugars, galactose, mannose, and sialic acid (N-acetylmuramic acid). Carbohydrate chains of mucins contain predominantly sialic acid and sulphate residues, although chains with blood group antigen properties contain about equal amounts of 6-deoxygalactose.
Saliva in Health and Disease, Chemical Biology of Saliva in Health and Disease

Table 2 Origin of the most important salivary proteins

<table>
<thead>
<tr>
<th>Glandular origin</th>
<th>Serum derived</th>
<th>Immune cell</th>
<th>Bacterial, unknown, mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha-amylase</td>
<td>Albumin</td>
<td>Metyloperoxidase</td>
<td>Alpha-1-macroglobulin</td>
</tr>
<tr>
<td>Blood-group proteins</td>
<td>Alpha-antitrypsin</td>
<td>Calprotectin</td>
<td>Cystatin peptidases</td>
</tr>
<tr>
<td>Cystatins</td>
<td>Blood-clotting factors</td>
<td>Cathepsin G</td>
<td>Cystein peptides</td>
</tr>
<tr>
<td>EGF</td>
<td>Lysozyme</td>
<td>Defensins</td>
<td>Defensins</td>
</tr>
<tr>
<td>Gustin</td>
<td>Mucins</td>
<td>Elastase</td>
<td>Elastase</td>
</tr>
<tr>
<td>Ractoferrine</td>
<td>Salivary peroxidase</td>
<td>Immunoglobulins</td>
<td>Fibrinogen</td>
</tr>
<tr>
<td>Lactoferrin</td>
<td>Proline-rich proteins</td>
<td>Immunoglobulins</td>
<td>Salivary chaperon Hsp70</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>Statherin</td>
<td>Immunoglobulins</td>
<td>Streptococcal inhibitor</td>
</tr>
<tr>
<td>Mucins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salivary peroxidase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proline-rich proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statherin</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...aceaminoglycosamine, galactosamin, and galactose (5). Other usual constituents of the carbohydrate chains of salivary glycoproteins are also N-acetylglucosamin, N-acetylgalactosamin, and glucuronic acid (5). The total amount of protein-bound carbohydrates in the saliva is 300–400 µg/mL, of which the amount of sialic acid is usually about 50 µg/mL (up to 100 µg/mL (8)). The most important function of protein-bound carbohydrates is the increase of viscoelasticity of the saliva, prevention of proteolysis through holding proteases at a distance, prevention of acid precipitation in case of several glycoporteins (i.e., acid soluble blood group antigens, mucins), and labeling/antigen function.

Lipids

Whole saliva contains about 10–100-µg/mL lipids (9). The most frequent lipids in the saliva are glycolipids (i.e., neutral and sulphated glyceroglucolipids), neutral lipids (i.e., free fatty acids, cholesterol ester, triglycerides, and cholesterol), and a somewhat lower portion of phospholipids (i.e., phosphatidylethanolamine, phosphatidylcholine, sphingomyelin, and phosphatidylserine) (10). Salivary lipids are mostly of glandular origin, but some (such as cholesterol and may be some fatty acids) are believed to diffuse directly from serum (11). Lipids originate from several membranes such as secretory vesicles, microsomes, lipid rafts, and other plasma and intracellular membrane fragments of lysed cells and bacteria, although the lower percentage of phospholipids indicate that the salivary lipids are not primarily of membrane origin (9). A large portion of salivary lipids is associated with proteins, especially to high molecular weight glycoproteins (i.e., mucins) and to PRPs (12). Salivary lipids may play a role in the acquired pellicle, dental plaque, calculus, sialolith, and caries formation.

Other molecules

As mentioned, many other molecules exist in the saliva, including nucleic acids (RNA, DNA), several hormones, growth factors and neurotransmitters, amino acids and their derivatives, urea, lactate, citrate, vitamins, creatinine, prostaglandins, several drugs, and chemical constituents of foods, cosmetics, tooth pastes, dental materials, and several other molecules originated from body and environment.

Chemical Biology of Saliva in Health and Disease

Saliva constituents play a role in several oral processes, and they perform important defense functions in the oral cavity. Moreover, saliva may be used for diagnostic purposes. The most important knowledge related to these fields will be summarized briefly in the following section.

Saliva and bacterial adhesion

The basis of bacterial adhesion is given by the acquired pellicle formation on tooth surfaces. This pellicle is a thin (~0.5–1 µm) layer of several salivary proteins with calcium hydroxide-binding properties. The most important such proteins are salivary amylase, cystatins (S, SA, and SN type), histatin (HRP1), mucin (MG1), acidic PRPs, statherin, and immunoglobulins (sIgA) (4, 7).

The surface binding of these proteins occurs mostly through ionic interaction of positively charged groups of the proteins’ polypeptide chain and the negatively charged tooth surface (globular proteins wring on the tooth surface during binding). Although tooth surface is negatively charged, in some cases negatively charged protein regions are responsible for binding (i.e., N-terminal region of PRPs). Calcium bridging (Ca²⁺ complex formation) between the negatively charged groups may be a mechanism of such binding (13).

First, bacterial adhesion (usually gram-positive cocci and filamentous bacteria) occurs primarily through a Ca²⁺ complex formation between carboxyl (COO⁻) and phosphate (HPO₄⁻) groups of bacterial surface and acquired pellicle. Although van der Waals’ forces and repulsive electrostatic forces are also present, some specific bacterial surface proteins also serve as adhesins for specific receptors on acquired pellicle. Pellicle-integrated immunoglobulins also bind bacteria specifically.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Saliva and bacterial biofilm (plaque) formation

After the adhesion of the first layer of bacteria (i.e., Streptococcus mutans) the bacterial accumulation process is initiated by the activity of secreted extracellular glucosyltransferases (GTFs) of S. mutans. In the presence of sucrose, GTFs synthesize several forms of high-molecular-weight branched extracellular glucans (i.e., dextran), which leads to sticky polysaccharide products resulted in stronger binding to the surface and facilitation of adhesion of more bacteria via glucan-binding proteins of the bacterial surface. Food rests in the saliva may serve as store of sucrose during this process (4).

Saliva and crystal growth homeostasis

In general, saliva (as well as plaque fluid) is supersaturated with respect to calcium-phosphate salts, and they prevent tendency to dissolve minor crystals of teeth. Moreover, precipitation of calcium-phosphate salts that include hydroxyapatite may also occur (remineralization) in early lesions of tooth surfaces injured by acidic bacterial products (i.e., lactic acid). Salivary fluoride facilitates calcium-phosphate precipitation, and such crystals (i.e., fluorapatite) show lower acid solubility properties that lead to an increased caries preventive effect. The increase of pH (i.e., buffer capacity and pH of saliva, as well as ureolysis in dental plaque) also facilitates crystal precipitation and remineralization (4, 13).

Similarly, supersaturated of saliva with respect to calcium-phosphate salts is the driving force of calculus (i.e., mineralized dental plaque) and salivolith (i.e., salivary duct “stones”) formation. In these cases, negatively charged phospholipids play a crucial role. Ca$^{2+}$ ions bind to the negative charges of such lipids, and inorganic phosphate associates with the bound calcium that forms a Ca-phosphate-phospholipid complex, which is an excellent nucleus of calcium-phosphate deposition. Salivary proteins may also play a role in this process because such complex formation occurs predominantly on lipids that are protein associated. The increase of pH facilitates these processes (13).

Because calcium-phosphate precipitation would lead to a “confluent growth” of tooth surfaces and intensive formation of dental calculus and salivolith, the precipitation must be controlled. For such purposes, calcium and/or hydroxyapatite-binding proteins such as calprotectin, histatins (HRP1), statherin, acidic PRPs, and cystatins (5, 14, SN type) are present in the saliva. All inhibit crystal growth, whereas statherin also inhibits spontaneous unseeded precipitation (nucleation inhibition). Interestingly, dental plaque-bound immunoglobulins also inhibit crystal growth during calculus formation. Mg$^{2+}$ ions also have some nucleation inhibitory effect (4, 13).

Saliva and surface protection

Besides taphant in acquired pellicle formation on tooth (den- ture, implant) surfaces, MUC5B type mucins cover all oral surfaces with a 10-20-μm thick layer. In addition, MUC5B type mucins form a hydrophilic viscoelastic gel (already in low concentration) that causes a high viscosity matrix of saliva. These properties of mucins (MUC5B), together with similar effects of glycosilated PRPs, accomplish the lubricating effect of saliva that defends against physical injuries during chewing (4, 14). Salivary proteins, especially basic PRPs, bicarbonate ions, and phosphate ions may also act as buffers against acids of nourishment and/or bacterial fermentation. PRPs and especially HRP5s are potent precipitators of trinys, which are a widespread phenolic plant compound (flavonoid) of nourishment with unpleasant taste and protein precipitating properties (15). Protease inhibitory effect of saliva (i.e., HRP5s against trypsin-like proteases, cystatins against cysteine proteinases, and SLPI against serine proteinases) may also serve as surface defense by decreasing the proteolytic degradation of surface proteins and salivary defense proteins. Salivary chaperone Hsp70 is also a potent defense protein against cell surface damage; moreover, Hsp70 can repair aggregated and/or denatured salivary proteins (2). Peroxidases also protect host cells that transform H$_2$O$_2$ (produced by microorganisms and during immunoinflammatory reactions) to reactive anion hypohydroxocyanate (OSCN$^-$) that has a stronger antibacterial effect but a smaller cell-damaging effect than H$_2$O$_2$ (16).

The diluting effect of saliva and the oral clearance (i.e., swallowing toward the stomach and/or expectoration) of many proteins, bound or free molecules, and microbes also serve as an effective surface defense mechanism.

Antimicrobial effects of saliva

A network of antimicrobial salivary defense includes numerous salivary proteins. Although some defense molecules are present in a rather low concentration in whole saliva, it should be considered that local concentrations of these proteins nearby the mucosal surfaces (mucosal transudate), periodontal sulcus (gingival crevicular fluid), and oral wounds and ulcers (transudate) may be much greater (2). Furthermore, the effects are additive, synergistic, and in many cases reinforced by immune and/or inflammatory reactions (2, 16, 17).

Some defense proteins are involved primarily in immune activation. Salivary immunoglobulins take part in elimination of bacteria fungi and viruses through specific immune binding and agglutination. Immunoglobulins act via the antibody-induced ozone formation (18). Molecular chaperone Hsp70 acts as a danger signal that leads to a specific immune answer and complement activation and takes part in the antigen presentation (bacterial, micolitic, and viral) (2). Cystatin C has chemonatotic properties, and it plays a role in antigen presentation of dendritic cells present in oral mucosa. Moreover, cystatin S, C, and D show antiviral activity; cystatin C, SA, and SN show antiapoptotic activity; and cystatin S shows antibacterial activity.

Other proteins are responsible for nonimmune elimination of microbes (4, 14, 17). Salivary amylase is proposed to perform inhibitory effect on growth of microorganism. Calprotectin has bactericidal and fungicidal properties. HRP5s show specific antibacterial and antifungal activities, i. actiniferine has bacteriostatic effect. Lysozyme is bactericidal for gram-positive bacteria. Secretory leukocyte proteinase inhibitor shows antibacterial, antifungal and antiviral activity. Defensins possess antimicrobial and cyotoxic properties. Mucins, especially MUC7, are highly affine to microorganisms, entrap and agglutinate bacteria, fungi and viral particles. Peroxidases have antifungal
and bacteriostatic properties through producing reactive anion hypothiocyanate. Acidic PRPs bind bacteria, basic PRPs bind fungi (e.g., Candida albicans), and viruses, whereas glycosylated PRPs bind bacteria and viruses that indicate a role of PRPs in clearance of these microorganisms toward the stomach (4, 16, 17). (Data are summarized in Table 3.)

### Table 3 The Most Important Salivary Proteins with Antimicrobial Properties

<table>
<thead>
<tr>
<th>Name</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunoglobulin</td>
<td>Specific immune binding, agglutination, antibody-induced ozone formation</td>
</tr>
<tr>
<td>Salivary chaperon Hsp70</td>
<td>Danger signal, complement activation, antigen presentation</td>
</tr>
<tr>
<td>Cystatin C</td>
<td>Chemotactic, antigen presentation</td>
</tr>
<tr>
<td>Alpha-amylase</td>
<td>Inhibition of microbial growth</td>
</tr>
<tr>
<td>Calprotectine</td>
<td>Bactericidal, fungicidal</td>
</tr>
<tr>
<td>Histatins</td>
<td>Antimicrobial, antifungal, antiviral</td>
</tr>
<tr>
<td>Lactoferrine</td>
<td>Bacteriostatic</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>Bacteriolytic</td>
</tr>
<tr>
<td>Protease inhibitor SLPI</td>
<td>Antibacterial, antifungal, antiviral</td>
</tr>
<tr>
<td>Defensins</td>
<td>Antimicrobial, cytotoxic</td>
</tr>
<tr>
<td>Mucins</td>
<td>Agglutination</td>
</tr>
<tr>
<td>Peroxidases</td>
<td>Bacteriostatic, antifungal, antiviral</td>
</tr>
<tr>
<td>Proline-rich proteins</td>
<td>A specific binding of bacteria, fungi, and viruses</td>
</tr>
</tbody>
</table>

### Saliva and wound healing

Besides prevention of wound infactions through the above antimicrobial effects, saliva plays other roles in the healing of oral wounds as well. Salivary EGF speeds up the healing process by its angiogenic and cell proliferating effects (19, 20). Other growth factors present in saliva (3) such as transforming growth factor beta, fibroblast growth factor, insulin-like growth factors, and nerve growth factor also contribute to the healing process. Furthermore, saliva contains several blood clotting factors (IXa, VIII, XI) at a level comparable to plasma, and saliva can replace platelets in the thrombin generation (5). This property of saliva is highly important in the oral wound healing because, although saliva dilutes blood-clotting factors of blood origin, blood-clotting can be initiated. A relatively high amount of salivary kallikrein (5) is suggested to play a role in vasodilatation around mucosal injuries to facilitate healing and defense of the injured area.

### Saliva in dental caries

Besides the presumable alterations of bacterial adhesion, plaque formation, and salivary defense mechanisms detailed previously, some more or less-specifically detectable changes of saliva are in connection with caries formation, and it may be used for recognizing risk patients and to maintain prevention. Decreased saliva flow rate, decreased buffer capacity, increased number of S. mutans and Lactobacilli in saliva are usually associated with increased caries prevalence. Similarly, decreased level of certain salivary proteins such as proline-rich proteins (PRP1, PRP3), histatin 1, and statherin is associated with significantly higher caries susceptibility (3, 21).

### Saliva and periodontal disorders

Several important effects of saliva on bacterial adhesion, plaque and calculus formation, and elimination of microorganisms are described previously. Besides these effects and measuring several marker proteins, saliva may be used for periodontal disorder screening. The levels of proteolytic granulocyte enzyme elastase, protease inhibitor alpha1-antitrypsin, and elastase inhibitor alpha-macroglobulin may increase considerably under gingivitis and/or periodontitis. Moreover, the level of alpha-macroglobulin is also a good indicator of an individual's periodontal status (22). Salivary level of 3-hydroxy-fatty acids (lipid constituent of lipopolysacharide endotoxin of several anaerobic bacterial are also good indicators of chronic periodontitis (23). Albumin may also correlate with gingival inflammation (24). Although the periodontal diseases be diagnosed only by dental examination, the latest data on the role of periodontal diseases in cardiovascular and cerebrovascular conditions and also in premature birth may increase the significance of “quick screening” from the saliva (3).

### Saliva and xerostomia

Five main reasons for subjective dry mouth sensation (xerostomia) exist, such as salivary gland disorders, systemic disorders, medication, radiation therapy, and aging. In healthy humans, the resting flow rate is around (or somewhat below) 1 mL/min, although in some conditions, like dehydration, sleeping, relaxation, or altered mental states (i.e., hypnosis, photo-acoustic stimulation) the flow rate may not exceed 0.25 mL/min (25). Under stimulation, the flow rates increase in healthy subjects to the usual value of 1.5–2.3 mL/min, but it may increase to 3.7 mL/min (26). Xerostomia usually appears when resting unstimulated whole saliva flow rate is less than 0.4–0.7 mL/min. In other cases (~25% of patients), the resting flow rate decreases, but the stimulated flow remains normal. In other patients (~22%), both resting and stimulated flow rate is normal (27). In serious cases, saliva demonstrates low pH and buffer capacity, increased total protein albumin and sodium concentration, decreased amylase protein ratio, and high lactobacilli and yeast.
activity and abnormally elevated salivary prostaglandin E2 was
of electrolytes (sodium, chloride, calcium, and phosphorus) and
in the saliva was also found. Cystic fibrosis induces elevation
of phosphate (but not the cystatin C). A degradation of the glycan moiety
protein concentration and decrease the cystatin S concentration
and carbamazepine medications increase the amylase and total
flow rate and pH usually remains normal. Phentoin, valproate,
and carbamazepine medications increase the amylase and total
protein concentration and decrease the cystatin S concentration
(but not the cystatin C). A degradation of the glycan moiety
of salivary mucins and other glycoproteins, and a decreased ef-
ficacy of mucine-induced aggregation of bacteria was also
shown under multiple antiepileptic medications (24). However,
it should also be considered that degradation and decreased bac-
terial aggregation was probably caused by poor oral hygiene in
this group.

Salivary changes under medication
Numerous kinds of medications such as anticholinergics, anti-
pressants, antidepressants, diuretics, benzodiazepines, anihyper-
tensive agents, muscle relaxants, analagetics, and antihistamines
have been reported to induce xerostomia, although some (an-
tipsychotics, benzodiazepines, and antihypertensive drugs) may
also induce sialorrhea (28). Antiepileptic drugs also induce sig-
ificant changes in saliva; however, in this case the level of
flow rate and pH usually remains normal. Phenotoin, valproate,
and carbamazepine medications increase the amylase and total
protein concentration and decrease the cystatin S concentration
(but not the cystatin C). A degradation of the glycan moiety
of salivary mucins and other glycoproteins, and a decreased ef-
ficacy of mucine-induced aggregation of bacteria was also
shown under multiple antiepileptic medications (24). However,
it should also be considered that degradation and decreased bac-
terial aggregation was probably caused by poor oral hygiene in
this group.

Salivary changes in systemic conditions
Anxiety and depression may lead to decrease in salivary flow
rate and consequent xerostomia. A cute stress conditions also
induce significant salivary changes such as a decrease in se-
cretory IgA (29), increase in salivary amylase (25, 30) and
molecular chaperone Hsp70 (25) concentrations, and prompt
changes of bacterial adherence to salivary mucins. In Sjögren’s
syndrome, low level of resting and stimulated flow rate as well
as increased salivary level of sodium, chloride, IgA, IgG, lacto-
ferrin, albumin, P2 microglobulin, lactoferrin, cystatine C, cystatine
S, prostaglandin E2, interleukin-6, soluble interleukin-2 recep-
tor, and kallikrein was reported. A decreased level of phosphate
in the saliva was also found. Cystic fibrosis induces elevation
of electrolytes (sodium, chloride, calcium, and phosphorus) and
lipid levels of submandibular saliva that lead to increased cal-
culus formation. An unusual form of EGF with poor biologic
activity and abnormally elevated salivary prostaglandin E2 was
also found in cystic fibrosis patients (3). Graft-versus-host dis-
 ease causes a mean reduction of 55-90% of salivary flow rate
with elevated concentration of sodium, magnesium, total pro-
tein, albumin, EGF, and IgG in the saliva, whereas the amount of
salivary IgA and IgM decreases, and no change exists in
potassium, calcium, and phosphate concentrations (31). (Data
are summarized in Table 4).

<table>
<thead>
<tr>
<th>Condition</th>
<th>Usual but not specific changes in the saliva</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>Decrease of flow rate</td>
</tr>
<tr>
<td>Depression</td>
<td>Decrease of flow rate</td>
</tr>
<tr>
<td>Acute stress</td>
<td>Decrease of IgA</td>
</tr>
<tr>
<td>Sjögren’s syndrome</td>
<td>Prompt changes of mucins’ adhesive properties</td>
</tr>
<tr>
<td>Cystic fibrosis</td>
<td>Increase of flow rate, phosphate</td>
</tr>
<tr>
<td>Graft-versus-host disease</td>
<td>Decrease of flow rate, sIgA, and IgM,</td>
</tr>
</tbody>
</table>

Salivary diagnostics of systemic diseases
In some cases, saliva can be used as a highly effective diag-
nostic tool of systemic conditions. Serum-free hormone levels
in the case of several nonpeptide hormones like cortisol, testos-
terone, estradiol, estradiol, progesterone, aldosterone, androstene-
don, dihydroandostendion, and insulin can be calculated from
salivary hormone levels (3, 32). Saliva levels of small peptide
type neurotransmitters such as met-encephalin, substance-p and
bet-endorphin (33), and metionin (a single amino acid derivative)
may also refer to plasma levels. Salivary levels of epinephrine,
norepinephrine, and dopamine although do not correlate too
much with serum levels, but increase specifically under stress
conditions. Monitoring of the systemic level of several medi-
cations is also possible from saliva. Similarly, medication and
drug abuse and level of active or passive smoking and ethanol
consumption can also be monitored. Screening of virus infec-
tions with detection of specific antibodies against viruses (i.e.,
hepatitis, HIV) in mucosal transudate enriched saliva is also a
simple, well-tolerated, and accurate method (3). Tumor markers
(i.e., c-erbB-e, p53 antigen, CA125) present in the saliva may
be also used for screening and early diagnosis of malignancies
that appear in several regions of the body (i.e., not exclusively
oral tumors) (3). The fact that, in addition to the normal salivary
transcriptome core (180 mRNA), a high amount (≥3,000) of
other mRNA is detectable under several systemic conditions,
and it will be of high diagnostic value in the future (32, 34).
Similarly, oral fluid also provides an available source of micro-
bial or human DNA, although the DNA content is rather low.
This finding is useful for biomarker profiling of oral bacteria,
oral, or systemic diseases and for forensic identification (32).
(Data are summarized in Table 5).
Saliva in Health and Disease: Chemical Biology of Saliva

**Table 5** The most important diagnostic possibilities of systemic diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Specific indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hormonal alterations</td>
<td>Nonpeptide hormones</td>
</tr>
<tr>
<td>Stress conditions</td>
<td>Cortisol, epinephrine, norpinephrine, dopamine</td>
</tr>
<tr>
<td>Abuse</td>
<td>Drug or derivatives</td>
</tr>
<tr>
<td>Infections</td>
<td>Virus-specific antibodies</td>
</tr>
<tr>
<td>Tumor (oral and other)</td>
<td>General and local tumor markers</td>
</tr>
<tr>
<td>Other disorders</td>
<td>Disorder-specific mRNA</td>
</tr>
</tbody>
</table>

**Chemical Techniques Used in Saliva Analysis**

The methodology of salivary analysis is wide ranging and includes nearly all techniques used commonly in other fields of chemical biology. The methods used most frequently for saliva analysis are summarized briefly below.

**Collection of saliva**

Sample collection should be made at standardized time, according to the diurnal cycle (and the response and recovery time) of the analyte. Subjects should not eat within 60 minutes prior to sample collection. For recovery of salivary glands, alcohol, caffeine, and dairy products should also be avoided. Resting saliva can be collected avoiding any chemical (i.e., acids), physical (i.e., pressure, warm, cold), biologic (i.e., taste, chewing), and psychologic (i.e., imagination of a meal) stimulation. Stimulated saliva is collected most widely with chewing stimulation (i.e., paraffin wax), and/or with taste stimulation (i.e., candy, lemon). Whole saliva can be collected simply by drooling into a vial with forward tilted heads or by allowing the saliva to accumulate in the mouth and then expectorate it into a vial. Isolated parotid saliva may be collected with direct cannulation of the parotid duct or with the use of parotid cap (a plastic container with a pocket that enables some negative pressure for stabilizing the device on the mucosal surface). Mixed saliva of submandibular and sublingual gland may be collected with direct cannulation of submandibular duct. Saliva of minor glands may be collected with pipettes or with small piece of absorbent. A fluid enriched in mucosal transudate and gingival crevicular fluid can be collected placing an absorbent pad between the cheek and the lower gum.

**Handling and storage of saliva**

Saliva is usually homogenized on a vortex mixer for one minute and precleared by centrifugation (i.e., 10,000 × g; 4°C; 10 min.) to remove food rests, bacteria, mucosal cells, and other particles. Saliva may also be cleared by filtration using small (i.e., 0.2 µm) pore size filters, but only small amounts, because the filter pores are blocked by high molecular weight components in saliva. Precleared saliva can be stored on ice (−4°C) without significant changes of enzyme activity (i.e., amylase) or protein degradation only for few hours. Addition of protease inhibitors is advantageous for time consuming analysis procedures. Freezing may lead to significant protein precipitation, even if quick freezing (i.e., liquid nitrogen) is used. Frozen sample can be stored at −20°C without any more damage for a few days only. Somewhat longer storage is possible in liquid nitrogen or at −80°C.

**Detection of ions**

The pH value can be measured with hydrogen-selective electrodes; however, it should be noted that the pH value of the saliva is dependent on the level of dissolved CO₂ thus–for a true pH value to be obtained—saliva must be collected without a loss of CO₂ to avoid measuring a pH value higher than real (5). The free (unbound) form of other ions like sodium and calcium can also be measured with ion selective electrodes. With the use of flame photometry the total amount of sodium, potassium and calcium can be measured in the saliva. With the use of atomic absorption spectrometry the total amount of calcium, magnesium, copper and some other ions (i.e., constituents of dentures like zinc, iron, cobalt, and chrome) is detectable. Photometric (calorimetric) methods can be used for measuring the total amount of chloride, calcium, bicarbonate, and phosphate. In case of calcium and phosphate a high percentage (~50% and up to 20%, respectively) may be bound to proteins and/or lipids (5, 35). Determination of the ratio between free and bound forms of calcium can be performed by comparing the data of flame photometry and electrode measurement. In case of phosphate, an exact discrimination is only possible after a careful isolation of proteins and lipids from saliva fluid.

**Analysis of salivary proteins**

The protein content of the saliva is determined usually with modified biuret reaction (Lowry method [36]). Qualitative protein analysis is carried out most frequently with gel electrophoresis, especially sodium-dodecyl-sulphate polyacrylamide gel electrophoresis, and isoelectric focusing in both cases either in one-dimensional (1-D) or in two dimensional (2-D) forms. Similarly, 1-D and 2-D high performance liquid chromatography (HPLC) is also used widely. These methods are combined frequently with immunologic methods such as Western-blot analysis (in the case of gels) and enzyme linked immunosay (ELISA, EIA) and radio immunosay in the case of whole saliva samples and saliva fractions (i.e., HPLC fractions). Measurements of enzymatic activity with the addition of substrates are common methods for determination of salivary amylase and lysozyme. Newer approaches use in-gel-tryptic-digestion of the separated proteins (7). Digestion is followed by extraction of the resulted peptides and fragmentation with tandem
mass spectrometry (MS/MS) in a predictable fashion that allows computational determination of the peptide sequence (37). The identification of the protein from which the peptide was derived can be reached by protein database searching. In the latest direct analysis of large protein complexes and multidimensional protein identification technologies, trypsin digestion is the first step, followed by chromatography of the peptide fragments (i.e., biphasic or triphasic microcapillary columns and HPLC) and MS/MS analysis (38).

### References

Saliva in Health and Disease, Chemical Biology of


Further Reading


See Also

Mass Spectrometry: Overview of Applications in Chemical Biology

Chromatography of Glycans

Chromatography of Lipids

Glycan-Protein Interactions

Lipoproteins, Chemistry of
Schizophrenia, Biological Mechanisms of
Nora I. Perrone-Bizzozero and W. Michael Bullock,
Department of Neurosciences, University of New Mexico School of Medicine.
doi: 10.1002/9780470048672.wecb669

Schizophrenia is a chronic and severely debilitating mental disorder that affects approximately 1% of the world’s population. Although schizophrenia has been recognized for over 100 years, the causes and pathophysiological mechanisms of this illness remained rather elusive until recently. Evidence obtained during the last 3 decades suggests that schizophrenia is a neurodevelopmental disorder that affects the structure and function of distributed brain regions. Multiple neurotransmitter systems have been implicated, as have both gray and white matter abnormalities. These structural alterations result in synaptic miscommunication at local neuronal circuits and long-distance functional disconnectivity, with both genetic and environmental factors contributing to these deficits. This article will discuss our current understanding of the biological and neurochemical bases of schizophrenia and will describe new pharmacological, genetic, and lesion models used for testing the mechanisms that underlie this devastating disease.

Schizophrenia was first identified in 1893 by the German psychiatrist Emil Kraepelin, who termed the disorder dementia praecox because of its early onset and irreversible mental decline. Then, 15 years later, the Swiss psychiatrist Eugen Bleuler recognized that the illness affected both the judgment and emotional state of the patient and referred to this disorder as schizophrenia, from the Greek schiz- to split and phren- mind. Bleuler identified the cardinal symptoms of the illness as loosening of associations, flat affect, social withdrawal, and ambivalence, and these criteria are still used for diagnostic purposes today. The onset of symptoms characteristically occurs during late adolescence and early adulthood. According to the Diagnostic and Statistical Manual IV (American Psychiatric Association, 1994), symptoms are categorized as positive or negative, and patients manifest these symptoms at various degrees during the course of their illness. Positive symptoms are manifestations of psychosis and include unusual behaviors such as paranoid or bizarre delusions, auditory hallucinations, and disorganized speech and thinking. In contrast, negative symptoms represent a loss of normal behaviors such as flat or blunted affect and emotion, poverty of speech (alogia), inability to experience pleasure (anhedonia), and lack of motivation (avolition). In addition, patients exhibit unremitting cognitive deficits. The clinical course and outcome of schizophrenia shows great variability; however, typically, positive symptoms fluctuate and negative symptoms remain more stable over time.

Alterations in Brain Structure and Function
Unlike neurodegenerative disorders such as Alzheimer’s disease and Parkinson’s disease, the brains of patients with schizophrenia reveal neither gross anatomical changes nor the presence of pathological structures such as amyloid plaques or Lewy bodies. Thus, for much of the 1900s, psychiatry textbooks classified schizophrenia as a “functional” psychosis, for example, a condition that had no underlying physical brain disease. The technological advances in the past 25 years made it possible for investigators to re-evaluate the biological bases of schizophrenia systematically and provided evidence of subtle but consistent neuropathological and molecular alterations. Although most of these studies focused on the dorsolateral prefrontal cortex (DLPFC) and hippocampus, other brain regions including the thalamus, cerebellum, and its connecting white matter tracts have been implicated in schizophrenia (Fig. 1).

Prefrontal cortex and anterior cingulate gyrus
Both postmortem tissue and neuroimaging analyses of the brain of patients with schizophrenia show small but reproducible gray matter reductions. In the prefrontal cortex (PFC), these changes were shown to be related to decreases in neuronal cell size and neuropil volume but not to a reduction in cell numbers (1). In
Hippocampus and entorhinal cortex

Convergent evidence from neuroimaging and neuropathological studies indicates that the structure and function of the hippocampus is compromised in schizophrenia. Patients with schizophrenia have decreased hippocampal volume (11, 12) and several cytoarchitectural abnormalities in both the hippocampus proper and entorhinal cortex (13). In addition, neuropathological examination of postmortem hippocampal tissue revealed a decreased number of GABAergic interneurons and several synaptic protein deficits in patients with schizophrenia (14–16). In contrast to the PFC, the levels of GAP-43 protein and the ratio of GAP-43 to synaptophysin mRNAs was decreased in the hippocampus of patients with schizophrenia (16, 17). These changes were correlated with decreased levels of another marker of excitatory synapses, the synaptic protein complexin II (18). Along with these structural and molecular alterations, patients with schizophrenia exhibit reduced activation of the hippocampus during the encoding and retrieval of episodic and relational information, two well-characterized hippocampal dependent tasks (19–21).

Cerebellum and thalamus

Traditionally, most research performed in the field of schizophrenia has focused on brain regions directly implicated in the symptomatology of the disease, namely the PFC and limbic areas. Work by Andreasen et al. (22) first implicated the cerebellum as an affected structure in schizophrenia through the cortico-cerebellar-thalamic-cortical circuit (CCTCC, see arrows in Fig. 1). Dysfunction in one area of this circuit such as the thalamus (23) is thought to affect all other areas of the circuit. As a component of the CCTCC, the lateral hemispheres of the cerebellum have been implicated in cognitive and emotional functioning (24). Retroviral tracing and neuroimaging studies demonstrated cerebellar–prefrontal connections (25), which may contribute to cognitive dysfunction in schizophrenia. Intrinsic to these connections, a forward modeling system of the cerebellum has been proposed in which information from the motor cortex or the PFC is transferred to the cerebellum, and the cerebellum acts as a predictor of the outcome for both motor and cognitive functioning (26). Clinically, patients exhibit cerebellar neurological signs (27), deficits in eyelid conditioning (28), and shortfalls in response timing (29). Neuroimaging studies have shown increases in blood flow and in glucose consumption in the cerebellum of schizophrenic patients relative to that of other brain regions (7, 30–32). In accordance with increased cerebellar activity, a recent study found that the levels of GAP-43 and brain-derived neurotrophic factor (BDNF), which are expressed in cerebellar granule cells in an activity-dependent manner, are upregulated in the patients (33). Additional molecular studies have shown decreased expression of the developmental marker reelin and the GABA synthesizing enzymes GAD65 and GAD67 and increases in the axonal chaperone protein semaphorin 3A (34–37). This finding is interesting considering the fact that GABA dysfunction in the cerebellum may lead to increases in granule cell firing and thus account for the increases seen in blood flow, glucose use, and
Schizophrenia, Biological Mechanisms of

Cortical Layers

Figure 2: Neurochemical and morphological alterations in the prefrontal cortex of patients with schizophrenia. The diagram shows some of the cell types in the layers I–VI of the dorsolateral prefrontal cortex, including glutamatergic pyramidal neurons (light gray) and GABAergic somatostatin-containing neurons, basket/wide arbor neurons and chandelier interneurons (dark gray), along with the changes in cell structure, gene expression, and neurotransmission observed in this region. These changes include decreased mRNA levels for several markers of GABAergic interneurons and reductions in the GAT-1 GABA transporter immunoreactivity in axon cartridges from chandelier cells with increased levels of postsynaptic GABAA receptors in pyramidal cells. Adapted from Ref. 9 and other references in the text.

White matter alterations

As described above and shown in Fig. 1, schizophrenia pathology is not restricted to a single brain region but affects multiple distributed areas. Brain regions considered important in the pathology of schizophrenia, such as the prefrontal cortex and temporal lobe, are linked normally to one another by tracts of dense and reciprocally afferent white matter, which thus suggests that white matter alterations could have a role in schizophrenia pathophysiology. Supporting this idea, white matter volume reductions have been reported in patients with schizophrenia (38–40). Analysis of frontal lobe white matter and corpus callosum of patients with diffusion tensor imaging revealed specific alterations in myelin structure (41). These changes were also observed in children and adolescent with schizophrenia (42), which indicates that white matter abnormalities may be present early in the disease. Postmortem tissue analysis at the light and electron microscopic levels also demonstrated the presence of myelin and oligodendrocyte abnormalities in schizophrenia (43–45). Along with these morphological abnormalities, microarray analysis of PFC tissue from subjects with schizophrenia also demonstrated downregulation of myelin-related and oligodendrocyte-related genes in the patients (46, 47). Finally, recent genetic studies identified polymorphisms in two important regulators of myelination, the transcription factor OLIG2 and the RNA-binding protein QKI, which are associated with increased risk for schizophrenia (48, 49).

Neurodevelopmental Hypothesis of Schizophrenia

As described above, the brains of patients with schizophrenia show alterations in the levels of the developmental markers reelin, GAP-43, and semaphorin 3A. In addition, evidence of disrupted neuronal distribution was found in several cortical areas and the hippocampal formation. These findings together with the absence of neuronal cell loss and concomitant reactive gliosis in these brain regions, both hallmarks of neurodegenerative disorders, led investigators to propose the neurodevelopmental hypothesis of schizophrenia (6, 50), which is still the prevailing theory in this disorder (51, 52). This hypothesis states that the illness is related to abnormal brain development and is supported by several pieces of evidence, including increased frequency of obstetric complications, viral infections, and other developmental stressors in patients with schizophrenia (see the sections below titled “Genetic and Environmental Factors” and “Developmental Models”) and the presence of soft neurological signs, cognitive impairment, and behavioral dysfunction in children long before the first psychotic episode.
Neurotransmitter Systems

As described in the previous sections, neuropathological studies demonstrated alterations in the levels of several synaptic proteins in the PFC, hippocampus, and cerebellum of patients with schizophrenia (13, 15, 53). These observations have led to the hypothesis that the clinical symptoms of schizophrenia are manifestations of abnormal neural circuitry and dysfunctional communication between different brain regions (22, 51). These abnormalities affect multiple neurotransmitter systems. Although dopamine dysfunction in schizophrenia is widely accepted, a growing body of evidence suggests the involvement of glutamate, GABA, and other neurotransmitters in schizophrenia.

Dopamine

The discovery that the first antipsychotic drugs in the early 1950s, such as chlorpromazine, work in vitro by blocking dopamine receptors led to the hypothesis that schizophrenia was the result of excessive dopamnergic neurotransmission (54, 55). Supporting this hypothesis, drugs that enhance dopamine action (e.g., cocaine, amphetamines, and L-DOPA) worsen the symptoms of schizophrenia. However, it is clear that 1) not all patients respond to neuroleptic treatment and 2) not all symptoms are reversed by the medication.

Classical treatments for schizophrenia involve the administration of typical antipsychotics, such as haloperidol and chlorpromazine, which primarily bind to dopamine D2 receptors with high affinity, and atypical antipsychotics, such as clozapine and risperidone, which bind to a broader range of receptors including serotonergic and noradrenergic receptors among others. Although antipsychotics are effective at relieving the positive symptoms via their actions on D2 receptors, they are not effective in ameliorating the negative and cognitive symptoms. Interestingly, recent studies suggest that patients with schizophrenia may have hypofunctional D1 dopamine receptors in the prefrontal cortex (Fig. 2) and that agonists to this subtype of receptor may be effective in treating the working memory deficits associated with this illness (56, 57). In addition to drugs that work on monoamine receptors, new drugs that target glutamate, GABA, and cholinergic receptors are now being tested for ameliorating the cognitive dysfunction in schizophrenia (58, 59).

Glutamate

The glutamate hypothesis of schizophrenia was derived from the fact that drugs that block the N-methyl-D-aspartate (NMDA) subtype of glutamate receptors, such as phencyclidine (PCP) and ketamine, cause schizophrenia-like symptoms in humans and animal models. Furthermore, these drugs mimic not only the positive (psychic) but also the negative and cognitive symptoms of the disease, which suggests that they act on the same basic pathophysiological mechanisms that are affected in schizophrenia.

NMDA receptor dysfunction has been characterized in different brain regions of patients with schizophrenia. Neuropathological studies revealed altered expression of receptor subunits in the prefrontal cortex (60), and single positron emission tomography studies have shown decreased NMDA receptor binding in the hippocampus of patients (61). Studies using NMDA receptor antagonists such as PCP, ketamine, and MK-801 additionally implicate hypofunction of these channels in schizophrenia (62). NMDA receptor antagonists have been shown to block NMDA channels located on GABAergic interneurons selectively (63, 64), which suggests that NMDA receptor dysfunction in a specific subset of these interneurons may be central to schizophrenia (62, 65).

GABA

GABA is the main inhibitory neurotransmitter in the brain, and dysfunction in certain subsets of GABAergic interneurons is one of the most consistent findings in the study of schizophrenia (9, 37, 66-68). Reductions in mRNA and protein levels of the 67 kD form of glutamic acid decarboxylase (GAD67), one of the GABA synthesizing enzymes, have been observed in the prefrontal cortex (69-71), the hippocampus (53, 72), the cerebellum (34, 35, 37), and other brain regions (73). As shown in Fig. 2, chandelier and basket interneurons in the PFC show decreased mRNA levels for the GAT-1 GABA transporter (74) and the calcium-binding protein parvalbumin (70), whereas other subtypes of interneurons have reduced levels of somatostatin (SST) mRNA (73). Posttranslational changes such as increases in GABA, u2 receptor density and GABA receptor radioligand binding in the PFC and anterior cingulate cortex were also observed (75-77). In addition to these findings, single nucleotide polymorphisms in the promoter region of the GAD67 gene were shown to be associated with reductions in gray matter in patients with childhood-onset schizophrenia (78). Considering the role of GABAergic interneurons in the modulation of excitatory output, it can be hypothesized that dysfunction in these cells may mediate some positive, negative, and cognitive symptoms seen in schizophrenia (79).

Other neurotransmitters: acetylcholine

Evidence of the involvement of acetylcholine in the pathophysiology of schizophrenia comes not only from the findings of decreased availability of cholinergic muscarinic receptors in patients (80, 81) but also from genetic studies that link specific polymorphisms in the gene for the alpha 7 nicotinic receptor (CHRNA7) with this illness. The CHRNA7 receptor is one of the ligand-gated ion channels that mediate fast cholinergic transmission at synapses. The CHRNA7 gene is located at chromosome 15q13-14, a locus implicated in the genetic transmission of schizophrenia (82, 83). Specific polymorphisms in CHRNA7 promoter region were shown to correlate with sensory gating alterations in patients with schizophrenia as measured by the P50 inhibition in auditory evoked response (84, 85). Furthermore, a recent study demonstrated that two additional single nucleotide polymorphisms (SNPs) in the CHRNA7 gene correlate with patterns of brain activation in schizophrenia patients during an auditory oddball task (86). The same study also linked an SNP in the gene coding for choline acetyltransferase, the acetylcholine synthesizing enzyme, with these abnormalities.
Genetic and Environmental Factors

Although the etiology of schizophrenia is not completely understood, it is becoming apparent that schizophrenia is a neurodevelopmental disorder that involves both genetic and environmental risk factors. The contribution of genetic factors was demonstrated by twin (87) and adoption studies (88) and by the higher prevalence of schizophrenia-like personality disorders in relatives of patients with schizophrenia (89). Other factors such as season of birth (90) and prenatal or perinatal complications such as ischemia (92) and viral infections (92) have also been identified as risk factors, although to a much lesser degree.

The influence of genetic factors is evidenced by the findings that about 50–75% of monozygotic twins with schizophrenia will have an affected twin and approximately 10% of first-degree relatives are also affected (93). Several genes have been associated with increased vulnerability for schizophrenia, including those encoding proteins associated with NMDA receptor function, synaptic plasticity, mitochondria energy metabolism, oxidative stress, development, and myelination (94, 95).

Recent studies demonstrate that specific polymorphisms in some of these genes correlate with cognitive and neuroimaging abnormalities in patients (96). The best example of these polymorphisms is an SNP coding for the substitution of a valine for a methionine in position 108 of the long form (Val108/158 Met in long form) of the catechol-O-methyltransferase (COMT, Val108/158 Met or SNP) protein. The amino acid substitution results in a protein that has increased stability and, thus, increased rate of dopamine inactivation (97). This polymorphism has been associated with impaired performance in working memory tests and abnormal patterns of prefrontal cortex activation in both patients with schizophrenia and healthy volunteers (98). In addition, polymorphisms in the genes for BDNF and the metabotropic glutamate receptor 3 (GRM3), among others, have been associated with subtle but consistent alterations in PFC and hippocampal structure and function (96).

Animal Models

Schizophrenia is a purely human disease, which makes it hard to model the behavioral manifestations of this illness in animals. Nonetheless, animal models have been shown to reproduce specific aspects of the illness such as its effects on brain structure and function. Currently, several animal models are available to investigators. These animal models can be classified as developmental, pharmacological, genetic, and lesion models. Examples of these animal models are described below, and a complete listing of current animal models can be found in a recent review (99).

Developmental models

Given the evidence of perinatal stressors in some patients with schizophrenia, animal models have been generated to examine the influence of these factors in adult behavior. These studies demonstrated that animals exposed to prenatal viral infections, maternal deprivation, and other stressors exhibit several behavioral abnormalities consistent with schizophrenia, including disrupted prepulse inhibition, enhanced response to amphetamine, and impaired social interactions (100–102). Furthermore, some models show molecular and morphological deficits in the neocortex and hippocampus that mimic the alterations seen in patients with schizophrenia (103).

Pharmacological models

Pharmacological models that exploit the GABA/glutamate system have proven useful in studying the underlying pathophysiology of schizophrenia. These models include the picrotoxin-induced antagonism of GABA receptors in rats (104) and the antagonism of NMDA receptors in both rodents and nonhuman primates (105–109). All these models affect the GABA/glutamate balance in different ways, but only the phenylethylamine model has shown both the GABAergic and NMDA receptor changes seen in patients with schizophrenia (108, 110, 111).

Because PCP administration leads to many symptoms inherent to schizophrenia, studies are now being conducted that administer the compound to rodents and primates to induce a schizophrenic-like phenotype (106, 112). A cute and chronic dosing regimen shows differential and often opposing effects in rodents. Immediately after administration of PCP to rats, neurons of the medial PFC show an initial excitation as seen by activation of early intermediate genes (112). This effect is likely because of the preferential blockade of receptors in GABAergic interneurons by PCP and other NMDA receptor antagonists (64). This initial activation then is followed by a period of cortical depression as described in glucose studies (114), presumably as a compensatory mechanism. A cute PCP administration also produces schizophrenia-like symptoms including social withdrawal (115), impaired sensory motor gating (116), and cognitive dysfunction (105, 107). Chronic intermittent exposure to low dose of PCP in rodents results in decreased metabolic activity in the prefrontal cortex, auditory cortex, hippocampus, and reticular nucleus of the thalamus (109), all regions affected in schizophrenia. Along with this decrease in metabolic function, decreases in parvalbumin expression were also seen (109), which mirror the chandelier and basket cell dysfunction seen in the prefrontal cortex of patients with schizophrenia (70). Taken together, the data suggest that the chronic intermittent PCP exposure model is one of the most functionally and neurochemically relevant animal models of unremitting schizophrenia.

Genetic models

Like the pharmacological model presented above, a genetic model also targets the NMDA receptor. This model was created by knocking down the NR1 subunit of the NMDA receptor, which is obligatory for receptor function, in mice so that only 5% of the protein is expressed (117). These animals, also known as NR1 hypomorphs, show NMDA receptor hypofunction and display several schizophrenia-like behaviors, such as reduced social interactions, increased locomotion, stereotypic movements, and sensorimotor gating deficits (117, 118). Interestingly, treatment of these mice with the atypical antipsychotic...
DISC1 protein exhibits brain and behavioral abnormalities suggestive of normal development (121), and mice that express mutant this chromosome 1q42.1 locus. This protein is known to be critical in schizophrenia (120), which has a balanced translocation in Scottish family with a high incidence of mental illness, including schizophrenia (120), which has a balanced translocation in this chromosome 1q42.1 locus. This protein is known to be critical for normal development (121), and mice that express mutant DISC1 protein exhibit brain and behavioral abnormalities suggestive of schizophrenia, such as impaired learning and memory processes and altered neuronal development (122-125).

Lesion models

Although no clear indication of a brain lesion is found in schizophrenia, developmental lesion models, such as the neonatal ventral hippocampal lesion (NVHL) and the neonatal amygdala data models, have been shown to reproduce several aspects of this illness (126, 127). For example, NVHL rats exhibit increased responses to dopamine agonists and MMDA receptor antagonists, which are manifested only after puberty. These animals also show impaired social interactions, altered sensorimotor gating, and cognitive deficits (128). At the molecular level, these animals show decreased numbers of GAD67 expressing interneurons in the medial PFC, which is similar to the findings observed in patients (71). Overall, this model also reproduces multiple aspects of schizophrenia behavior and pathophysiology.

Concluding Remarks

In summary, the work reviewed in the previous sections suggests that schizophrenia is a neurodevelopmental disorder that affects the structure and function of distributed brain regions and their connecting white matter, with both genetic and environmental factors contributing to these alterations. As shown in Fig. 1, affected regions include frontal lobe and limbic system structures involved in cognition and emotion and areas that participate in sensorimotor integration such as the thalamus and cerebellum. Structural and molecular abnormalities in these regions result in synaptic alterations at local neuronal circuits and long-distance functional disconnectivity. Besides dopamine, multiple neurotransmitter systems have been implicated including glutamate, GABA, and acetylcholine. Based on these findings, drugs targeting specific subtypes of these receptors are now being tested in animal models and patients. It is expected that these new developments will help researchers not only to understand the etiology and basic pathophysiological mechanisms that lead to schizophrenia but also to develop better treatment strategies for this devastating illness.

References

20. Weins AP, Schatter SL, Goff DC, Rauch SL, Aritz P, Fischman AJ, Hecker S. Impaired hippocampal recruitment during...


Schizophrenia, Biological Mechanisms of


Further Reading


See Also

Brain Development, Neurochemistry of Neurotransmitter Release Synaptic Chemistry


Lewis DA, Gonzalez-Burgos G. Neuroplasticity of neocortical circuits in schizophrenia. Neuropsychopharmacology. 2008;33:141–165.


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Systems Approach to Studying Disease

Gregory D. Foltz and Leroy Hood, Institute for Systems Biology, Seattle, Washington
doi: 10.1002/9780470048672.wecb587

The emerging field of systems biology promises to transform our understanding of the molecular basis of human disease. Recent technological advances allow for multiparameter measurements to be integrated across global genomic and proteomic platforms to inform predictive and probabilistic gene and protein regulatory networks. A systems approach to disease is based on the idea that disease-perturbed gene and protein regulatory networks differ from their normal counterparts, and that these differences are predictive of the disease course and response to therapy. The ability to detect disease-related perturbations in individual patients will transform health care over the next decade from our current reactive medicine to a new medical practice that is predictive, personalized, preventive, and participatory (P4 medicine).

The completion of the Human Genome Project fundamentally transformed contemporary approaches to medicine and disease. At the root of this transformation is the concept that biology is an informational science based on a digital code, encoded in the genome, from which all biologic processes are derived. This realization has several important implications for the study of disease. First, the digital code of the human genome is knowable and can be defined, interrogated, and compared in disease and healthy states. A second type of biologic information is that which emerges from the environment to modify the digital genomic readout. Thus, the integration of the digital genomic information and the environmental information across the development of organisms, their physiologic responses to the environment, and their responses to disease is the heart of what has come to be known as systems biology and systems medicine. This field provides a basic blueprint of disease from which hypotheses can be formulated regarding etiologies and possible therapeutic entry points. Second, the digital nature of genomic data, including the complex hierarchy of molecular (biologic) networks, and their dynamical response to environmental stimuli is amenable to modern computational and analytical techniques. Third, because of the parallel advancement of high-throughput genomic and proteomic technologies, the genomic code and its derivative dynamic molecular networks are increasingly accessible and testable at the individual level. Fourth, the digital code and the dynamics of its encoded networks can be compared across model organisms allowing for the targeted manipulation of gene networks implicated in disease states. A key point in the new medicine, then, is the idea that drugs can be designed to reengineer networks (and this will take multiple drugs) to make them behave in a more normal manner—quite a different concept from the idea that one drug should destroy or enhance the activity of one particular target. Finally, unbiased systems analyses result in our understanding of the emergent properties that are not predicted a priori, which leads to new insights into disease states. As a result of these influences, medicine is evolving from an observational, population-based, reactive approach to a more quantitative, predictive, and individualized approach to understanding disease and disease treatment based on rational genomic analysis. This evolution will ultimately transform traditional medicine into a new field of “systems medicine” that will provide personalized, predictive, preventative, and participatory treatment of disease in individual patients.

Systems Biology: An Introduction

Understanding a systems approach to disease first requires an introduction to the emerging field of systems biology. Systems approaches to understanding biologic complexity emerged as a result of several key transformative technological and theoretical advances that created new ways of thinking about biologic analysis and the scientific infrastructure required for integrated experimental models. These advances include 1) the development of high-throughput platforms for the rapid acquisition of global data sets. These platforms, which include high-speed DNA sequencers, DNA microarrays, mass spectrometry-based global proteomic technologies, antibody arrays, and the study of metabolomics with nuclear magnetic resonance (NMR) and mass spectrometry, have fundamentally enabled systems biology. 2) The completion of the Human Genome Project, which
A Systems Biology Approach to Disease

A systems approach to disease is derived from two very simple hypotheses. First, the functions of living organisms are executed by biologic networks of two different types: 1) protein networks (generally protein/protein interactions) that use biological information to carry out functions such as signal transduction, metabolism, development, or physiologic responses, and 2) gene regulatory networks (transcription factors controlling layered networks of other transcription factors) that take input biologic information from, for example, signal transduction networks—integrate and modulate it—and then output it to the protein networks mediating, for example, development of physiologic responses. Second, disease states and (b) to integrate these data to generate predictive mathematical models of disease behavior and response to therapy.

Systems medicine requires the integration of many different types of data into models that have predictive behavior. We discuss below several of the high-throughput platforms that are generating large-scale data for a systems approach to disease.

High-Throughput Platforms for Systems Analysis

Genomics

Complete genome sequencing provided the foundation for systems biology by enabling investigations that could examine specific molecular hypotheses in the context of a completely defined catalog of genes and proteins for humans. The first complete genome of a free-living species, Haemophilus influenzae, was completed in 1995 (1) and was followed quickly by other important human bacterial pathogens (2–4), yeast (5), human (6), mouse (7) and chimpanzee (8). By 2005, only a decade later, more than 1,000 complete genomes have been completed. This massive expansion of sequence data was enabled by
technological improvements in the Sanger sequencing method (9), which allowed for increased automation and throughput (10, 11). Despite these advances, significant infrastructure and cost is required for sequencing even a relatively small genome. As a result, most genome sequencing is still performed at large dedicated genome centers and most completed genomes represent only one or a few sampled organisms (12). Recent developments in microfluidics, image processing, and enzymology promise to increase massively the speed and capacity of DNA sequencing at a significantly reduced cost. Currently, this “next-generation” high-throughput DNA sequencing (from companies such as 454, Solexa, Applied Biosystems, and Helicos) is being used to characterize cancer-associated mutations across extended patient populations (13). Once individual genome sequencing is achieved economically, comprehensive identification of individualized markers of disease susceptibility and treatment response will enable predictive and preventative strategies to treat and prevent disease.

One powerful application that has emerged from large-scale genomic sequencing is the identification and characterization of polymorphisms, either single nucleotide polymorphisms (SNPs) or, more commonly, simple sequence repeats, in genes that identify and predict variations in biologic response, behavior, and predisposition to disease states (by DNA arrays and hybridization or by DNA sequencing of various types). Most common diseases are thought to result from a mixture of genetic and environmental factors. Many factors demonstrate a complex genetic predisposition thought to result from the contribution of small variations in several genes. More than four million putative SNPs have been identified in the human genome. Because they are highly abundant, occurring on average every few hundred base pairs in the genome, and, relatively stable, they provide useful markers for linkage analysis of genes involved in the pathogenesis of complex disease. The development of high-throughput genotyping methods makes genome-wide linkage disequilibrium mapping of SNPs a viable approach to the study of complex disease susceptibility. A nontrivial application of the SNP analysis is the identification of genetic variants that influence a patient’s response to a drug, which usually is used to predict pharmacological efficacy or the likelihood of harmful side effects. An important extension of this application is the identification of SNPs that alter cellular responses to biologic signaling during normal development and function. This largely unexplored area offers vast potential for defining the genetic basis of normal biologic variation in the development of disease.

Transcriptomics

A advance in genome sequencing technology enabled the systematic measurement and comparative analysis of complete transcriptional programs of cells and tissues at various points in time and at any given developmental, pathological, or functional stage. The most widely used methodologies for transcriptome analysis include DNA microarrays, serial analysis of gene expression (SAGE), microarray-based massively parallel signature sequencing (MPSS), and the massively parallel sequencing by synthesis (SBS). DNA microarrays are a powerful tool for high-throughput identification and quantitation of nucleic acids in biologic systems. DNA arrays typically consist of thousands of short gene-specific DNA molecules spatially arranged on a solid surface. Nucleic acid-specific hybridization allows for the precise identification and quantitation of transcript levels in a cellular system at two or more different states. DNA microarrays can either be spotted arrays of oligonucleotides (25-60 bp in length) or cDNA molecules, or they can be oligonucleotide arrays produced by piezoelectric deposition or in situ synthesis. Recently, custom-designed arrays have become available containing up to 400,000 oligonucleotides. This flexibility in array design allows for economical approaches to study specific disease states in systems biology.

In principle, oligonucleotide arrays are more specific than the cDNA array and have the capability to distinguish between single-nucleotide differences. This method has the advantage of distinguishing between transcripts derived from individual members of multigene families and alternatively spliced variants. The widespread acceptance of DNA array technology has led to the development of applications beyond transcriptome analysis that have an impact on the study of disease. DNA array technology has been used in genotyping studies to identify SNPs and to confirm the sequence identity of known regions of DNA. Currently, DNA array applications include promoter analysis, ChiP-on-chip studies of protein-promoter binding site occupancy, mutation analysis, comparative genomic hybridization, and genome resequencing. DNA sequences can also be tagged or labeled in such a way that they can be identified in solution. A powerful new application of sequencing technology, massively parallel signature sequencing (MPS or SBS), combines advances in microfluidics, enzymology, and image processing technologies to allow up to 1,000,000 different sequences of up to 35 residues to be determined simultaneously per sample (14).

Proteomics

Proteomics can be defined as the global characterization of proteins in complex mixtures, including protein identity, abundance, processing, chemical modifications, interactions in protein complexes, and subcellular localization within a cell or tissue. Currently, no proteomics technology approaches the throughput and level of automation of genomic technology. Strategies for protein identification and quantification can be divided into MS-based techniques, designed to provide unbiased global measurements of protein abundance, and antibody-based techniques designed to identify known proteins in a biologic sample. These and other strategies to reduce sample complexity, differentially label protein samples, and improve relative quantitation of proteins by MS analysis and antibody arrays promise to enhance a systems approach to disease.

Proteomics: mass spectrometry techniques

The standard approach to proteome analysis is based on the separation of complex protein samples by two-dimensional gel electrophoresis (2DE) and the subsequent identification of selected separated protein species by one of a variety of mass
specrometric techniques (15). This approach is limited funda-
mentally because specific classes of proteins are not rep-
resented in the gels (e.g., membrane proteins, small proteins, or
very basic proteins) or are undetectable because of their limited
abundance. Furthermore, this method remains labor intensive
despite automation of 2DE gel computerized pattern matching,
protein extraction and digestion, and mass spectrometry-based
analysis. In addition, the enormous dynamic range of protein
abundances found in biologic systems, ranging from 1 to 10^6
copies or greater in cells and up to 1 to 10^{12} in serum, is a
major impediment for detecting low-abundance proteins. Im-
proved throughput is provided by direct analysis using tandem
mass spectrometry (MS/MS) of peptides generated by the di-
gestion of complex, unseparated protein mixtures (16). The key
feature of this method is the ability of a tandem mass spec-
trometer to collect sequence information from a specific pep-
tide, even if numerous other peptides are concurrently present
in the sample. This collection is accomplished in the instru-
ment by the isolation of the peptide ion of interest from other
peptides, fragmentation of the peptide ion in a collision cell
collision-induced dissociation, CID), and the acquisition of the
fragment ion masses in a computer. It is these fragment ion
masses that represent unique identifiers for a peptide and the
sequence of the peptide, and therefore, the identity of a protein
determined by correlating the CID spectrum with the contents
of sequence databases (17). Recently, protein separation tech-
niques have been enhanced by the use of multidimensional liq-
uid chromatography (LC) followed by specific protein/peptide
capture strategies (18).

The development of technologies for global comparative
measurements of proteins or protein moner-metal complexes, the proteome, and their function and structure from cells or tissues of different
states (e.g., healthy vs. Disease) is a fundamental requirement
for a systems approach to disease. Stable isoform labeling of
proteins or proteins enables high-throughput relative quantification of proteins using MS on a scale approaching several thou-
sand per sample. The general strategy involves differentially
labeling proteins or proteolytic peptides with stable isotopes,
mixing of labeled samples at a 1:1 ratio, followed by com-
bined sample processing and subsequent MS analysis. As the labeling reagents possess almost identical chemical properties,
the labeled peptides appear closely paired in the LC and MS
processes. Relative quantification is achieved by comparing ion
signal intensities or peak areas of isotope-encoded peptide pairs
observed in the corresponding mass spectra. Current methods
in current for the isotope encoded tag for relative and absolute quantitation (iTRAQ).

The isotope-coded affinity tag (ICAT) technique involves dif-
ferential labeling of two different protein populations on the side
chain of reduced cysteinyI residues using one of two chem-
ically identical but isotopically different ICAT reagents (19).
By incorporating a biotin affinity tag into the ICAT reagents,
selective isolation and purification of labeled peptides substan-
tially reduces sample complexity. The ICAT approach has been
applied successfully to the systematic identification and quantifi-
cation of proteins contained in the microsomal fraction of cells
(20). It has also been applied, in conjunction with DNA mi-
croarray analysis, to identify differential expression profiles of
hematopoietic progenitor cells (21). A major drawback of ICAT
is that it only labels the fraction of proteins containing cysteine residues. An alternative approach, SILAC, involves
growing two populations of cells, under identical conditions ex-
cept that the culture medium for one population contains 20
essential amino acids in their naturally occurring isotopic forms
(“light” population), whereas the other population is grown
in medium where one or more amino acids are replaced by
heavy isotopic analogs (22). The incorporation of a heavy
amino acid into a peptide, which is referred to as metabolic la-
beling, results in a known mass shift relative to the peptide
that contains the light version of the amino acid. The advan-
tage of using metabolic labeling is that it allows mixing of
labeled and unlabeled cells before the fractionation and purifi-
cation steps and therefore avoids introduction of any errors in
relative quantification in subsequent sample preparation. Fur-
thermore, all peptides within the sample can be analyzed, not
just those containing cysteine residues, increasing confidence in
both identification and quantification.

Isobaric tag for relative and absolute quantitation (iTRAQ) is
a multiplexed approach that allows up to four samples to be ana-
yzed simultaneously by MS in the same experiment (23). Pe-
tides are labeled on the free amine groups at the amino terminus
and on lysine residues. Unlike other isotopic labeling strategies,
the iTRAQ label reagents are designed to provide quantitative
information during peptide fragmentation. This technique mod-
difies peptides by linking a mass balance group (carbonyl group) and
and a reporter group (based on N-methylpiperazine). Designed
to be isobaric (having the same mass), the iTRAQ reagents
are chromatomographically indistinguishable in the LC step, which
causes the ion peak for each of the identical labeled peptides
to be detected simultaneously by the mass spectrometer. When
MS/MS is used for analysis, the reporter group of the iTRAQ
labeled and unlabeled cells before the fractionation and purifi-
cation steps and therefore avoids introduction of any errors in
relative quantification in subsequent sample preparation. Fur-
thermore, all peptides within the sample can be analyzed, not
just those containing cysteine residues, increasing confidence in
both identification and quantification.

Proteonomic reducing complexity
Protein concentrations in biologic systems span 10^{12} orders of
magnitude, whereas the most common available mass spec-
trometry-based methods only allow for the identification of
proteins spanning approximately three orders of magnitude in
concentration from a given sample. Several methods have been
advanced that select for specified fractions of the proteome in
order to reduce the complexity of the sample sufficiently
to identify biologically interesting proteins. Protein glycosyla-
tion, one of the most common posttranslational modifications, is
characteristic of secreted proteins and cell-surface markers but
not found on the predominant serum proteins such as albumin.
Recent approaches selecting for N-linked or O-linked glycosylated peptides using affinity capture techniques or solid-phase extraction followed by stable isotope labeling enrich for these biologically active proteins. Zhang et al. examined glycosylated proteins from several tissues, cells, and plasma and compared the glycoproteins identified in the tissues and cells to those identified in the plasma. A significant overlapping was observed. This study demonstrates that tissue-derived proteins are indeed present and detectable in the plasma via direct MS analysis of captured glycopeptides, proving the feasibility of MS-based approach for plasma protein discovery and analysis. A significant improvement on this technique has been the capture of glycopeptides (rather than glycoproteins) (27).

Multiple reaction monitoring (MRM) is a highly selective, highly sensitive mass spectrometry approach for detecting the presence of particular peptide species in a complex mixture such as plasma (28). A specific tryptic peptide is selected as a stoichiometric representative of the protein from which it is cleaved. This peptide is quantified by MRM against a spiked internal standard (a synthetic stable isotope-labeled version of the peptide) to yield a measure of protein concentration. In principle, such an assay requires only knowledge of the masses of the selected peptide and its fragment ions and an ability to make the stable isotope-labeled version. This method can quantify reliably protein concentrations over a dynamic range of 4.5 orders of magnitude in human plasma using a multiplexed approach. MRM assays coupled with enrichment of proteins by immunodepletion and size exclusion chromatography (29), or enrichment of peptides by antibody capture have also been reported (30). Stable isotope standards and capture by anti-peptide antibodies (SISCA PA) has been shown to extend the sensitivity of a peptide assay by at least two orders of magnitude and with additional development appears capable of extending the MRM method to cover the full known dynamic range of plasma (i.e., to the pg/mL level) (30). In systems approaches to disease, many important bioactive proteins are presumed to be secreted in the blood as key regulators of systems processes. These and other strategies to reduce or overcome the complexity of serum hold great promise for identifying key proteins active in disease states.

Proteomics: antibody-based arrays

An alternative strategy to global proteome analysis using MS is the use of antibody-based array techniques (31-33). A variety of methods have been developed based on antibody binding, all of which are limited by 1) dependence on the affinity and specificity of the antibodies employed for detection, 2) relatively high cost of generating monoclonal antibodies, and 3) potential cross reactivities in complex protein mixtures. Despite these limitations, antibody arrays have the advantage of providing a quantitative and comparative platform for rapid screening of proteomes from different disease states such as lung (34), pancreatic (35), and prostate cancer (36). One emerging approach with tremendous promise is surface plasmon resonance (SPR), which enables real-time, label-free measurement of protein abundance (37). SPR is a physical phenomenon that occurs when electromagnetic waves, such as light, are reflected off a thin metal film at specific incident angles and wavelengths. A fraction of the light energy (either polychromatic, many colors, or monochromatic, one color) interacts with and transfers to the surface plasmons, thus reducing the reflected light intensity at a sharply defined angle or at a specific wavelength. Any modifications on the metal surface, such as occurs with the interaction between antibody and antigen, will affect the SPR condition and can be used to detect and monitor specific molecular interactions. Sensitivity of SPR for low-abundance proteins is estimated to be in the picogram/cenimeter squared range. Current SPR-based chips have 800 unique antibodies arrayed at approximately 4 µm spatial resolution. Significant advantages of this method are that 1) protein abundance can be monitored in real time allowing for assessment of binding dynamics, and 2) slides can be regenerated allowing for cost-effective screens of multiple samples. Another recently developed technique, DNA-encoded antibody libraries (DEALs), is a highly sensitive measurement technique that can detect protein and single-stranded DNA simultaneously on a single chip (38). DNA-encoded antibodies are labeled with single-stranded DNA oligomers. DNA-encoded antibodies and secondary (fluorescently) labeled antibodies are added to the biologic sample containing the protein of interest. The entire complex is then captured by nucleic-acid hybridization onto a spot that was pretreated with the complementary single-stranded DNA oligomer. This approach has been used for the rapid detection of multiple proteins within a single microfluidic channel with a lower detection limit of 10 fM, which is 150 times more sensitive than the analog ELISA.

Computational Approaches

A systems approach to disease requires the integration of vast amounts of quantitative biologic data generated by global genomic and proteomic analyses in order to 1) identify comprehensively key molecular components defining disease and healthy states, and 2) determine how these components interact in biologic networks in a predictable way. Initial efforts primarily integrated and analyzed large databases of gene expression data to identify subsets of genes with predictive value for disease stratification and prognosis. A variety of methods have been applied to disease diagnoses, including approaches based on support vector machines and relative expression reversals, among many others. Application of these methods has led to the discovery of molecular classifiers of varying degrees of accuracy to identify prognostic signatures for breast cancer, ovarian cancer, colon cancer, prostate cancer, and brain cancer (39, 40).

With the development of protein-protein and protein-DNA interaction databases, gene expression data can be mapped onto interaction networks to identify relevant biologic pathways active in specific disease states or experimental perturbations. This type of approach is useful for assigning disease-specific relevance to differentially expressed genes or molecular pathways and has been applied in several human diseases, most notably cancer. Although these interaction networks are very useful tools for visualizing large data sets, they are not computable, predictive network models, which are those that hold
the most promise for predictive medicine and drug development. One approach uses an integrated framework, Pointillist (41), for combining diverse data sets and inferential algorithms to generate model networks, which are incorporated into Cytoscape (42) for visualization and simulation. The integration methodology of Pointillist can handle data of different types and sizes (e.g., interactions, protein expression, or gene expression) to create a higher confidence interaction network than that resulting from a single data set alone. A novel aspect of this methodology is that it does not require a "gold standard" set of data to be used for training nor does it make assumptions about the underlying statistical distributions in the form of parametric models. This process involves designing an efficient deterministic optimization algorithm to minimize the numbers of misses and false positives via an iterative element by element procedure. The methodology is general purpose so that it can be applied to integrate data from any existing and future technologies (43).

Other approaches for inferring genetic regulatory networks include parsimonious linear regression models, probabilistic Boolean networks, and Bayesian networks from expression data (both steady-state and time-course) (44). Probabilistic Boolean networks are robust in the face of biological and measurement uncertainty and offer the ability to characterize and simulate global network dynamics using the inferred model structure (45). It also provides a natural way to determine the influences of particular genes on the global network behavior (46). Thus, the model can be used to predict the effects of perturbations on network dynamics, which is an important goal for understanding disease development and treatment response. These and other predictive models stemming from mathematical descriptions of biochemical reaction networks and statistical influence models are critical for identifying disease-perturbed networks in disease states. Dynamic and predictive network models have been developed for important signaling networks in disease such cancer. Such approaches are now used to predict response to network perturbations in mammalian systems using an algorithm called Reconstruction of Accurate Cellular Networks (ARACNe) (47). These and other computational modeling approaches will play a key role in the identification of disease-perturbed networks, identifying potential therapeutic targets. With the development of comprehensive databases, hypothesis-driven global analysis methods, and predictive network models, systems biology has matured as a predictive science, which is capable of generating testable hypotheses based on network models with predictable behaviors. Applying this systems approach to disease holds great promise for the development of new therapies.

P4 Medicine: Personalized, Predictive, Preventative, and Participatory

Health care will be transformed over the next decade from our current reactive medicine to a new medical practice that is predictive, personalized, preventive, and participatory (P4). Personalized medicine will reflect the fact that each human differs from one another by approximately 6 million DNA bases—and has unique predispositions to differing combinations of diseases. The genome analyses and blood molecular fingerprints will permit each individual to be assessed individually and to be treated individually. Preventive medicine will emerge from the realization that a systems approach to disease will bring deep insights into the disease-perturbed networks and the fact that drugs can be eventually used to reengineer network behavior. Thus, the strategy for choosing drug targets will change in a fundamental manner. In a similar vein—once an individual's DNA has predicted a high likelihood say for brain cancer at the age of 50 or older—one could design drugs to prevent the relevant brain networks from ever becoming perturbed—if the drugs are taken 10 or 20 years before onset—and this would be preventive medicine. Finally, participatory medicine originates from the fact that given more information patients will be able to participate more fully in choosing their own health trajectories. This participation will also require the education of physicians as to the nature of P4 medicine. This transformation, if focused and properly leveraged, can both immensely improve the lives of people and reduce substantially the growing burden of health-care costs in modern countries.

The transformation that is currently on the horizon, from our current reactive medicine to a medicine that is predictive, personalized, preventive, and participatory, will impact much more than medical science. It will affect national economies, social policy, and the spectrum of business relationships, opportunities, and constraints. The science that is driving this change has not yet impacted medicine, but the changes are on the horizon. The changes in the science have been building for the past 20 years, beginning with the human genome project. In coming decades the health-care system will not only generate billions of bits of data for each individual patient but must learn how to use this data effectively—for the individual and for the collective knowledge of medical susceptibilities and response to therapies that will be enabled. This transformation will be driven by new systems strategies for studying disease, powerful new measurement technologies (e.g., nanotechnology), and revolutionary new computational and mathematical tools for dealing with the enormous amounts of information that will be gathered and for converting it into hypotheses about health and disease.

References


22. Ong SE, et al. Stable isotope labeling by amino acids in cell

21. Tian Q, et al. Integrated genomic and proteomic analyses of


19. Gygi SP, et al. Quantitative analysis of complex protein mix-

18. Wolters DA, Washburn MP, Yates JR 3rd. An automated mul-

17. Sadygov RG, et al. Code developments to improve the efficiency


15. Patterson SD, Aebersold RH. Proteomics: the first decade and


13. Marqués M, et al. Genome sequencing in microfabricated high-

12. Hall N. Advanced sequencing technologies and their wider impact

11. Prober JM, et al. Integration of automated MS/MS spectra interpre-

10. Madabhushi RS. Separation of 4-color DNA sequencing extension

9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-


Further Reading


See Also

Functional Genomics
Informatics for Proteins
Proteomics
Systems Biology
ADME Properties of Drugs
Li Di and Edward H. Kerns, Wyeth Research, Princeton, New Jersey

doi: 10.1002/9780470048672.wecb003

ADME properties have tremendous impact on the success of drug candidates. They are a set of fundamental physico-chemical and biochemical properties of drug molecules that can be affected by physiologic conditions in vivo: solubility, permeability, stability, metabolism, and others. ADME properties can be improved through structural modification by medicinal chemists. They are important both in vitro and in vivo. A successful drug must possess a balance of good potency and drug-like properties. High-throughput ADME assays have been developed and implemented widely in the pharmaceutical industry to reduce the attrition of drug candidates. The data are used to identify potential liabilities, guide structural modification, prioritize lead series, select compounds for in vivo studies, and diagnose in vivo assay results. Studies have demonstrated that early screening of ADME properties is a successful approach to enhance the quality of drug candidates.

Drug discovery typically starts from an unmet medical need (1), followed by target identification-validation, HTS (high-throughput screening) and hit finding, as well as hit-to-lead and lead optimization (Fig. 3). Once an active compound is identified, preclinical development starts. If the compound is efficacious and safe in animals, it will be evaluated in human clinical trials after an IND (investigational new drug) is filed. Phase I clinical trials involve safety and pharmacokinetic studies in a small group of healthy volunteers. Phase II involves efficacy studies in a small group of patients. Phase III continues safety and efficacy studies in a large group of patients. If a compound is safe and efficacious in patients, an NDA (new drug application) is filed for FDA review and approval. Once an NDA is approved, the drug can be marketed. The safety of the drug continues to be monitored in postmarketing studies. Fifty discovery programs, each with hundreds to thousands of compounds prepared, will lead to 10 development candidates and result in one marketed drug (2). The process takes an average of 12 years, costs more than $800 million, and has a success rate of only 10% in the clinic (2, 3). Drug discovery and development is costly and high risk; it takes a long time. A study in the early 1990s showed that many reasons explain why drugs fail in development, such as toxicity, lack of efficacy, or market reasons (Table 1). It was striking to see that, at that time, more than 40% of the drugs failed development because of poor biopharmaceutical properties (3, 4). As a result, a major initiative in the pharmaceutical industry developed to address ADME/TOX (absorption, distribution, metabolism, excretion, and toxicity) issues during drug discovery so that property information can be generated in parallel with activity data and so that successful drugs can be developed with good potency and drug-like properties (Fig. 2) (5). In the same time period, other new technologies also were implemented in drug discovery, such as HTS, combinatorial chemistry, molecular biology, genomics, and proteomics (6, 7). They all have a tremendous impact on the way drug discovery is conducted today. The term ADME originally referred to in vivo processes that govern pharmacokinetics (PK). It has been expanded to include in vitro physico-chemical and biochemical properties that affect in vivo PK. This article discusses various in vitro ADME properties that influence the overall success of a drug candidate in vivo. For detailed discussion of PK and metabolism, please refer to the other articles in this encyclopedia and further reading list. It is important to point out that suitable ADME properties not only are essential for the success of in vivo studies but also critical for the success of in vitro assays (5).

Traditionally, ADME/TOX assays were not high-throughput (8). Each project team only had a few slots per month for ADME/TOX profiling. If a team were making a lot of compounds, they would have to rely solely on potency and selectivity to triage compounds for in vivo studies, and the compound selected potentially could have poor properties and not be worthwhile to pursue. Nowadays, ADME/TOX assays are high throughput. Depending on company philosophy and available

Table 1 Reasons for drugs to fail development

<table>
<thead>
<tr>
<th>Reason for failure</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toxicity</td>
<td>22%</td>
</tr>
<tr>
<td>Lack of efficacy</td>
<td>32%</td>
</tr>
<tr>
<td>Market reasons</td>
<td>6%</td>
</tr>
<tr>
<td>Poor biopharmaceutical property</td>
<td>41%</td>
</tr>
</tbody>
</table>

For detailed discussion of PK and metabolism, please refer to the other articles in this encyclopedia and further reading list. It is important to point out that suitable ADME properties not only are essential for the success of in vivo studies but also critical for the success of in vitro assays (5).
Solubility

Solubility impacts every aspect of drug discovery and development. In vivo, an orally administered solid drug has to disintegrate, dissolve, pass through the gastrointestinal membrane into the portal vein, go through the liver, and finally reach systemic circulation. Solubility, along with permeability and metabolism, affects oral absorption and oral bioavailability. For IV dosing, a compound needs to be completely soluble in a dosing vehicle. Thus, good solubility is essential for the success of an IV formulation. Low solubility can cause abnormal PK profiles, such as second peak phenomena, which is caused by precipitation and redissolution. Formulation development for insoluble compounds can be costly and time-consuming, and it can lead to delayed introduction of lifesaving drugs to the market.

Solubility not only affects in vivo assay results, but it also has tremendous impact on in vitro assays (10, 11). Compounds that have poor aqueous solubility can produce erratic assay results, erroneous SAR (Structure-Activity Relationship), discrepancy between assays, and artificially low potency because of the right shift of IC50 curves. They can cause low HTS hit rates, low confirmation of hits from bench-top assays, and underestimation of toxicity. For some compounds, the potency and properties just cannot be measured because of the extremely low concentration. Low solubility can cause a lot of frustration and loss of productivity in drug discovery.

High-throughput kinetic solubility

Several high-throughput solubility methods have been developed and are used widely in the pharmaceutical industry to provide solubility information for thousands of discovery compounds per year. The data are used to provide an early alert for potential solubility issues so that strategies can be developed to address them. High-throughput methods often are used to measure kinetic solubility. In kinetic solubility methods, solid material is not necessarily in equilibrium with the solution. Compounds are dissolved in DM SO, which is added to aqueous buffers. The concentration of the compound in solution is determined by direct UV measurement (13), turbidity (13), nephelometric (14), or flow cytometry (15). Kinetic solubility usually is higher than equilibrium solubility for three reasons: 1) a small amount of DMSO is present in the buffer (~1%), 2) pre-crystallization tends to be amorphous materials, which have a higher solubility than crystalline solid, and 3) equilibrium is not fully established, and the solution is supersaturated. Because compounds are dissolved in DM SO first, the samples do not have memory of the crystal form, so kinetic solubility is not sensitive to crystal form. Kinetic solubility provides an optimistic estimate of the equilibrium solubility and insights for discovery experiments in which compound is first dissolved in DM SO.
Equilibrium solubility

Equilibrium solubility typically is used in late-stage drug discovery and development, when crystal forms of the material are well characterized. However, a recent trend exists to move the assay much earlier in drug discovery to screen compounds in dosing vehicles and to select compounds for in vivo studies. Equilibrium solubility typically is measured by adding a solvent to solid material and shaking for 24 hours. When the solution reaches equilibrium, it is filtered; the filtrate is diluted according to the sample size and analyzed using LC-UV (liquid chromatography-ultraviolet) or LC-MS (liquid chromatography-mass spectrometry). Equilibrium solubility also is used for formulation development and regulatory filing.

Solubility in bioassay medium

Solubility in bioassay medium usually is initiated when a team suspects that bioassay results might be incorrect because of low solubility. The assay is performed by using the exact protocol used in the bioassay, including buffer composition, dilution protocol, incubation time, and temperature. The solution is filtered, and the filtrate is diluted accordingly and quantified using LC-MS. The solubility data are used to guide assay optimization and address questions related to bioassays, such as erratic assay results, erroneous SAR, irreproducible data, and discrepancy between assays (11).

Permeability

Drugs have to go through many biologic barriers before they can reach the therapeutic targets. The barriers can be the gastrointestinal membrane for oral absorption, the blood-brain barrier for brain penetration, and the cell membrane for cell-based assays. Many transport mechanisms exist for absorption (16). Most drugs are absorbed by passive diffusion across the lipid membrane. Certain compounds can be absorbed by active transport. This transport requires the compounds to have certain structural motifs that can be recognized by transporters, such as amino acids, peptides, nucleosides, and so forth. Polar compounds with a small molecular weight (<200 Dalton) can be absorbed by the paracellular route through the junctions between the cells. Endocytosis can enhance uptake of highly lipophilic compounds by forming vesicles that engulf the molecules. Efflux occurs when efflux transporters, such as Pgp, pump the molecules out of cells. Pgp efflux can impair brain penetration. It also affects oral absorption, especially for highly potent compounds, when the low concentration in the gastrointestinal tract is not likely to saturate the efflux transporters.

Strategies to improve permeability include increasing lipophilicity with an optimal Log D value between 1 and 3, reducing hydrogen-bonding capacity, and decreasing polarity and rotatable bonds (5, 8). Prodrugs also are used to enhance the permeability of compounds.

Several software programs are available to predict permeability and absorption. The simple descriptors of Log P and PSA (polar surface area) gave good predictability for oral absorption (17). Several assays described below commonly are used in the pharmaceutical industry to determine the permeability of drug candidates.

PAMPA

PAMPA stands for parallel artificial membrane permeability assay. The assay was developed in 1998 (18). It has been used widely in the industry to measure permeability (13). The assay uses a 96-well filter plate coated with lipids (e.g., 2% phosphatidylinositol in dodecane) as a permeability barrier between two aqueous wells. The donor well contains the test compound in solution, and the acceptor well contains only buffer at the beginning of the assay. Molecules cross the lipid barrier from the donor to the acceptor well, driven by the concentration gradient. After incubation for a period of time, the concentrations in the donor and acceptor wells are determined and the effective permeability is calculated. PAMPA is a single-mechanism assay. It accounts for passive diffusion only. Different variations of the original PAMPA assay have been developed, including a different lipid composition, different filter material and thickness, additives in the donor and acceptor to create "sink" conditions (19), and addition of cosolvent to solubilize compounds that are otherwise insoluble. Specific PAMPA models have been optimized to predict oral absorption [Double Sink PAMPA (19), BAMPA (20), brain penetration (PAMPA-BBB (21)], and skin permeability [PAMPA-Skin (22)]. PAMPA has good correlation with Caco-2 (human colonic adenocarcinoma cell(s) and multi-drug resistant-Madin Darby Canine Kidney) for compounds that predominantly permeate by passive diffusion (23). PAMPA has much higher throughput and lower cost than Caco-2 and MDCK assay. PAMPA is applied widely in drug discovery to predict permeability. Computational Quantitative Structure-Activity Relationship models also have been developed for variations of PAMPA (24).

Caco-2

Caco-2 is a monolayer cell permeability assay. It was designed originally as an oral absorption model, and now it also is used commonly to investigate transporters. Bidirectional transport can be determined from apical to basolateral direction (A→B) and from basolateral to apical direction (B→A). Efflux ratio is the permeability from B to A direction divided by permeability from A to B direction. A high efflux ratio (e.g., > 2.5) is indicative of an efflux substrate. Inhibitors (e.g., verapamil for Pgp) can be added to confirm whether a compound is a substrate for a specific efflux transporter. The system commonly is used to diagnose efflux. Caco-2 not only expresses efflux transporters (Pgp, MRP) but also uptake transporters, such as PEPT1. It is used to screen compounds for substrates of PEPT1 to increase absorption via transporters. Caco-2 is a multiple mechanism assay including passive diffusion, paracellular route, active transport, as well as efflux and metabolism. To increase throughput, a 96-well Caco-2 assay has been developed (25). The Caco-2 assay involves a 21-day cell culture, which takes significant resources. An automated cell seeding and feeding workstation can be used to increase efficiency.
MDCK

MDCK is a monolayer cell transport assay. Similar to Caco-2, bidirectional transport can be determined. MDRI-MDCKII cells are transfected with the human MDRI gene to express human Pgp. The assay is a very sensitive Pgp assay and is used widely in the industry to identify Pgp substrates, especially for CNS (central nervous system) targets (26). MDCK only requires 3 days of cell culture, which significantly reduces resources needed for the assay.

Stability

Drugs encounter many stability challenges after oral dosing. In the gastrointestinal tract, compounds can degrade in different pH environments: acidic pH in the stomach and neutral to basic pHs in the small and large intestine. They can be metabolized in the intestine and liver by various metabolic enzymes. In blood, compounds can be hydrolyzed by hydrolytic enzymes. Physico-chemical and metabolic stability is essential for drug candidates to reach the therapeutic target with an efficacious concentration.

Metabolic stability

Metabolism occurs in two phases. Phase I metabolism oxidizes or reduces the drug molecule or exposes the functional groups to increase water solubility. It involves oxidation, reduction, hydrolysis, hydration, isomerization, and others. CYP450 is a major metabolic enzyme family. Figure 3 shows an example of Phase I metabolism. Phase II metabolism involves synthesis and conjugation reactions, such as glucuronidation, sulfation, methylation, and others. Phase II metabolism increases the water solubility of the compound or detoxifies a reactive species. Examples of Phase II metabolism are shown in Fig. 4.

Metabolism affects clearance and, therefore, half-life and oral bioavailability. It dictates how often and how much (dose regimen) a drug should be dosed (5, 8, 27). In vitro metabolic stability assays can be used to predict in vivo clearance and oral bioavailability, guide structural modification to enhance stability, develop structure–metabolism relationships, and triage compounds for in vivo studies. Strategies to improve Phase I metabolic stability are as follows: 1) Block the labile sites, 2) remove the labile groups, 3) reduce lipophilicity, and 4) introduce polar functional groups. For Phase II metabolism, several approaches can be applied to reduce metabolism: 1) introduce electron-withdrawing groups and steric hindrance, 2) employ isosteric replacement of hydroxyl and carboxylic acid groups, and 3) use prodrugs.

The microsomal stability assay is the most common platform for assessing metabolic liability of a compound. High-throughput assays have been developed to screen large numbers of compounds. Several strategies have been developed to increase throughput and shorten data turnaround time, including the single-time point approach, sample pooling, online extraction, parallel analysis of multiple samples, and the use of a combined ESI–APCI ion source (ESCI) to increase sample coverage. A assay conditions are critical for the success of a microsomal stability assay (28). Substrate concentration cannot be too high (<1 μM), otherwise it can saturate the enzymes. The concentration of DM SO should be kept less than or equal to 0.2%. High DM SO concentration can inhibit the CYP450 enzyme. Microsomal protein concentration needs to be low (0.5 mg/mL) because nonspecific binding can affect assay results. Microsomal batches can vary greatly from batch to batch and from vendor to vendor; therefore, quality control is essential to generate consistent data. Insoluble compounds can produce artificially high stability estimates (29) because of precipitation and nonspecific binding to plasticware. Cosolvent can be used to overcome solubility effects.

Hepatocytes also are used for metabolic studies (30). Other metabolic enzymes are present in hepatocytes, which are not found in microsomes. It was found that hepatocytes give better predictions for Phase II glucuronidation than microsomes, even though both systems underestimated the extent of glucuronidation in vivo (31). Molecules have to permeate hepatocyte cell membrane to be metabolized, which is a closer mimic of in vivo. Hepatocytes are used commonly for studies of metabolic enzyme induction.

Plasma stability

Plasma stability is important for compounds that contain hydrolyzable functional groups, such as esters, amides, carboxylic acids, sulfonamides, lactams, and lactones. It also is used commonly to screen prodrugs and to select antedrugs. Plasma stability is important for compounds that contain hydrolyzable functional groups. They can be hydrolyzed by hydrolytic enzymes. Plasma stability affects clearance and, therefore, half-life and oral bioavailability. It dictates how often and how much (dose regimen) a drug should be dosed (5, 8, 27). In vitro plasma stability assays can be used to predict in vivo clearance and oral bioavailability, guide structural modification to enhance stability, develop structure–metabolism relationships, and triage compounds for in vivo studies. Strategies to improve plasma stability are as follows: 1) Block the labile sites, 2) remove the labile groups, 3) reduce lipophilicity, and 4) introduce polar functional groups. For plasma stability, several approaches can be applied to reduce metabolism: 1) introduce electron-withdrawing groups and steric hindrance, 2) employ isosteric replacement of hydroxyl and carboxylic acid groups, and 3) use prodrugs.

The microsomal stability assay is the most common platform for assessing metabolic liability of a compound. High-throughput assays have been developed to screen large numbers of compounds. Several strategies have been developed to increase throughput and shorten data turnaround time, including the single-time point approach, sample pooling, online extraction, parallel analysis of multiple samples, and the use of a combined ESI–APCI ion source (ESCI) to increase sample coverage. A assay conditions are critical for the success of a microsomal stability assay (28). Substrate concentration cannot be too high (<1 μM), otherwise it can saturate the enzymes. The concentration of DM SO should be kept less than or equal to 0.2%. High DM SO concentration can inhibit the CYP450 enzyme. Microsomal protein concentration needs to be low (0.5 mg/mL) because nonspecific binding can affect assay results. Microsomal batches can vary greatly from batch to batch and from vendor to vendor; therefore, quality control is essential to generate consistent data. Insoluble compounds can produce artificially high stability estimates (29) because of precipitation and nonspecific binding to plasticware. Cosolvent can be used to overcome solubility effects.

Hepatocytes also are used for metabolic studies (30). Other metabolic enzymes are present in hepatocytes, which are not found in microsomes. It was found that hepatocytes give better predictions for Phase II glucuronidation than microsomes, even though both systems underestimated the extent of glucuronidation in vivo (31). Molecules have to permeate hepatocyte cell membrane to be metabolized, which is a closer mimic of in vivo. Hepatocytes are used commonly for studies of metabolic enzyme induction.

Plasma stability

Plasma stability is important for compounds that contain hydrolyzable functional groups, such as esters, amides, carboxylic acids, sulfonamides, lactams, and lactones. It also is used commonly to screen prodrugs and to select antedrugs. Plasma stability affects clearance and, therefore, half-life and oral bioavailability. It dictates how often and how much (dose regimen) a drug should be dosed (5, 8, 27). In vitro plasma stability assays can be used to predict in vivo clearance and oral bioavailability, guide structural modification to enhance stability, develop structure–metabolism relationships, and triage compounds for in vivo studies. Strategies to improve plasma stability are as follows: 1) Block the labile sites, 2) remove the labile groups, 3) reduce lipophilicity, and 4) introduce polar functional groups. For plasma stability, several approaches can be applied to reduce metabolism: 1) introduce electron-withdrawing groups and steric hindrance, 2) employ isosteric replacement of hydroxyl and carboxylic acid groups, and 3) use prodrugs.

The microsomal stability assay is the most common platform for assessing metabolic liability of a compound. High-throughput assays have been developed to screen large numbers of compounds. Several strategies have been developed to increase throughput and shorten data turnaround time, including the single-time point approach, sample pooling, online extraction, parallel analysis of multiple samples, and the use of a combined ESI–APCI ion source (ESCI) to increase sample coverage. A assay conditions are critical for the success of a microsomal stability assay (28). Substrate concentration cannot be too high (<1 μM), otherwise it can saturate the enzymes. The concentration of DM SO should be kept less than or equal to 0.2%. High DM SO concentration can inhibit the CYP450 enzyme. Microsomal protein concentration needs to be low (0.5 mg/mL) because nonspecific binding can affect assay results. Microsomal batches can vary greatly from batch to batch and from vendor to vendor; therefore, quality control is essential to generate consistent data. Insoluble compounds can produce artificially high stability estimates (29) because of precipitation and nonspecific binding to plasticware. Cosolvent can be used to overcome solubility effects.

Hepatocytes also are used for metabolic studies (30). Other metabolic enzymes are present in hepatocytes, which are not found in microsomes. It was found that hepatocytes give better predictions for Phase II glucuronidation than microsomes, even though both systems underestimated the extent of glucuronidation in vivo (31). Molecules have to permeate hepatocyte cell membrane to be metabolized, which is a closer mimic of in vivo. Hepatocytes are used commonly for studies of metabolic enzyme induction.
tamin basic amines and they are mostly CNS drugs. About 50% of drugs are metabolized by CYP3A, 10% by CYP2C, and 4% by CYP1A2. For the reasons above, the four most important CYP450 isozymes for drug–drug interaction studies are 3A4, 2D6, 2C9, and 1A2.

PhRMA (Pharmaceutical Research and Manufacturers of America) recently provided general guidelines on in vitro and in vivo drug–drug interaction studies, and leaders in the field are working toward standardization of the assays (36). The current acceptance criteria for CYP450 inhibition is \( \frac{C_{\text{max}}}{K_i} < 0.1 \) or \( I_{50} > 10 \mu M \) for early drug discovery.

Several methodologies have been developed to determine CYP450 inhibition. Some of these methodologies are implemented widely in the pharmaceutical industry (37).

**Fluorescent-based assay**

The fluorescent assay uses singly expressed recombinant human CYP450 isozymes. Substrates are coumarin-based or fluorescein-based compounds that can be metabolized by human CYP450 isozymes to form fluorescent metabolites (38). If a compound is an inhibitor of a CYP enzyme, the metabolic formation will decrease and lead to the reduction of the fluorescent signal, from which percent inhibition is calculated.

The assay is amendable to high throughput. Both the 96- and 384-well format are used commonly in the industry for screening of drug discovery compounds. The assay is used widely in the pharmaceutical industry to assess drug–drug interaction potential. MetaChip (Solidus Biosciences, Inc., Troy, NY) microarray technology has been developed with 11,200 reactions on a single chip and >1000-fold reduction in assay volume (39). The limitation of fluorescent-based assays is that if a test compound is autofluorescent or forms fluorescent metabolites, it can cause false-negative results. As such, fluorescent-based assays are not recommended for FDA submission. If a compound is a fluorescent quencher, false-positive results can be generated.
ADME Properties of Drugs

LC–MS-based assay

LC–MS-based assays use “drug-probes” that are specific substrates for certain isozymes. The formation of metabolites is monitored by using LC–MS–MS. Human liver microsomes are used commonly in the assay. It is important to keep the protein concentration low so that substrate conversion is less than 10% to minimize metabolism and generate relevant kinetic information (40). To increase throughput, a cocktail approach, which uses a mixture of probe substrates for the different isozymes and human liver microsomes, is helpful (41), or the cocktail assay, which uses a mixture of rh-CYP isozymes and a mixture of drug probes, may be used (42, 43). LC–MS–MS-based assay is used typically for regulatory submission or in vivo drug–drug interaction studies (36).

Luminescence-based assay

In the luminescence-based assay, luciferin derivatives are used as substrates and incubated with recombinant human CY P450 isozymes (44). Luciferin is released as a metabolite, which is converted to de-carboxyluciferin with light emission in the presence of luciferase and ATP. The assay is amendable to high-throughput screening of many compounds in 96-, 384-, or 1536-well format. It has the advantages of no fluorescent interference, a large dynamic range, and high sensitivity. Compounds that enhance or inhibit light emission directly can cause false negatives or false positives.

Other ADME/TOX Properties

Several other ADME/TOX properties are profiled commonly in drug discovery, such as uptake and efflux transporters, blood–brain barrier, plasma protein binding, lipophilicity, pKa, and pharmacokinetics. These ADME/TOX properties are discussed in other articles in this encyclopedia and elsewhere (5).

References

Further Reading


Skeett P. Introduction to Drug Metabolism. 2003. Nelson Thornes, Glos, United Kingdom.


See Also

ADME Properties of Drugs

ATP-Binding Cassette (ABC) Transporters: Small Molecules Passive Diffusion Across Membranes Systems Biology Approach to Metabolism Xenobiotic Metabolism, Chemistry of


Allergy and Asthma, Small Molecules to Target

Garry M. Walsh, University of Aberdeen, School of Medicine, Aberdeen, Scotland, United Kingdom
doi: 10.1002/9780470048672.wecb676

Current therapies for asthma and allergy are aimed at controlling disease symptoms. For most asthmatics, inhaled anti-inflammatory therapy is effective, but a subset of patients remains symptomatic despite optimal treatment creating a clear unmet medical need. Although considered a less-serious condition than asthma, allergic diseases are common and considerably impact on the quality of life of affected individuals. Innovative disease-modifying therapeutics are needed for both allergy and asthma. Biopharmaceutical approaches may identify small molecules that target key cells and mediators that drive the inflammatory responses that underlie the pathogenesis of allergy and asthma.

Asthma and allergy are important conditions whose complex pathology is influenced by both environmental and genetic factors. Although effective anti-inflammatory therapy is available for both conditions, this therapy is used when symptoms develop and is not always effective in all subjects with asthma and/or allergy. Asthma and allergy share some common pathological features including the overexpression of inflammatory mediators that result in leukocyte accumulation and changes in structural cell function. We have considerable knowledge regarding the cells and mediators that control the pathogenic changes in asthma and allergy. This information has informed the identification of small molecules that target aspects of the inflammatory cascade in asthma and allergy. This article will review the background and status of these novel compounds.

Biological Background

Asthma

Asthma is now one of the most common chronic diseases in developed countries and is characterized by reversible airway obstruction, airway hyper reactivity (AHR), and airway inflammation. Key pathological features include infiltration of the airways by activated lymphocytes and eosinophils; damage to, and loss of, the bronchial epithelium; mast cell degranulation; mucous gland hyperplasia; and collagen deposition in the epithelial sub-basement membrane area. Asthma pathology is associated with the release of myriad proinflammatory substances that include lipid mediators, inflammatory peptides, chemokines, cytokines, and growth factors. In addition to infiltrating leukocytes, structural cells in the airways, which include smooth muscle cells, endothelial cells, fibroblasts, and airway epithelial cells, are all important sources of asthma-relevant mediators (1). This complex scenario means that potential targets for therapeutic intervention are many and varied, and the task of successful therapy is challenging.

Anti-inflammatory therapy in asthma is largely reliant on glucocorticoids (GC)—particularly in their inhaled form—with or without the addition of short- or long-acting bronchodilators. The cysteinyl leukotriene-receptor antagonists are also an established part of the asthma armamentarium. Overall, they are less effective than inhaled GC, but some patients show a striking improvement, and a GC-sparing effect has been demonstrated. Although the symptoms of most asthmatics are satisfactorily controlled by regular use of inhaled GC, their use often raises concerns with respect to compliance, particularly in children and adolescents. Increasing evidence shows that long-term use of inhaled GC, especially in high doses, can cause systemic adverse effects, including adrenal suppression, reduced growth and reduced bone–mineral density, as well as local side effects such as dysphonia. Moreover, a significant subgroup of asthmatic patients responds poorly or not at all to high-dose inhaled or systemic GC treatment. As few alternative treatments are available, these patients can be difficult to treat and may require frequent hospitalization (1). Thus, identification of potential targets for therapeutic intervention is an important goal in asthma research.

Allergy

Allergic inflammation is characterized by an immediate immunoglobulin-E-dependent mast cell and basophil degranulation leading to the release of mediators such as histamine that is responsible for most immediate manifestations of allergic disease. Other mediators include platelet-activating factor,
prostaglandins, and cytokines, which include interleukin (IL)-3, IL-4, and tumor necrosis factor-α (TNF-α). Cytokines are particularly important in driving the subsequent late-phase reaction that results in the accumulation of leukocytes at the sites of inflammation. These inflammatory processes lead to various manifestations of allergic disease such as intermittent and persistent allergic rhinitis, conjunctivitis, atopic dermatitis, urticaria, and a degree of airway inflammation observed in allergic (extrinsic) asthma. The burden of allergic disease worldwide is such that it represents a serious public health problem that attracts considerable effort to identify effective and safe therapies.

Histamine has a prominent and diverse role in the pathophysiology of allergic disease. Therapeutic intervention is therefore commonly focused on blocking its interaction with H1 receptors. The H1 histamine receptor is a heptahelical transmembrane molecule that transduces extracellular signals to intracellular second-messenger systems via G proteins. H1 histamines act as inverse agonists that combine with the H1-receptor, which stabilizes it in the inactive form and shifts the equilibrium toward the inactive state (2). Although effective in the treatment of allergic rhinitis first-generation antihistamines, such as chlorpheniramine and promethazine, had a tendency to cross the blood–brain barrier resulting in unwanted side effects, particularly sedation and impaired psychomotor activity (3). In contrast, the second-generation histamine H1 receptor antagonists are highly effective and well-tolerated treatments for allergic disease and are among the most frequently prescribed drugs in the world (4). More recently, several novel anti-histamines, which include fexofenadine, desloratadine, and levocetirizine, have been developed that are either metabolites or active drugs or enantiomers with improved potency, duration, and onset of action, together with increased predictability of active drugs or enantiomers with improved potency, duration, and onset of action, together with increased predictability and safety. Anti-inflammatory effects in vitro and in vivo independent of H1-receptor blockade have been described for most antihistamines, and these may enhance their therapeutic benefit (3). Other treatments available include anticholinergics, decongestants, intranasal corticosteroids, leukotriene receptor antagonists, or desensitization with modified allergens. It is noteworthy that some of these treatments are often used in combination rather than as monotherapy (e.g., combination of an antihistamine with a nasal decongestant or with a leukotriene antagonist).

Small Molecules in Asthma

Interleukin-5

Much inflammation in asthma is thought to be a consequence of the inappropriate accumulation of eosinophils and the subsequent release of their potent proinflammatory arsenal that includes such diverse elements as granule-derived basic proteins, mediators, cytokines, and chemokines (5). Interleukin (IL)-5 is crucial to the development and release of eosinophils from the bone marrow, their enhanced adhesion to endothelial cells that line the postcapillary venules, and their persistence, activation, and secretion in the tissues. Several animal models of asthma including the use of primates have provided good evidence that inhibiting the effects of IL-5 using specific mAb-inhibited eosinophilic inflammation and AHR (6). Given its central role in regulating eosinophil development and function, IL-5 was therefore chosen as a potentially attractive target to prevent or blunt eosinophil-mediated inflammation in patients with asthma.

However, several clinical trials have reported disappointing clinical outcomes after treatment of asthmatic patients with an anti-IL-5 mAb. The first study was designed to validate the safety of the humanized anti-IL-5 mAb mepolizumab (7); it faced criticism of lack of power (8) and the validity of patient selection (9). A later placebo-controlled study (10) found that treatment of mild asthmatic patients with mepolizumab abolished circulating eosinophils and reduced airway and bone marrow eosinophils, but it reported no significant improvement of clinical measures of asthma. Critically, lung biopsy samples from the treatment group contained intact tissue eosinophils together with large quantities of eosinophil granule proteins; these findings likely explain the lack of clinical benefit after mepolizumab treatment. Similar findings were reported with the anti-IL-5 mAb SCH52709 in patients with severe asthma who had not been controlled by inhaled corticosteroid use. These authors reported profound reduction in circulating eosinophils but no significant improvement in either asthma symptoms or lung function (11). Compared with placebo, treatment of mild atopic asthmatics with mepolizumab significantly reduced the expression of the extracellular matrix proteins tenascin, lumican, and procollagen III in the bronchial mucosal reticular basement membrane. In addition, significant reductions were observed in both airway eosinophils that express mRNA for transforming growth factor-beta (TGF-β1) and the concentration of TGF-β1 in bronchoalveolar lavage (BAL) fluid (12). TGF-β1 is implicated in airway remodeling in asthma, and eosinophils are important sources of this growth factor, which thereby contributes to tissue-remodeling processes in asthma by regulating the deposition of ECM proteins. Mepolizumab may prove useful in preventing this. An alternative to the use of humanized anti-IL-5 mAb is the use of molecular modeling of the IL-5 receptor α-chain to develop specific receptor antagonists. Recently, such a compound (YM-90709) has been shown to be a relatively selective inhibitor of the IL-5R (13), and intravenous injection of YM-90709 inhibited infiltration of eosinophils into the BAL fluid of allergen-challenged BDF1 mice (14).

Interleukin-4

Another cytokine important in eosinophil accumulation is IL-4, which together with its close relative IL-13, is important in IgE synthesis by B cells. Both cytokines signal through a shared surface receptor, IL-4Rα, which then activates the transcription factor STAT-6 (15). Studies with soluble IL-4Rα given in a nebulized form demonstrated that the fall in lung function induced by withdrawal of inhaled corticosteroids was prevented in patients with moderately severe asthma (16). However, despite these promising findings subsequent trials have not been as successful; consequently, this treatment is no longer being developed. Other approaches for blocking the IL-4 receptor include administration of antibodies against the receptor and mutant IL-4 proteins. For example, a peptide-based...
vaccine for blocking IL-4 was recently developed by anti-
genic prediction and structure analysis of IL-4 receptor com-
plex. Vaccine construction involved a truncated hepatitis B core antigen as carrier with the peptide inserted using gene engi-
nearing methods. Compared with control animals, immunized allergen-challenged ovalbumin (OVA)-sensitized mice had sig-
nificant reductions in Immunoglobulin-E (IgE), eosinophil accu-
mulation in BAL, goblet-cell hyperplasia, tissue inflammation, and methacholine-induced respiratory responses (17).

Interleukin-13

IL-13 has been found in BAL after allergen provocation of asthmatic subjects, which strongly correlated with the increase in eosinophil numbers. mRNA expression for IL-13 was de-
tected in bronchial biopsies from both allergic and nonaller-
gic asthmatic subjects. In animal models, IL-13 mimics many proinflammatory changes associated with asthma (18). It is therefore another potential therapeutic target for the resolution of
airway inflammation. Two receptors for IL-13 have been described—IL-13 Rα1 and IL-13 Rα2. The latter exists in sol-
uble form, has a high affinity for IL-13, and can thus “mop up” secreted IL-13. In mice, IL-13 Rα2 blocked the actions of
IL-13, which included IgE production, pulmonary eosinophilia, and AHR (19). A humanized IL-13 Rα2 is now in clinical de-
vvelopment as a novel therapy for asthma. A mouse-based study reported that intratracheal administration of human IL-13
induced leukocyte infiltration in the lung, AHR, and goblet-cell metaplasia with allergic eosinophilic inflammation in the esoph-
agus. An anti-human IL-13 IgG4 mAb (CAT-354) significantly reduced many of these parameters. In contrast, another study
using mice sensitized by intranasal application of ovalbumin as a model of asthma/allergy found that the inhibition of the
IL-4/IL-13 system efficiently prevented the development of the
asthmatic phenotype, which includes goblet-cell metaplasia and
airway responsiveness to methacholine, but it had little effect
on established asthma (20). A humanized anti-IL-13 mAb called
IMA638 significantly reduced eosinophils, neutrophils, eotaxin,
and RANTES in BAL fluid from cynomolgus monkeys sensi-
tized to A. Ascaris suum after segmental antigen challenge, com-
pared with levels observed in control animals (21). IMA638
also gave dose-dependent inhibition of the antigen-induced late
responses and airway hyperresponsiveness in a sheep model of
allergic asthma that used animals with natural airway hypersen-
sitivity to A. Ascaris suum antigen (22).

As stated above, IL-4Rs is the signaling component of the
heterodimeric receptor complex shared by both IL-4 and IL-13. Therefore, it represents an attractive target to antagonize the
effects of both cytokines, as this approach may be more effec-
tive than targeting either IL-4 or IL-13 alone (23). Recently,
a recombinant human IL-4 variant pitrakinra (Aerovant) was
developed that competitively inhibits the IL-4 Rα receptor com-
plex to interfere with the actions of both IL-4 and IL-13. In
two independent small-scale parallel-group Phase IIa random-
ized, double-blind, placebo-controlled clinical trials, patients
with asthmatic asthma were treated with pitrakinra or placebo given
either as a single subcutaneous dose or via nebulization twice
daily. Active or placebo treatments were given for 4 weeks be-
fore the patients were given an inhalated inhalation challenge.
Compared with placebos, allergen challenge-induced decreases in
FEV 1 were significantly attenuated after 4 weeks of inhala-
tion of pitrakinra. The frequency of spontaneous asthma attacks
that required rescue medication use was also diminished in the
study in which pitrakinra was given subcutaneously (24). These
important findings support the hypothesis that dual inhibition of
IL-4 and IL-13 can affect the course of the late asthmatic re-
response after experimental allergen challenge. More large-scale
clinical trials on patients with day-to-day asthma are required to
establish fully whether pitrakinra is an effective and safe asthma
treatment.

Interleukin-9

A new TH2 cytokine, IL-9, and its receptor are found in asth-
matic airways in increased levels (25). IL-9 has several proin-
flammatory effects on eosinophils, which includes enhancement of
eosinophil IL-5 receptor expression, differentiation in the
bone marrow, and prolonged survival through inhibition of
apoptosis (26). Transgene expression of IL-9 in the lungs of mice resulted in lymphocytic and eosinophilic infiltration of
the lung, airway epithelial hyperplasia with mucus production,
and mast cell hyperplasia, as well as production of IL-4, IL-5,
and IL-13 (27). Treatment of OVA-challenged mice with an
anti-IL-9 antibody significantly prevented AHR in response to
a methacholine challenge together with reductions in numbers of
eosinophil and levels of IL-4, IL-5, and IL-13 in BAL (28).

Tumor necrosis factor-α

TNF-α is expressed in asthmatic airways and may play a key
role in amplifying airway inflammation through activation of
transcription factors such as NF-κB and AP-1. TNF-α has proin-
flammatory effects on eosinophils, neutrophils, T cells, and
endothelial cells. It is thought to contribute to AHR, airway re-
modeling, and GC resistance in asthma; therefore, it represents
a potential target for therapy. Humanized anti-TNF mAb (in-
fliximab) and soluble TNF receptor blockers (etanercept) have
been developed, and preliminary clinical studies have shown
significant improvements in lung function, airway hyperreactiv-
ity, and exacerbation rate, particularly in patients with severe
asthma refractory to GC treatment. However, some clinical stud-
ies reported negative findings, so heterogeneity seems to exist
in response to TNF-α antagonism. In addition, concerns exist
regarding potential side effects in some subjects treated with
anti-TNF therapy, which include higher rates of solid organ
malignancies or latent TB reactivation (29). Small-molecule
inhibitors of TNFα-converting enzyme have also been synthe-
sized; and these inhibitors may be attractive therapeutic targets
for asthma (30).

Immunoglobulin-E inhibitors

IgE plays a central role in the pathogenesis of diseases associ-
ated with immediate hypersensitivity reactions, which include
allergic asthma. IgE-dependent biological actions are a result of
its binding to high-affinity (FcεRI) receptors on mast cells and ba-
osphils and to low-affinity (FcεRII) receptors on macrophages,
dendritic cells, and B lymphocytes. Allergen molecules then
crosslink adjacent Fab components of IgE on the cell surface, which thereby activates intracellular signal transduction. In mast cells, this action leads to the release of preformed mediators and the rapid synthesis and release of other mediators responsible for bronchoconstriction and airway inflammation. Therefore, blocking the action of IgE using blocking antibodies that do not result in cell activation is an attractive approach.

Omalizumab (rhAb-E25) is a humanized monoclonal antibody directed to the FcRγRII binding domain of human IgE. It inhibited early-phase and late-phase allergen-induced asthmatic reactions and reduces serum-free IgE concentrations to less than 5% of baseline; it has now progressed through clinical development (31). A large Phase II trial studied fortnightly intravenous administration of omalizumab for 20 weeks in 317 patients (32), whereas two Phase III trials, which included over 500 patients each, studied omalizumab given subcutaneously every 2-4 weeks for 12 months (33, 34). Ayres and colleagues (35) examined the effects of Omalizumab in patients with moderate-to-severe allergic asthma whose symptoms were poorly controlled by high doses of inhaled GC. Omalizumab was administered for 12 months and benefited these patients as shown by a 30% reduction in their asthma deterioration-related incidents. A nother study reported that Omalizumab treatment of subjects with both persistent rhinitis and difficult-to-treat asthma resulted in significantly reduced asthma exacerbations and improved quality of life in those patients who received anti-IgE therapy over the 28-week study period (36). Omalizumab has also been shown to be beneficial as an add-on therapy in patients who have inadequately controlled, severe persistent asthma (37). Consistent findings from these trials showed that omalizumab is an effective therapy for patients with symptomatic moderate to severe allergic asthma despite treatment with GC and rescue medication. It reduced the frequency of exacerbations and improved symptom control while allowing a reduction in the use of GC and β2-agonists. It also improved patient quality of life and produced a significant improvement in lung function as measured by PEFR and FEV1. Omalizumab seems to be well tolerated with few side effects reported in these studies with no reports of circulating antibodies against omalizumab. Although more long-term studies are needed to elucidate the benefit and safety of anti-IgE therapy in asthma, its niche may be in the treatment of patients with severe asthma who are dependent on oral corticosteroids.

Eotaxin

Chemokines are a family of small, secreted proteins that control migration of monocytes, lymphocytes, neutrophils, eosinophils, and basophils. Eotaxin is an inducible, secreted chemokine that promotes selective recruitment of eosinophils from the blood into inflammatory tissues. It was first described in 1993 when intranasal injection of naive guinea pigs with BAL fluid from antigen-challenged guinea pigs resulted in the recruitment of eosinophils (38). Uniquely, the major characteristic of eotaxin is its selective ability to act on eosinophils. CCR3, which is a seven-transmembrane-spanning G protein-coupled receptor for eotaxin-1, is highly expressed on eosinophils and mediates the biological effects of other eosinophil chemokines, such as eotaxin-2, eotaxin-3, MCP-3, MCP-4, and RANTES. Furthermore, CCR3 is expressed not only on eosinophils but also on basophils (39), mast cell subpopulations (40), and activated TH2 cells (41), which might explain the coordinated recruitment of these cell types to sites of allergic inflammation. CCR3 is also expressed by airway epithelial cells (42), and although the bronchial epithelium consists of structural nonmigratory cells, expression of the CCR3 receptor may represent an autoregulatory feedback mechanism to monitor chemokine production. Furthermore, eotaxin produced by the epithelium may be sequestered by the CCR3 receptor and presented to infiltrating cells, which thereby enhances their activation; this phenomenon is observed with IL-8 and its receptor. Several clinical studies have suggested a pivotal role for CCR3 ligands/CCR3 receptor in the eosinophilic inflammation characteristic of atopic dermatitis, asthma, and allergic rhinitis; thus, blockade of this receptor may have pronounced beneficial effects in these diseases (43). Several small-molecule CCR3 antagonists have been described. In this regard, N-ureidoalkyl) benzziperidines have been identified as potent CCR3 antagonists, which inhibit eosinophil chemotaxis and calcium mobilization in the micromolar to nanomolar concentration range (44). The small-molecule selective CCR3 antagonist YM-344031 potently inhibited the binding of eotaxin-1 and RANTES to human cells transfected with CCR3, ligand-induced Ca2+ influx, and chemotaxis, together with inhibition of eotaxin-1-induced eosinophil shape change. Furthermore, both immediate and late-phase allergic skin reactions in a mouse model were significantly inhibited by orally administered YM-344031 (45). Another small-molecule selective CCR3 antagonist YM-355179 inhibited eotaxin-induced intracellular Ca2+ influx, chemotaxis, and eosinophil degranulation. It also inhibited eosinophil infiltration into airways of ovalum-challenged guinea pigs and prevented eosinophilic inflammation and airway remodeling. CCR3 antagonist treatment resulted in a marked reduction of eosinophils in the bronchoalveolar lumen and in airway wall tissue, whereas infiltration of lymphocytes or macrophages remained unchanged. Furthermore, antagonizing CCR-3 reduced AHR, goblet-cell hyperplasia, and airway remodeling as defined by subepithelial fibrosis, and it increased accumulation of myofibroblasts in the airway wall of chronically challenged mice. Therefore, antagonizing CCR-3 represents a novel approach and promising asthma or allergy therapy (47). Furthermore, evidence from animal models suggests that IL-5 and eotaxin may work in a synergistic fashion to promote the release of mature eosinophils from the bone marrow (48). Thus, it might be that combination therapies of CCR3 antagonist and humanized anti-IL-5 mAb might prove an effective approach to limit or prevent eosinophilic toxic in the asthmatic lung.
Targeting Inflammatory Cell Accumulation

Asthma pathology is characterized by excessive leukocyte infiltration that leads to tissue injury. Leukocyte extravasation, migration within the interstitium, cellular activation, and tissue retention are controlled by cell-adhesion molecules (i.e., selectins, integrins, and members of the immunoglobulin superfamily). Numerous animal studies have demonstrated essential roles for these cell-adhesion molecules in inflammation, which include L-selectin, P-selectin, and E-selectin, ICAM-1, and VCAM-1, together with many of the β1 and β2 integrins. For example, compared with wild-type mice, inhaled allergen challenge of P-selectin deficient mice resulted in fewer eosinophils in BAL fluid with significant reductions in AHR.

Selectins

Selectins are responsible for the early “rolling” adhesive events between leukocytes and the endothelial cells that line the postcapillary venules. The small-molecule selectin antagonist Bimosiamose (TBC1269) is a synthetic computer-designed selectin antagonist targeted against all three selectins (49). Therefore, these therapies represent potentially important therapeutic targets and these families of adhesion molecules have been under intense investigation by the pharmaceutical industry for the development of novel therapeutics.

Integrins

Selectins are also vital in early adhesive neutrophil interactions in a sheep model of allergic asthma, pretreatment with an L-selectin monoclonal antibody significantly reduced both the early and late airway response and significantly reduced AHR (49). Therefore, these therapies represent potentially important therapeutic targets and these families of adhesion molecules have been under intense investigation by the pharmaceutical industry for the development of novel therapeutics.

Immunotherapy

Asthma and allergy are manifestations of an imbalance in cytokine and signaling pathways that mediate inflammatory and structural changes within the affected tissues. Developing treatments include strategies to alter the cytokine/chemokine balance or to skew the cytokine profile away from T helper 2 (Th2) responses and toward Th1 responses. Immunotherapy may potentially attenuate symptoms by disease modification through the induction of tolerance to common environmental allergens rather than by suppressing inflammation. A major advantage is the potential for a positive effect to remain for several years after the end of the treatment period. The use of allergen-specific immunotherapy is not a new approach to asthma therapy, and until relatively recently the crude nature of the allergen extracts available meant that its use was limited by unwanted side effects, such as anaphylaxis. Strategies to overcome these problems include the use of hypoallergenic isoforms, recombinant allergens, or DNA vaccines (58). For example, in mice with chronic airway inflammation maintained by repeated ovalbumin inhalation, mucosal administration of CpG DNA oligonucleotides significantly reversed airway hyperreactivity together with both acute and chronic markers of inflammation (59). A newer approach is the use of short, synthetic, allergen-derived peptides that induce T cell tolerance but cannot crosslink IgE on mast cells or basophils and induce anaphylaxis. The main effect seems to be a shift from a Th2 to Th1 profile and induction of regulatory cytokines such as IL-10 and TGF-β. One study that used patients with asthmaic reactions to cats demonstrated that treatment with a desensitising vaccine based on many short, overlapping, HLA-binding, T cell peptides derived from Fel d 1 inhibited both early- and late-phase reactions to a subsequent whole-allergen challenge. Changes in immunological parameters included modulation of the proliferation of blood
monocellular cells together with their production of IL-4, IL-10, and IL-13, and of interferon-γ (60). A more recent study from the same group demonstrated that treatment of cat-allergic asthmatics with FeL-derived T-cell peptides significantly improved clinically relevant outcome measurements, which include reductions in late asthmatic reactions to inhaled whole cat dander and significant improvements in asthma-related quality of life (61). Although from these studies are encouraging, we require larger, dose-ranging studies before firm conclusions on clinical efficacy of peptide allergen therapy can be made. Although it is conceivable that targeting T cells may induce Th1-type autoimmune pathology in humans, to date no evidence indicates that this occurs. In addition, nonallergic mechanisms are also likely to be important in the pathogenesis of asthma; thus, it is possible that novel treatments targeted at the allergy fraction of the phenotype may not be as efficacious as hoped.

References

27. Temann U-A, Ray P, Flavell RA. Pulmonary overexpression of IL-9 induces Th2 cytokine expression, leading to immune pathol- Allergy, Small Molecules to Target
Allergy and Asthma, Small Molecules to Target

patients sensitive to cats: a randomised controlled trial. Lancet 2002;360:47.


See Also

Small Molecule Inhibitors, Design and Selection of Small Molecule Drugs that Target Chronic Obstructive Pulmonary Disease (COPD)

Chemotaxis

Integrin signalling
Lead optimization in drug discovery has changed significantly over the past five years and no longer is fragmented into separate hit-to-lead and lead optimization phases. Chemical lead optimization from high-throughput screening (HTS) to clinical candidate identification is now one seamless process that draws on new technologies for accelerated synthesis, purification, and screening of directed, iterative compound libraries. Advances in high-throughput screening technologies allow detection of new allosteric modes of target modulation, which provides new chemotypes and target opportunities. With the incorporation of drug metabolism and pharmacokinetics (DMPK) inputs early in the lead optimization work flow, molecules are not optimized solely for target potency and selectivity. Moreover, "closed-loop" work flows are in place such that synthesis and primary screening operate on a 1-week turnaround for up to 48 compounds/week with DMPK data cycling every other week to guide compound design, which provides expedited timelines for the development of proof-of-concept compounds and clinical candidates with limited human resources.

The competitive drug discovery environment, whether in industry or academia, requires constant innovation and refinement as a prerequisite for success. A combination of market, patient, and regulatory concerns requires that new chemical entities act on truly novel, and therefore not clinically validated, molecular targets. With the high attrition rates and limited human resources, drug discovery efforts must focus on a large and diverse collection of molecular targets and judiciously employ enabling technologies and new paradigms to develop simultaneously multiple early stage programs. Importantly, the goal at the outset of a nascent program is to provide rapid target validation in vivo with a novel small molecule or to deliver a quick kill for the program so that resources can be reassigned. Coupled with these concerns is the need to establish intellectual property early in the development process, as chemical space is shrinking at an alarming rate (1).

Historically, the scope, mission, and technology platforms of lead optimization groups varied considerably across the drug discovery industry, which led to highly variable success rates (2–5). Some organizations had defined “hand-off” criteria and fragmented lead optimization into a hit-to-lead phase and a chemical lead optimization phase. Hit-to-lead focused on optimizing screening hits, usually by library synthesis (solution phase and/or solid phase), for target potency with minimal concern for selectivity, ancillary pharmacology, and pharmacokinetics (PK). Leads that met certain potency criteria and displayed robust structure-activity relationships (SAR) then would be "handed-off" to a second group for the lead optimization phase, wherein more classic medicinal chemistry (single compound synthesis and intense drug metabolism and pharmacokinetics (DMPK) profiling) would occur (2–5). In the past five years, lead optimization in drug discovery has changed significantly and no longer needs to be fragmented into separate hit-to-lead and lead optimization phases (2–5). Major advances have been made in HTS technologies, which have enabled detection of novel modes of target modulation.
Once limited to detection of classic agonists and antagonists by HTS, kinetic imaging plate readers allow for the HTS identification of positive and negative allosteric modulators of both known and novel targets, which offers new chemotypes as well as improved selectivity and safety profiles (6, 7). Chemical lead optimization, from evaluation of screening hits to clinical candidate identification, now can be a seamless process that draws on new technologies for accelerated synthesis, purification, and screening. Directed, iterative compound libraries now are employed throughout the lead optimization continuum with single compound synthesis restricted to an “as needed” basis. With the incorporation of DMPK inputs at the initiation of a lead optimization program, molecules are not optimized solely for target potency and selectivity but also for the optimization of protein binding and pharmacokinetics and for diminishing CYP inhibition (1–11). Moreover, “closed-loop” workflows are in place such that chemical synthesis and primary screening data operate on a 1-week turnaround for hundreds of compounds/week, with DMPK data cycling every other week to guide compound design and to provide expedited timelines for the development of proof-of-concept compounds to validate/kill novel molecular targets and to deliver clinical candidates with limited human resources. To avoid the negative stigma of combinatorial chemistry, both industrial and academic laboratories, this new paradigm for lead optimization is coined “technology-enabled synthesis” or “TES”; however, a more accurate moniker would be “technology-enhanced medicinal chemistry” (12–21).

**Advances in High-Throughput Screening Technologies**

The first step toward a successful lead optimization campaign begins in a state-of-the-art screening facility, which is ideally based on a philosophy that values the ability to automate complex biologic assays to allow screening of difficult-to-screen targets and to detect novel mechanisms of target modulation. Historical HTS paradigms valued the use of automation only to increase throughput; however, the focus now is to execute fully complex tasks with high precision. Modern HTS facilities employ automated screening systems composed of state-of-the-art liquid handling, plate readers, incubators, and other instruments to support a wide variety of cell-free and cell-based assays that range from enzyme assays on purified proteins to phenotypic screens on model organisms like C. elegans and zebrafish embryos (6, 7, 21–24). Advances in analysis software allow for information-rich assay forms, primarily in cell-based or organism-based environments, with read modes based on either parallel acquisition of kinetic data that use instruments like the Hamamatsu FDSS kinetic imaging plate readers (21, 25) or on object-based screening that uses high spatial resolution devices like automated microscopes or the BlueShift Isocyte (21, 26). Both of these read modes yield complex, information-rich data sets. The analysis and storage of such data can be challenging; however, the success of a lead optimization campaign is linked directly to the ability to acquire, synthesize, store, and present compounds as well as to the ability to collect/analyze data from the biologic systems for which we hope to discover proof-of-concept compounds and clinical candidates (6, 7, 21–26).

**A triplicate screen to identify classic and allosteric modes of target modulation**

Miniaturization of assays that employ kinetic imaging plate readers allow for the development of robust high-throughput calcium mobilization-based assays that detect the activation/inhibition of molecular targets through both classic and allosteric modes of target modulation. For instance, we can measure receptor-induced intracellular release of calcium by using an imaging-based plate reader that makes simultaneous measurements of calcium levels in each well of a 384-well plate. In a novel triplicate-screening paradigm (Fig. 1), either vehicle or a test compound was added to cells expressing a G Protein-coupled receptor (GPCR) that has been loaded with a fluorescent dye, Fluo-4. After a 2.5-minute incubation period, an imaging-based plate reader that makes simultaneous measurements of calcium levels in each well of a 384-well plate. In a novel triplicate-screening paradigm (Fig. 1), either vehicle or a test compound was added to cells expressing a G Protein-coupled receptor (GPCR) that has been loaded with a fluorescent dye, Fluo-4. After a 2.5-minute incubation period, a nearly maximal (EC80) concentration of orthosteric agonist was added, followed by a nearly maximal (EC80) concentration added 1 minute later. In this manner, we can screen for and identify classic agonists/antagonists, allosteric potentiators, and antagonists simultaneously, which maximizes the efficiency of each screen and delivers a diverse collection of hits for chemists to optimize. This paradigm affords the medicinal chemists with options, both in terms of a modulatory mechanism for their therapeutic target and in terms of a chemotype, for the lead optimization campaign in a manner previously unavailable (27).

Of course, technology has not advanced only for the screening of GPCRs but also for kinases and ion channels. Kinase screens now employ both low and high concentrations of ATP to identify both ATP-competitive and allosteric inhibitors. Numerous technology platforms recently have appeared for ion channels targets, such as highly automated Ion-Works and Q-patch, which avoid the need for burdensome and slow single patch-clamp experiments (21–27).

**Solution-Phase Parallel Synthesis for Lead Optimization**

The chemical technologies and platforms for chemical lead optimization have undergone a major paradigm shift in the past 10 years. In the 1990s, hit-to-lead efforts were driven by combinatorial chemistry and characterized by large (1000–10,000 member) solid-phase libraries that required months to synthesize and characterize (28). Often, by the time the library was ready for screening, the SAR of the program and/or lead series had moved on, and the value of the library was minimal (28, 29). As a result, most pharmaceutical companies disbanded their combinatorial chemistry groups, and lead optimization relied primarily on single-compound synthesis or small collections...
Figure 1  Triplicate screen to identify allosteric modes of target modulation. (a) Vehicle with an EC20 and EC80 of agonist, (b) waveform profile of an agonist, (c) waveform profile of an antagonist (flowup necessary to distinguish orthosteric versus allosteric antagonist), and (d) waveform profile of a potentiator, aka, positive allosteric modulator. A single screen generates an entire spectrum of hits.

Less than 12 of compounds. Driven to make the lead optimization process more efficient, the concept of solution-phase parallel synthesis began to gain favor, and technologies rapidly began to develop to create this new approach (30–34). In the last five years, major advances were made in the availability of polymer-supported reagents and scavengers and in the advent of precision-controlled, single-mode microwave synthesizers for organic synthesis. Along with the development of robust mass-directed, preparative HPLC purification, platforms have revolutionized and accelerated lead optimization (20, 30–36).

Solution-phase parallel synthesis (SPPS)

Key to the success of SPPS was the development of “scavenging reagents.” Scavenging (quenching) reagents are highly effective tools for the rapid purification and isolation of the desired product(s) from a solution-phase reaction by forming either covalent or ionic bonds with excess reactants and/or reaction by-products. In general terms, scavenging can be considered a “phase-switching” technique wherein a chemo-selective reaction is employed to switch the phase of one product relative to another by virtue of a “tag” attached to the scavenging reagent [30–34]. Three major classes of scavenging reagents are categorized by the nature of the phase-tag: solid-phase polymers, ionizable functional groups, and fluoroalkyl chains [30–34, 37]. In a typical scenario, an excess of reactant B is combined with A to provide product P along with B and other reaction by-products X in a homogeneous solution-phase reaction. Then, B and X are chemo-selectively removed from solution in a subsequent “scavenging” step with a scavenging reactant 1 linked to a phase tag. After separation of the resulting phases, the product, P, is obtained in high purity by simply evaporating the solvent (Fig. 2).

The most commonly used tags are solid-phase polymers, and hence, a wealth of literature centers on the applications of polymer-supported scavenger reagents to transfer a captive species from the organic liquid phase to the solid phase for removal by filtration. Indeed, this approach has gained widespread acceptance because of the commercial availability of a diverse array of electrophilic and nucleophilic polymer-supported scavenging reagents and scavengers can be used simultaneously (30–34).

A rather commonly used tagging strategy involves linking a scavenger to an ionizable functional group, such as a COOH (pKa < 5) or an NR2 (pKa > 10). In this instance, the captured species can be phase transferred selectively by either pH-adjusted liquid/liquid extraction or by solid-phase extraction (SPE) on an ion-exchange cartridge that leaves the desired product either in the organic liquid phase or in the SPE cartridge eluent (34). SPE is a very attractive method for purification because a crude reaction simply is applied to a disposable silica plug and grafted with either a sulfonic acid (SCX—strong cation exchange) or a tertiary amine (SAX—strong anion exchange), and neutral molecules are eluted off with methanol, whereas ionizable functional groups are retained on the SPE cartridge. Unfortunately, this strategy impacts the diversity of a library by limiting the presence of ionizable groups to either neutral or orthogonally charged library members (34).
Lead Optimization in Drug Discovery

Relying on the affinity that fluoroalkyl chains have for each other and the phobia that they exhibit toward both organic molecules/solvents and aqueous solvents, researchers began examining fluorous tags as a means of phase switching (35). Initially, efforts centered on “heavy” fluorous tags (60% or more fluorine content by molecular weight; for example, 18 or more difluoromethylene, CF₂, groups) that used liquid/liquid phase separation to isolate fluorous-tagged molecules from untagged organics. Typically, a three-phase liquid/liquid extraction, which requires an organic layer, an aqueous layer, and a fluorous layer (a perfluorohexane such as FC-72), delivers pure material. More recently, fluorous solid-phase extraction (FSPE) that employs fluorous silica gel (reverse-phase silica gel with a fluorocarbon bonded phase) has been developed to separate effectively both “heavy” fluorous-tagged molecules as well as “light” fluorous-tagged molecules (4 to 10 CF₂ groups) from untagged organics. The FSPE columns, referred to as FluoroFlash (Fluorous Technologies, Pittsburgh, PA) columns, retain the fluorous-tagged material when eluted with a fluorophobic solvent, such as 80/20 MeOH/H₂O, which allows the untagged organic molecule to elute rapidly from the column. Homogeneous reaction kinetics, generally with respect to charged and neutral functional groups and a variety of efficient phase-separation options, have spurred a dramatic increase in the development of fluorous scavenging reagents and protocols (37).

Microwave-assisted organic synthesis (MAOS)

MAOS, fueled by the development of precision-controlled, single-mode microwave reactors, has a profound impact on organic and parallel synthesis. Reaction times typically are cut by orders of magnitude, and it is usual to observe a diminution in side product formation. Moreover, MAOS reactions tend to be general in scope and lend themselves to the synthesis of libraries to develop SAR rapidly. These advantages, easily appreciated when considering established routes with successful reactions, are even more valuable when working out robust conditions for a synthesis. Exploratory reactions can be conducted in minutes to hours instead of days, and speculative, higher-risk ideas can be pursued with minimal time investment. Indeed, MAOS allows any chemistry to be pursued in parallel and allows chemistries that historically were avoided for library synthesis (multicomponent reactions, organometallics, transition-metal catalyzed couplings, etc.) to be completed successfully in minutes (20, 38, 39).

Beyond the speed advantage, two additional merits of MAOS and modern reactors should be highlighted: precision and reaction scope. As has been noted in these pages and elsewhere, the benefits of MAOS have been studied in multimode “kitchen microwaves” for decades; what prevented acceptance in the wider community was irreproducibility because of a lack of pressure and temperature control (Fig. 3). In addition, kitchen microwaves employ multimode resonators that lead to a heterogeneous field and local “hot” spots (Fig. 3a); despite this disadvantage, early work demonstrated the use of MAOS. Modern systems (Fig. 3b) provide a homogeneous field and precise control of temperature and pressure, and they bare little resemblance to kitchen microwaves. Importantly, MAOS relies on dipolar oscillations and ionic conduction, for example, molecular friction, to generate heat and to afford uniform heating of the sample. In contrast, conventional thermal heating relies on heat transfer from the walls of a reaction vessel and affords nonuniform heating of the sample (Fig. 3c). This uniform heating and rapid time to set temperature delivers reproducible results with fewer side products and, as a result, higher chemical yields (20).

Also, MAOS technology significantly has impacted library design and synthesis. For instance, when presented with a small heterocycle as a hit from an HTS, MAOS technology allows one not only to synthesize and evaluate substitutions on the parent heterocyclic scaffold rapidly but also to synthesize and evaluate multiple heterocyclic templates with diverse substituents in parallel (Fig. 4). Therefore, a single library will contain multiple heterocyclic cores with varying degrees of basicity and topology, while broadening the generic scope for a composition of matter patent (38).

Mass-directed preparative HPLC

Despite the purity obtainable by SPPS scavenging (typically >90%) and the high purities obtained by MAOS (also typically >90%), modern lead optimization programs require >95% purity of all compounds that contribute to the development of
SAR and that advance into DMPK assays. For years, many labs employed UV-directed preparative HPLC and often multichannel units to increase throughput. Although this approach worked, purification of a single sample might lead to 30–40 fractions per sample, which then required analysis by analytical Liquid Chromatography/Mass Spectrometer (LCMS) to identify which fractions contained the desired product (34, 40). In 2000, several vendors launched preparative LCMS units that offered mass-directed fractionation. Now, purification of a single crude sample afforded only one or two pure fractions—a significant advance. Additional modifications for library purification included DMSO slugs to bracket sample injections or “at-column dilution” to provide robust chromatography and prevent in-line sample precipitation before the column. These modified systems were capable of purifying, in a single pass, 60 to 80 compounds per day with purity levels exceeding 98%; however, the systems required an expert chromatographer to develop custom gradients for each sample in a library (34, 40).

Recently, several vendors launched an analytical-to-preparative LCMS software package that addressed the need for a dedicated, expert chromatographer to operate each prep LCMS instrument. With analytical-to-preparative software, a file containing the compound ID and exact mass for each sample in a library to be purified is uploaded into the preparative LCMS system, which then electronically accesses the analytical LCMS data and extracts the retention time of the mass of interest from the crude sample chromatograms. The preparative LCMS system analytical-to-preparative software then calculates a customized gradient for each sample in a library and therefore reduces the need for an experienced chromatographer to achieve excellent first-pass purification results. This feature also allows the instrument to run overnight unattended and additionally increases operational efficiency.

Postpurification sample handling and compound characterization

Modern parallel synthesis laboratories, for example, high-throughput medicinal chemistry laboratories, have borrowed a page from the automotive industry and have developed highly efficient assembly lines for postpurification sample handling and compound characterization. Automated weighing systems with bar code readers scan and record weights on unique bar-coded vials into which pure compounds from the preparative LCMS systems are transferred for concentration in a sample evaporator. After the dry-down step, the bar-coded vials with pure, solid sample are transferred to a liquid handling robot. This instrument scans each bar code, weighs the vial, and determines the net weight of the pure product. This data file is merged with a registration file that contains the molecular weight of the compound, and the system software then calculates the volume of DMSO required to dilute the samples to a preset concentration.
Lead Optimization in Drug Discovery

![Diagram](image)

**Figure 4** MAOS to access rapidly diverse heterocyclic scaffolds from a common intermediate.

**Lead Optimization in Drug Discovery**

**Expedited Drug Metabolism and Pharmacokinetics**

Lead optimization involves more than just optimizing for target potency. New technology allows early lead optimization campaigns to address multiple parameters and inputs for each round of iterative library synthesis. These inputs allow for the rapid development of potent compounds with drug-like profiles, as opposed to just potent compounds. These data provide “quick kills” to individual leads or series and allow the lead optimization effort to re-direct resources toward more productive leads (1-5).

**High-throughput in vitro drug metabolism assays**

Significant effort has been applied to the miniaturization and DMSO compatibility of in vitro drug metabolism assays. Now, a 48-member library in a 96-well plate of DMSO stock solution can be evaluated rapidly in cytochrome P450 inhibition assays (CYP3A4, 2D6, 2C9), protein binding assays (rat, dog, and human), logP, HERG binding, and other standard assays (8-11, 42, 43). Previously, chemists were forced to choose which compounds to evaluate in these assays, and often only the most potent analogs would be selected; the most potent analogs were not necessarily the ones with the most promise as clinical candidates. Being able to acquire these data for an entire library provides opportunities to pursue leads within a series with the most balanced potency and DMPK profiles. This ability is of critical importance in lead optimization to ensure that drug-like leads are being pursued and additionally refined. Similarly, cassette dosing of compounds in liver microsomes and hepatocytes enables evaluation of an entire library in short order (8-11, 42, 43). This timely evaluation is especially valuable in the lead optimization of a backup clinical candidate wherein the clinical candidate is the positive control and new compounds (typically five to six per cassette) are viewed qualitatively as more or less stable than the first clinical candidate. Once an in vitro/in vivo correlation can be established for a given series, these rapid cassette experiments can drive a lead optimization program and require only intermittent in vivo experiments (8-20, 42, 43).

**Pharmacokinetic cassettes**

Resource and practical constraints prohibit acquiring single rat and dog pharmacokinetics (PK) (i.v. and p.o.) for every member of a library; however, the compound with the best PK may not be the most potent analog in a library, and knowing this data is crucial for lead optimization. In fact, chemical lead optimization programs have been guided solely by optimization of PK. With the success of in vitro cassette paradigms for microsomal and hepatocyte stability, the concept recently was extended to in vivo PK. In rats and dogs (Fig 5) so that an entire library could be evaluated in vivo employing a limited number of animals (8-11). However, some caveats exist. Combinatorial oral dosing to determine oral bioavailability (%F) in cassette format generally proved not very reproducible nor in agreement.
Lead Optimization in Drug Discovery

Figure 5  Pharmacokinetic evaluation of libraries: cassette dosing to evaluate clearance rates relative to a bid control compound.

with single PK experiments. In contrast, intravenous cassette dosing in both rats and dogs proved highly reproducible and
within the error of a single PK experiment and is a valuable tool to determine qualitative rates of clearance between five to
six new compounds and an internal control of known clearance (8-11). PK cassettes employ an overall low dose of test
compounds to minimize potential drug-drug interactions. These rapid cassette experiments prioritize which compounds from a
library then should be studied in single i.v./p.o. single animal PK studies. As shown in Fig. 5, compound 2 has qualitatively
lower intrinsic plasma clearance rate than an internal control compound with known bid (twice daily predicted dosing) PK
(Cl = 12 mL/min/kg), whereas the other four compounds in the cassette have higher clearance and need not be studied more.

Expedited, “Closed-Loop” Work Flow for Lead Optimization

Combining all above-mentioned technologies and paradigms
for synthesis, screening and DMPK evaluation affords an ag-
gressive, expedited process for chemical lead optimization
(1, 12–21). This protocol allows one to two synthetic chemists
to support a chemical lead optimization effort with acceler-
ated timelines that deliver proof-of-concept compounds within 6
months and clinical candidates within 12 months of the initiation
of a lead optimization campaign.

Library design

Independently, the technologies and strategies described herein
provide improvements for chemical lead optimization; however,
when they become closely aligned with screening and DMPK
resources in a “closed-loop” paradigm, the impact on drug
discovery is exponential [Fig. 6]. Starting from an HTS hit,
considerable attention is devoted first to library design, without
question the most important component of a successful lead
optimization effort. Library design changes over the course of
a lead optimization campaign. The initial design strategy is to
explode SAR around a screening hit and to be as diverse as
possible with respect to monomer input and analog synthesis to
rapidly identify productive changes for additional optimization.
In addition, this component of lead optimization is conducted
often in parallel, wherein a single chemist simultaneously will
synthesize diversity libraries around four to six hits to identify
expeditiously the best leads for additional optimization. After
this initial diversity-oriented explosion, library design must
become more focused to impact drug discovery goals: Random
libraries do not accelerate programs. It is important to approach
directed library design from a medicinal chemistry perspective
and to assemble the library as a collection of single compounds
designed to address a particular issue. For example, the design
of a 24-member library should involve careful thought regarding
what four single compounds would be synthesized first to
increase potency, improve PK, and so forth. Then, for each
of the first four analogs synthesized, designers should consider
what the next four analogs should be if the first changes were
productive or nonproductive. This exercise in library design
generates quality data that drive a lead optimization program
toward proof-of-concept compounds and clinical candidates
very quickly (1, 12–21).

Division of labor

A further key feature of the “closed-loop” approach to lead opti-
mization involves division of labor and the transfer of samples
from medicinal chemists to the analytical chemists. In this
paradigm, the medicinal chemists design and synthesize the
compound libraries (24 to 96 compounds) and obtain analyti-
cal LC/MS reports for each member of the library. At this
point, the medicinal chemists transfer the crude samples to the analytical chemists who purify the libraries by mass-directed preparative HPLC to >98% by using analytical-to-preparative software and perform all postpurification sample handling and coordinate submission of samples, in a 96-well plate format, to the biologists and DMPK personnel for screening (vide supra) (12-20). If resources allow, this division of labor affords opportunities for the medicinal chemists to focus on library design, develop and optimize new chemistries, and pursue multiple lead series in parallel (1, 12-21).

Data turnaround

The success of this paradigm hinges on rapid screening and dissemination of data to the medicinal chemists so that the next iteration of library synthesis can be initiated. To facilitate this, the delivery of compounds is coordinated with the biologists and assays are run the same day the compound libraries are delivered. Biologic data then is returned within 24 hours of the receipt of the libraries. This allows lead optimization to operate on a 1 week turnaround between the initiation of chemical synthesis and the generation of primary assay data. Secondary and/or selectivity data typically tail primary data by 1 to 2 days. As these data trigger the need for DMPK information, DMPK data typically follow a week after the initial assay data is obtained. Overall, this expedited process parallels traditional singleton medicinal chemistry work flows but generates data on hundreds of compounds in the time it used to take to evaluate just a few compounds. Moreover, this protocol allows one to two synthetic chemists to support a chemical lead optimization effort with accelerated timelines delivering proof-of-concept compounds within 6 months and clinical candidates within 12 months of the initiation of a lead optimization campaign. It is important to note that this lead optimization paradigm requires collaboration, close and frequent communication with biology and DMPK colleagues, sophisticated databases to store the volumes of data generated, and a major investment in technology (1, 12-20).

References


Figure 6  Expedited ‘‘closed-loop’’ lead optimization paradigm.
Lead Optimization in Drug Discovery


Further Reading


See Also

48. ADMET Properties of Drugs

49. High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology

50. Solution-Phase Synthesis of Biomolecules

51. Synthetic Chemistry: Building Molecules to Modulate Biological Systems

52. Target Family-Based Compound Library: Optimization, Target Selection and Validation
The use of metal ions in diagnostic medicine is widespread. Metal complexes and nanoparticles are used clinically in magnetic resonance imaging, nuclear medicine (gamma scintigraphy and positron emission tomography), and X-ray imaging. Preclinically, additional advances are being made in these fields through the creative application of metal complexes and metal-containing nanoparticles. This article briefly describes different clinical imaging modalities and the physico-chemical properties required for imaging agents. Commercial agents are used to illustrate these properties and to describe the clinical state of the art. Selected examples of exciting preclinical imaging agents that have shown efficacy in vivo also are described.

Imaging agents are defined here as compounds that either are required to generate an in vivo image or provide additional contrast to an existing in vivo image. This article describes the use of metal-containing compounds for clinical or preclinical in vivo imaging. This topic covers modalities such as magnetic resonance imaging (MRI), nuclear imaging (gamma scintigraphy, single photon emission computed tomography (SPECT), and positron emission tomography (PET)), X-ray imaging, and computed tomography (CT). Ultrasound contrast agents are used clinically, but these agents do not require a metal ion. For space constraints, cell tracking studies by MRI, reporter gene imaging with PET, and in vitro imaging will not be discussed. The focus will be on compounds used in clinical medicine and those with potential for clinical use. References are limited to 40, so where possible, relevant reviews have been cited. This article aims to provide a sense for the properties (chemical, magnetic, nuclear, etc.) required for an imaging agent, the current clinical practice, and the current promising directions in preclinical research.

Overview of Imaging Agents and Imaging Modalities

General requirements

Imaging agents are used for noninvasive in vivo diagnosis. They generally have found usage where the diagnostic information yielded impacts a life-threatening and/or expensive clinical decision or where the imaging test replaces a more invasive, risky procedure. An example of this use is imaging myocardial perfusion to test for the presence of coronary artery disease. Absence of disease sees the patient sent home, whereas presence of disease results in cardiac catheterization and possible percutaneous or bypass intervention.

For approval, the agent must demonstrate efficacy compared with an accepted gold standard (biopsy, other established imaging modality, or patient outcome). Because imaging agents themselves offer no therapeutic benefit, the safety requirement is quite high. Potential side effects must be minimized, and for nonendogenous metals, such as gadolinium, it is critical that the entire dose be eliminated from the body. The effective concentration range of the agent depends on the technique. For X-ray agents, millimolar tissue concentrations are required to generate contrast. For such highly concentrated agents, it is important to minimize the osmolality of the injected solution to prevent osmotic shock. Nuclear agents are effective at nanomolar or lower concentrations, and toxicity is usually less of an issue; however, receptor-targeted agents should not show any agonist activity.

Metal-essential versus metal-tagged imaging agents

For the compounds described here, the metal is essential for providing image contrast or for creating the image. However, the biologic distribution of the agent can depend primarily on the metal (metal essential) or the distribution can be dependent on a conjugated targeting vector, such as an antibody (metal-tagged) (Fig. 7). In all modalities, metal-essential imaging agents exist. These compounds often have nontargeting and are used to image the blood vessels or organs in an elimination pathway.
nonspecific protein binding. A linker can represent points to alter lipophilicity, charge, and especially with smaller targeting vectors. The complex and the linking radiation, has excellent spatial resolution (routinely 1 mm³ to create CT contrast. MRI does not require ionizing contrast agents are required at very high concentrations to the patient is high, soft tissue contrast is poor, and exogenous contrast is needed. PET imaging can provide quantitative data on receptor occupancy. However, spatial resolution is much worse with these nuclear techniques which enables the imaging of low-concentration targets, like neurotransmitters, with these techniques. PET imaging can provide quantitative data on receptor occupancy. However, spatial resolution is much worse with these nuclear techniques (1-5 mm clinically) than with CT or MRI and the images are susceptible to attenuation and motion artifacts. Nuclear agents typically must be prepared on the day of use because of their short half-lives. These agents also bring a radiation dose to the patient.

Clinical hybrid PET-CT and PET-MRI systems are being developed now. These approaches combine the high-resolution anatomical imaging of CT or MRI with the sensitive molecular targeting of PET.

X-ray or CT Contrast Agents

Planar X-ray imaging is the oldest and most common radiographic technique. CT (formerly, computed axial tomography—CAT scan) is a method in which a series of two-dimensional X-ray images are acquired around an axis of rotation and these images are combined mathematically to generate a three-dimensional image. Contrast in X-ray imaging derives from heavier atoms (Ca, P) and denser tissue (bone) absorbing and scattering the X-rays. Compounds administered to provide additional contrast are termed contrast agents, contrast media, or X-ray dyes. These compounds contain atoms with a large atomic number, Z, such as iodine (Z = 53).

The most common X-ray agents are based on highly soluble tri-iodo aryl derivatives. However, some metal-containing X-ray contrast media exist. The most common is barium sulfate, administered as a suspension to provide contrast in imaging the gastrointestinal tract. Barium (Z = 56) is quite opaque to X-rays, and the sulfate salt is very insoluble, which results in negligible oral bioavailability. The insolubility of BaSO₄ obviates its use for intravenous applications, and soluble Ba(II) complexes are quite labile, releasing the toxic free metal ion.

Gadolinium MRI contrast agents have been used clinically for X-ray contrast (2). These agents have been used mainly with patients known to have an allergy to iodinated contrast. The use of Gd(III) complexes has derived from the excellent safety profile of these agents, the ability to formulate them at the molar concentrations required for injection, the high atomic number of Gd, and the fact that they are available clinically. These compounds are given intravenously or intra-arterially for angiographic imaging.

Animal studies have been reported with other Gd(III) compounds and where the Gd(III) was replaced by other lanthanides, such as Yb(III) or Dy(III). The liver-specific agent [Gd(EOB-DTPA)]²⁻ and its Yb(III) and Dy(III) analogs were evaluated in rat, rabbit, and dog models and could provide good contrast to distinguish liver tumors (2). Recently, bis-methyl sulphide nanoparticles coated with the biocompatible polymer polyvinylpyrrolidone were described (3). Analogous to the iron oxide particles (see below), these nanoparticles were shown to provide contrast in the blood pool and in lymph nodes. Coated gold nanoparticles also have been studied in an animal model (4). Earlier efforts involved metal clusters such as polyoxotungstates; Yu and Watson give an excellent review of metal-based X-ray contrast media up to 1999 (5).

MRI Contrast Agents

Clinical MRI typically involves imaging the hydrogen atoms in water and fat. These molecules are the most abundant molecules in the body, and hydrogen has the highest NMR sensitivity of elements found in the body. A strong magnetic field is required for detection (clinical fields range from 0.3 to 7.0 tesla, most common is 1.5 T). Applied magnetic field gradients provide spatial encoding that allows an image to be reconstructed.

In CT, contrast occurs solely because of the attenuation of the X-ray beam by its interaction with tissue. Contrast in MRI depends on how the image is acquired. Proton density images reflect the different amounts of water content in different tissues. Chemical shift imaging can distinguish water from fat, based on the different chemical shifts of these protons. Contrast also can be altered in other ways: by the exploitation of differences in water diffusion in tissue, by chemical exchange, and by magnetization transfer effects, to name a few. The majority of scans used clinically derive contrast from differences in the relaxation times T₁, T₂, and T₂*.

When nuclei are excited by a radiofrequency (RF) pulse, T₁ is the time constant for the realignment of spins with the static magnetic field (along the z-axis, the longitudinal relaxation). T₁ ranges from about 300 ms for fat to several seconds for cerebrospinal fluid. If the delay between repeated acquisitions is not long enough, then magnetization will not have recovered.
MRI contrast agents provide increased contrast by acting on T1, and TR, different degrees of soft tissue contrast can be obtained. In a spin-echo sequence, an initial 90° pulse is followed by a delay, TE. Then, the dephasing nuclei are refocused because of a 180° pulse and the signal of this spin echo is acquired. Tissue with T2 long compared with TE will produce more signal than tissue with T2 < TE, which gives rise to T2-weighted contrast.

By changing the pulse sequence or the imaging variables TE and TR, different degrees of soft tissue contrast can be obtained. MRI contrast agents provide increased contrast by acting on T1, T2, and T2*. Paramagnetic agents—gadolinium and manganese

The most widely used clinical MRI contrast agents are based on Gd(III). A powerful approach is shown in Fig. 77. These neutral or anionic complexes (as the meglumine or sodium salt) are highly soluble and formulated at 0.25–1.0 M. They consist of a Gd(III) ion sequestered by an octadentate ligand with a water coligand. Gd(III) has seven unpaired electrons and a symmetric 5/2, 6S ground state (M(nll) = 5/2, 5S state). The high spin number and relatively slow electronic relaxation make these complexes potent nuclear magnetic resonance (NMR) contrast agents. The octadentate ligand is required for high thermodynamic stability to prevent the potentially toxic Gd(III) ion from being released in vivo. The ligand also directs the assembly of the complex (metal-essential). The coordinated water is critical for transmitting the relaxation effect. The water ligand is in fast exchange (10−10−12 s−1 at 37°C) with bulk water. Relaxation of this water coupled with fast exchange results in an overall shortening of the bulk water relaxation time, which affects image contrast.

MRI agents are characterized in terms of relaxivity. Relaxivity is defined as the extent to which the contrast agent can change the relaxation rate of solvent water protons normalized to the concentration of the metal ion M (Equation 1)

\[ r_1 = \frac{1}{T_1} - \frac{1}{T'_1} \left[ \frac{M}{[M]} \right] \]  

where T1 is the relaxation time in the absence of contrast agent. Increased relaxivity enables detection at lower concentrations of agents. Gadolinium and manganese complexes shorten both T1 and T2. However, because tissue T1 is usually much longer than T2, these contrast agents affect T1 on a much greater percentage basis and are referred to as T1 agents. T2 agents provide positive (bright signal) contrast.

Some relaxivities of Gd, Mn, and Fe agents are listed in Table 77. For discrete chelates, differences in relaxivity occur mainly because of differences in inner-sphere hydration number (q), in molecular tumbling, and in the rate of water exchange from the inner coordination sphere. Small paramagnetic molecules tumble in solution too fast to frequencies generating a fluctuating magnetic field that can induce nuclear relaxation. However, for imaging studies done at 65 MHz, relaxation is not as efficient as it could be. Compare the relaxivities of Gd(DTPA) and Mn (20 or 64 kHz) because when Mn is bound to albumin it tumbles much more slowly and relaxation is more efficient. Gd(DTPA) does not bind to proteins and has a lower relaxivity. At higher fields, the frequency match of tumbling with Larmor frequency decreases for protein-bound Mn (30 kHz) and relaxivity drops as well. Strategies for slowing molecular tumbling through protein binding, through increased molecular size, or through increased rigidity have proved useful for increasing relaxivity (6).

For fast-tumbling complexes, it is fast motion and not slow water exchange that limits relaxivity. In Table 77, relaxivities for the first four compounds are very similar because they have similar size and molecular weight and tumble in solution at about the same rate, even though they have water exchange rates differing by a factor of 10 (Gd(III) vs. Gd(III)–BMA). When tumbling is slowed, water exchange can be limiting. The water exchange rate also can be adjusted by changing the donor atoms or chelate geometry (7). Amide oxygen donors generally decrease water exchange relative to the carboxylate analog, whereas phosphate donors will increase it. Slow water exchange can limit relaxivity, but it only becomes pronounced when molecular tumbling is slowed.

The hydration number, q, also can be increased, but it may make the complex less stable to dilution (10). Raymond and coworkers (8, 9) have described a series of 2 or 3 complexes with fast water exchange that are resistant to transmetallation and/or anion binding. Increasing q also should be coupled with slowed tumbling to achieve very high relaxivity. Alme et al. (10) recently have demonstrated this with a novel 3–2 complex bound to albumin.

A more common approach is to assemble multiple chelates either covalently (11) (oligomer, polymer, and dendrimer) or non-covalently (12, 13) (micelle, liposome, and emulsion). These approaches all yield higher molecular relaxivities because of the assembly, but the per-ion relaxivity also is increased because motion is slowed. Fast internal motions can limit these relaxivity gains, but this limitation can be overcome by rigidifying the structure in some way (14). The iron oxide nanoparticles (MION) is used. USPIO and SPION are superparamagnetic, which means that they become magnetic in the presence of an applied field but revert to being nonmagnetic.
when taken out of the field. USPIO have a single Fe₃O₄ core about 4–5 nm in diameter, and it is coated to make the particle biocompatible. SPIO have more than one crystal of iron oxide and are larger than USPIO but still submicron. The small size allows SPIO and USPIO to be formulated as suspensions and administered intravenously. Large particles have been used for GI tract imaging after oral administration.

No inner space water molecules in iron particles exist, and the relaxation of water derives from the water molecules diffusing near the particle. Some generalities exist about relaxivity in these particles. For the USPIO, longitudinal relaxivity \( r_1 \) can be quite high, and these can function as effective \( T_1 \) agents, especially at low fields. The \( r_{1\beta} \) ratio for USPIO is significantly larger than for gadolinium complexes, and \( r_{2\beta} \) increases with increasing magnetic field. When an aggregation of crystals exists, which is the case in SPIO, \( r_1 \) decreases while \( r_2 \) increases. Thus, both for the particles themselves as well as for aggregates of particles, the ratio of \( r_{2\beta} \) typically increases as the size of the particles or aggregates increases, although the \( T_2 \)-relaxivity as a function of particle size can be quite complicated.

These agents are used mainly as \( T_2 \)-agents where they provide negative image contrast (destroy signal) wherever they localize. The mechanism is predominantly through altering the local magnetic susceptibility. \( T_1^* \). Gadolinium agents need to interact directly with water to cause relaxation. The \( T_2^* \) effect of iron oxides is a through space effect and means that even if the agent is compartmentalized in a small space, it still can provide significant contrast. A good example of this effect is in brain perfusion imaging. Even though the blood vessels represent only about 3% of the brain, the \( T_2^* \) effect of the contrast agent extends well beyond the vessels and yields robust contrast. Lanthanides have been coated with different biocompatible reagents, such as dextran and citrate. The coating and particle size both impact biodistribution. Recent efforts (16, 17) have focused on using chemically functionalizable coatings.

**Clinical applications**

The first generation of gadolinium-based agents are tracers of extracellular fluid (ECF) agents. The various ECF agents are almost indistinguishable from an efficacy perspective and, as a result, often are referred to by physicians as “gadog” or “gadolinium.” They are used to detect or rule out lesions in the brain because of the disrupted blood-brain barrier in the presence of tumor. They are used to generate blood vessel images immediately when injection (dynamic first-pass angiography) occurs as well as for assessment of cerebral, myocardial, and renal perfusion. Although nonspecific, the ECF agents are used widely in clinical practice and are by far the most commonly used MRI contrast agents.

The ECF agents are eliminated almost exclusively via the kidneys. The next generation of contrast agents were designed for liver imaging (18). Three different approaches are seen in approved agents. SPIO such as AMI-25 (ferumoxides, Feridex, Bayer Healthcare, USA and Endorem, Guerbet, France) and SHU-555A (ferucarbotran and Resovist, Bayer-Schering, Germany) are recognized by the reticuloendothelial system and transported to Kupffer cells in the liver. The SPIO make the normal-functioning liver appear dark. Tumors and fibrotic lesions do not take up the particles and appear bright on a \( T_2 \)-weighted image (Fig. ??). The Gd-based agents Gd–EOB–DTPA and Gd–BOPTA have mixed renal and hepatobiliary clearance. The complexes are taken up by hepatocytes and transported through the liver and gall bladder. Here, the healthy liver appears bright on a \( T_1 \)-weighted image, while lesions are not enhanced. The manganese compound Mn–DPDP (mangafodipir and Teslascan) is unstable with respect to transmetalation (19). The complex partially dissociates in plasma. Both the complex and free Mn²⁺ are taken up rapidly by the liver. In hepatocytes, the manganese ion binds to cellular proteins, which results in increased relaxivity. Like the Gd-based liver agents, the normal liver appears bright on a \( T_1 \)-weighted scan.

Although ECF agents are used for blood vessel imaging, the imaging must be done immediately after injection (so-called dynamic MRA). The blood concentration drops quickly with time as the agent both distributes outside the vessels into the interstitial space and is cleared via the kidney. There was a push to develop “blood pool” agents (20) that are present only in the vascular space to maximize blood-tissue contrast and that provide persistent contrast in the blood such that slightly longer, higher resolution scans can be used (the rapid imaging required with ECF agents sacrifices spatial resolution). Three approaches have been employed resulting in compounds that have progressed to the clinic or to clinical trials:

1. **Serum albumin targeting:** MS-325 was designed to bind reversibly to serum albumin. Albumin binding serves to localize the agent in the vascular space; it increases the relaxivity of the protein-bound complex, and the unbound fraction is available for renal clearance. MS-325 is approved for use in the EU and other countries. The bile acid–GdDTPA conjugate gadodocoletic acid (a.k.a. B22956, Fig. ??) being developed by Bracco functions similarly; Phase II clinical trials have demonstrated its efficacy.

2. **Increased size, Fig. ??:** A different approach was taken with the compound Gadomer (also called Gadomer-17, Bayer Schering Pharma). Gadomer is a dendrimer with 24 gadolinium DO3A-monomamide chelates on the periphery of the dendrimer. Its size is large enough to restrict the dendrimer to the vascular compartment and to increase relaxivity relative to ECF agents, but it is still small enough to be renally eliminated. A related approach was taken by Guerbet with the compound P792 (gadomelitol, Vistarem, Guerbet, France). P792 is retained in the vascular space because of its size; the relaxivity is high because the gadolinium sits at the center of the molecule, which gives it a long correlation time.

3. **USPIO:** The ultrasmall iron oxide particles have good \( T_2 \) relaxation properties, especially at lower fields (1.5 T). Their small size and coating enables them to evade the RES and remain in the blood pool; they are too large to extravasate. The dextran-coated formulation NC100150 was evaluated in multicenter trials and found to be effective, but ultimately, it...
was abandoned. The citrate-coated V50P particle has entered Phase II trials.

The USPIO AMI-227 (ferumoxtran, Combidex, Advanced Magnetics, USA or Sinerem, Guerbet, France) has undergone clinical trials for lymph node imaging. In addition to being useful for angiography, these particles are trafficked slowly to the lymph nodes. A day after injection, normal-functioning lymph nodes appear dark on a T2-weighted image. If a tumor is present in the node, then it appears brighter because the cancer cells do not take up the iron particle. These particles also are taken up by macrophages and may prove useful as a marker of inflammatory response.

EP-2104R is a peptide-based agent that targets fibrin. Fibrin is an abundant component of blood clots (thrombi). The fibrin-specific peptide is conjugated to several gadolinium chelates to create a high-relaxivity, thrombus-specific contrast agent that worked well in animal models (21) of coronary, cardiac, and pulmonary thrombosis. Figure 77 shows an EP-2104R-enhanced image of a right atrial thrombus in a porcine model and highlights the difference between positive, T1-weighted contrast agents and negative, T2-weighted contrast agents (e.g., Fig. ??). Recently, this compound has entered clinical trials.

Newer agents

A part from EP-2104R, the compounds described above rely on distribution and passive targeting for their imaging efficacy. Newer contrast agents are targeted actively. Some examples that have been studied in animal models are described below and shown in cartoon form in Fig. ??.

The Westerdal group (16, 17) have developed a platform technology based on iron oxide particles called cross-linked iron oxides (CLIO). These particles have amine groups on the surface that can be used to attach multiple targeting vectors as well as fluorochromes for multimodality (MRI, near IR) imaging. Examples from this group include particles targeted to vascular cell adhesion molecule (VCAM) or for atherosclerotic plaque detection, particles targeted to E-selectin in cancer, particulates targeted to mucin-1 expression (MUC-1) in cancer, and derivatized particles with annexin V for imaging apoptosis. Iron oxides usually provide higher sensitivity than gadolinium-based agents and, as mentioned above, extend their contrast-enhancing effect through space.

The Wickline/Lanza group (22) uses a perfluorocarbon emulsion as a contrast agent platform. Gadolinium chelates with lipophilic tails can incorporate into the emulsion such that thousands of Gd(III) ions per particle exist. Targeting is achieved by derivatizing a targeting vector (antibody, small molecule, peptide, etc.) with a lipid chain to incorporate into the emulsion. They have demonstrated targeting to thrombus and also to angiogenesis by targeting the integrin αvβ3.

Considerable effort has been expended for imaging atherosclerosis using a variety of approaches (12, 23), in addition to those described above: gadolinium chelates noncovalently associated with high density lipoprotein (HDL); immunomicelles containing gadolinium chelates targeted to the macrophage scavenger receptor, and micelles formed from gadolinium chelates with perfluorocarbon chains.

Another approach has been to make the contrast agent responsive to its environment. The molecular parameters affecting relaxivity (hydration, tumbling time, and water exchange) can be altered by an enzymatic transformation, by pH change, or when binding an analyte (e.g., Ca or glucose). For instance, Meade and coworkers (24) developed a gadolinium complex that changed its relaxivity in the presence of the enzyme beta-galactosidase (Fig. ??). The complex has a galactose residue conjugated to it that blocks the access of water into the inner coordination sphere of the gadolinium. When the sugar is cleaved by the enzyme, relaxivity is enhanced and can be detected by MRI. Querol and Bogdanov (25) recently reviewed these amplification strategies.

Radiopharmaceuticals

A radiopharmaceutical is a pharmaceutical with a radioactive element incorporated in it. Radiopharmaceuticals that emit gamma rays are used for imaging, whereas those that emit alpha particles, beta particles, or Auger electrons are used for therapeutic purposes, mainly in cancer. Because most of the periodic table consists of metals, it is not surprising that many radionuclides with useful imaging or therapy properties are metallic. Therapeutic radiopharmaceuticals are outside the scope of this article but will be discussed where imaging is used to guide the therapy. The nuclear imaging agents described will be split into gamma emitters, such as 99mTc, and positron emitters, such as 64Cu.

Gamma ray (γ-ray) emitters

The most widely used metal-based imaging agent is based on the metastable isotope technetium-99m. The hexakis(sodium) 99mTc haloferrate(V) complex, Tc-99mHMPAO (Cardiolite, sestamibi) (Fig. ??), is used routinely for myocardial perfusion imaging to evaluate coronary artery disease. Imaging is done either in two dimensions or in three. Planar imaging with a gamma camera is a 2-D projection of the gamma rays emitted from a 3-D source (the patient) and is analogous to planar X-ray imaging. SPECT generates 3-D images by using a gamma camera to acquire multiple 2-D images from different angles and by reconstructing the image.

For gamma imaging, the ideal agent should have a half-life for decay on the order of the examination, and it should be eliminated from the body with a half-life on the order of the examination. This minimizes the radiation exposure to the patient. The isotope should be readily available for widespread use and should only emit γ-rays. The emission energy should be high enough (>50 keV) to penetrate tissue but low enough (<30 keV) that it is efficiently detectable by the camera. The agent should be easily prepared and on administration should be rapidly taken up by the area of interest with quick background clearance. 99mTc is the best isotope for γ-imaging based on its physical properties and availability; 64Cu is the...
using a Tc-tagged bombesin peptide fragment.

A generator provides a readily available on-site source of 99mTc. A complex is used to label white blood cells for infection imaging. Illustrates tumor-targeted planar γ-emission imaging in a mouse model. An active research effort into Tc-tagged imaging agents is underway (26-29). For tissue penetration and detection, 141 keV. 99mTc is a decay product of 99Mo (t1/2 = 66h). The technetium generator consists of reactor-produced 99Mo that is immobilized on an anion exchange column encased in lead. Each day pertechnetate, 99mTcO4-, is produced, which can be eluted off the column. The generator provides a readily available on-site source of 99mTc to hospital radiopharmacies.

Applications of nuclear medicine imaging agents

The most common nuclear medicine procedure is myocardial perfusion imaging. Here, the imaging agent is administered while the patient is exercising (treadmill or bicycle) or given a drug to increase blood flow. The imaging agent is taken up rapidly by the heart muscle and retained there. Imaging is performed later, and the image reflects blood flow during exercise. Areas of reduced flow are indicative of coronary artery disease. Originally, the tracer used was the thallous ion as the chloride salt, 201Tl. 201Tl+ enters myocytes through potassium ion channels. Thallium still is used, but the majority of exams are done with Tc agents, the most common being 99mTc-sestamibi and 99mTc-tetrofosmin (Moyvane, GE Healthcare, USA) (Fig. 7). Figure 7 also shows approved brain and kidney perfusion agents. Bone imaging is done routinely to search for metastases. Bone-seeking agents are structurally ill-defined 99mTc complexes of pyrophosphate (PYP, also useful in imaging myocardial infarction), methylenediphosphonate (MDP), or hydroxymethylenediphosphonate (HDP). The 99mTc complex of DTPA is used in perfusion imaging. 99mTc also is co-ordinated directly to micron-sized aggregated particles of serum albumin for lung perfusion studies. 111In forms a lipophilic insoluble complex with β-hydroxyurea (linoleic acid), and this complex is used to label white blood cells for infection imaging. Ga either as the citrate or nitrate behaves as a mimic of Fe2+ and has been used in imaging inflammation and tumors.

Initially, imaging agents were discrete metal-essential complexes. Increasingly, the metal is complexed by a bifunctional (prelabeling technique). This process avoids exposing a potentially sensitive moiety to harsh reduction/chelation conditions and may be useful for proof-of-concept studies. However, for clinical use, usually it is preferred to conjugate the ligand first to the targeting group and to react this purified product with the Tc(V) complex and then couple to the targeting group (postlabeling technique). This process avoids exposing a potentially sensitive moiety to harsh reduction/chelation conditions and may be useful for proof-of-concept studies. However, for clinical use, usually it is preferred to conjugate the ligand first to the targeting group and to react this purified product with technetium for ease of use in the hospital setting.

Another widely used approach to targeting 99mTc is through the hydrazinonicotinic (HY-NIC) technique (28). This approach

Receptor-targeted radiotherapy is a growing field. Imaging plays a key role in the efficacy of these therapeutic radiopharmaceuticals (30). Not all tumors express a given receptor, and in those that do, receptor density can vary in time and as a response to treatment. Imaging tumor uptake with a diagnostic analog can select patients for therapy and be used to monitor treatment. Imaging also can help with dosimetry because it is critical to deliver enough radiation to kill the tumor while minimizing damage to other organs or nearby bone marrow. 90Y is a pure γ-emitter used in therapy. The chemistry of Y4+ and in3+ is similar enough that 111In often is incorporated for imaging studies. 177Lu is another isotope used in therapy, and it also emits a gamma with suitable energy for imaging.

Bioconjugate approaches for Tc and In labeling

The bioconjugate chemistry of technetium is more specialized (26-28) because of how the isotope is produced and because of the rich redox chemistry of technetium (oxidation states from −1 to +7). Pertechnetate eluted from the generator sometimes must be reduced and chelated. For clinical applications, a kit can be used to stabilize a given oxidation state (Fig. 7) gives a flavor. Many permutations exist of the co-ligands used, reducing agents, and methods for conjugation to the targeting vector; these are the subject of excellent review articles (26-29). For conciseness, the three most common bifunctional approaches are described.

The oxotechnetium core, [Tc=O]3+, in Tc(V) is readily accessible from stannous chloride reduction of pertechnetate. The Tc(V) oxidation state can be stabilized by thiolate, amide N, and amine N donors. As a result, many examples exist of 𝑁3,𝑁,𝑁 ligands (see, e.g., Figs. 7 and 7) that form the base of square pyramidal complexes with Tc(V) where the oxo group is the pyramid apex. As Tc(V) is capable of deprotonating an amine N-H and coordinating to the amide nitrogen, amino acids may make useful building blocks for N,N,N,N-ligands. The MAG3 ligand in Fig. 7 is assembled from mercaptoacetate coupled to three glycines. The resultant [Tc(O)(MAG3)] complex has a pendant carboxylate group that can be activated and coupled to the amine of a targeting vector. One can appreciate that ligands assembled from other amino acids would generate functionalizable side chains for coupled to the Tc(V)-deprotonate shown in Fig. 7. In principle, one can form the Tc(V) complex and then couple to the targeting group (prelabeling technique). This process avoids exposing a potentially sensitive moiety to harsh reduction/chelation conditions and may be useful for proof-of-concept studies. However, for clinical use, usually it is preferred to conjugate the ligand first to the targeting group and to react this purified product with technetium for ease of use in the hospital setting.

Another widely used approach to targeting 99mTc is through the hydrazinonicotinic (HY-NIC) technique (28). This approach

Metallic Medicine, Imaging Agents

The most common nuclear medicine procedure is myocardial perfusion imaging. Here, the imaging agent is administered while the patient is exercising (treadmill or bicycle) or given a drug to increase blood flow. The imaging agent is taken up rapidly by the heart muscle and retained there. Imaging is performed later, and the image reflects blood flow during exercise. Areas of reduced flow are indicative of coronary artery disease. Originally, the tracer used was the thallous ion as the chloride salt, 201Tl, which enters myocytes through potassium ion channels. Thallium still is used, but the majority of exams are done with Tc agents, the most common being 99mTc-sestamibi and 99mTc-tetrofosmin (Moyvane, GE Healthcare, USA) (Fig. 7). Figure 7 also shows approved brain and kidney perfusion agents. Bone imaging is done routinely to search for metastases. Bone-seeking agents are structurally ill-defined 99mTc complexes of pyrophosphate (PYP, also useful in imaging myocardial infarction), methylenediphosphonate (MDP), or hydroxymethylenediphosphonate (HDP). The 99mTc complex of DTPA is used in perfusion imaging. 99mTc also is co-ordinated directly to micron-sized aggregated particles of serum albumin for lung perfusion studies. 111In forms a lipophilic insoluble complex with β-hydroxyurea (linoleic acid), and this complex is used to label white blood cells for infection imaging. Ga either as the citrate or nitrate behaves as a mimic of Fe2+ and has been used in imaging inflammation and tumors.

Initially, imaging agents were discrete metal-essential complexes. Increasingly, the metal is complexed by a bifunctional chelator that is tagged to a small molecule, peptide, or antibody-targeting group. Clinical examples include 99mTc-Acibromab, a Fab’ monoclonal antibody fragment directed toward the carcinoembryonic antigen (CEA) expressed on most colorectal cancers; 111In-DTPA labeled capromab antibody for imaging of prostate cancer and prostate metastases; 111In-DTPA conjugated to a synthetic peptide analog of somatostatin (111In-pentetreotide, Octreoscan, Covidien, USA) is used for the detection of cancers overexpressing the somatostatin receptor; 99mTc-P829 (NeoTect, Bayer Healthcare, USA), is a peptide-based analog of somatostatin with the 99mTc chelator built into the peptide (Fig 7). An active research effort into Tc-tagged imaging agents is underway (26-29); Figure 7 illustrates tumor-targeted planar γ-imaging in a mouse model using a Tc-tagged bombesin peptide fragment.

Receptor-targeted radiotherapy is a growing field. Imaging plays a key role in the efficacy of these therapeutic radiopharmaceuticals (30). Not all tumors express a given receptor, and in those that do, receptor density can vary in time and as a response to treatment. Imaging tumor uptake with a diagnostic analog can select patients for therapy and be used to monitor treatment. Imaging also can help with dosimetry because it is critical to deliver enough radiation to kill the tumor while minimizing damage to other organs or nearby bone marrow. 90Y is a pure γ-emitter used in therapy. The chemistry of Y4+ and in3+ is similar enough that 111In often is incorporated for imaging studies. 177Lu is another isotope used in therapy, and it also emits a gamma with suitable energy for imaging.

Bioconjugate approaches for Tc and In labeling

The bioconjugate chemistry of technetium is more specialized (26-28) because of how the isotope is produced and because of the rich redox chemistry of technetium (oxidation states from −1 to +7). Pertechnetate eluted from the generator sometimes must be reduced and chelated. For clinical applications, a kit can be used to stabilize a given oxidation state (Fig. 7) gives a flavor. Many permutations exist of the co-ligands used, reducing agents, and methods for conjugation to the targeting vector; these are the subject of excellent review articles (26-29). For conciseness, the three most common bifunctional approaches are described.

The oxotechnetium core, [Tc=O]3+, in Tc(V) is readily accessible from stannous chloride reduction of pertechnetate. The Tc(V) oxidation state can be stabilized by thiolate, amide N, and amine N donors. As a result, many examples exist of 𝑁3,𝑁,𝑁 ligands (see, e.g., Figs. 7 and 7) that form the base of square pyramidal complexes with Tc(V) where the oxo group is the pyramid apex. As Tc(V) is capable of deprotonating an amine N-H and coordinating to the amide nitrogen, amino acids may make useful building blocks for N,N,N,N-ligands. The MAG3 ligand in Fig. 7 is assembled from mercaptoacetate coupled to three glycines. The resultant [Tc(O)(MAG3)] complex has a pendant carboxylate group that can be activated and coupled to the amine of a targeting vector. One can appreciate that ligands assembled from other amino acids would generate functionalizable side chains for coupled to the Tc(V)-deprotonate shown in Fig. 7. In principle, one can form the Tc(V) complex and then couple to the targeting group (prelabeling technique). This process avoids exposing a potentially sensitive moiety to harsh reduction/chelation conditions and may be useful for proof-of-concept studies. However, for clinical use, usually it is preferred to conjugate the ligand first to the targeting group and to react this purified product with technetium for ease of use in the hospital setting.

Another widely used approach to targeting 99mTc is through the hydrazinonicotinic (HY-NIC) technique (28). This approach
Involves reacting an activated ester of HYNIC acid with an amine group of the targeting moiety (peptide or protein) and purifying it. A Tc(V) oxo complex is prepared by reduction of TcO$_4$$_{-}^-$. In the presence of a coligand such as ticine, mannitol, or glucophosphate. Sometimes triphenylsulfonic acid/phosphine is used as a reducing agent and coligand. The HYNIC derivative then is reacted with this Tc(V) complex and the hydrazine group coordinates to the Tc$_3$ oxidation state. The complex is water-soluble and, subsequently, tridentate ligands have been employed. The single amino acid chelator found to be less stable in vivo and, subsequently, tridentate ligands have been employed. The single amino acid chelator (SAAC) approach (34) involves a protected unnatural amino acid that contains a tridentate chelator as shown in Fig. ?? The SAAC can be incorporated into the targeting peptide at any point using solid-phase peptide synthesis and represents a small perturbation to the peptide compared with the N-terminus of the peptide or coligand. The HYNIC derivative then is reacted with this conjugate (35). Often a DOTA derivative is used to chelate the Cu$^{+}$ ions and the hexadentate NOTA ligand ($\text{Cu}^{+}$)$_2$ has been reported that involve bifunctional 64Cu compounds that cause the complex to fall apart, and the copper is trapped inside the cell. The redox potential of Cu$^{+}$ATSM is lower than that of Cu$^{+}$TSM, and this redox potential is believed to be the root of the selectivity of Cu$^{+}$ATSM for hypoxic tissue. Copper complexes also are used as bifunctional chelators (37), however, transmetallation remains an issue. The Cu$^{+}$DOTA complex is known to lose copper in vivo. Increasing the macrocycle size by two carbons (cyclocyclam) to give TETA results in a more stable/inert complex. Cross-bridged macrocycles also have been used to give more inert complexes. However, it is important to have facile incorporation of the copper into the chelate, and some ligands giving more stable complexes do not exhibit efficient/mild labeling kinetics. The coordination chemistry here still is evolving (37). Several in vivo studies have been reported that involve bifunctional $^{64}$Cu compounds that are targeted, for example, to vascular endothelial growth factor (VEGF) receptor (38), to malignant melanoma (39), or to the $\alpha$V$\beta$3 integrin (40) for tumor detection or for monitoring cancer therapy.

Positron emission tomography (PET)

Radioisotopes that decay by positron emission also are detectable with a gamma camera. The positron undergoes an annihilation reaction with an electron in tissue resulting in two gamma rays (511 keV) being emitted at 180° apart. PET scanning involves acquiring a series of 2-D images at different orientations and reconstructing them to give a 3-D image. PET imaging has the benefit of higher resolution than SPECT and also offers the potential for getting quantitative data. This quantitative aspect has been exploited in obtaining absolute measures of blood perfusion as well as in determining receptor occupancy numbers. A resurgence of PET has occurred in recent years mainly because of the success of $^{18}$F-fluorodeoxyglucose (FDG) in oncology and the development of the combined PET/CT scanner. PET/CT provides a high resolution CT anatomical image onto which the lower resolution molecular image of the PET tracer is superimposed.

The majority of PET research involves $^{18}$F ($t_{1/2} = 110$ min) and $^{11}$C ($t_{1/2} = 20$ min). With considerable effort, these isotopes can be incorporated into organic molecules such as hormones or drugs. The benefit is that the tracer does not alter the pharmacokinetics of the molecule being labeled in the same way that the conjugation of a metal complex will. Nevertheless, applications of metallic PET tracers exist. Rubidium-82 ($t_{1/2} = 1.27$ min) has a very short half-life but is easily obtained from a strontium-82 generator. The rubidium ion is a potassium analog, and $^{82}$Rb$^+$ is used in myocardial perfusion studies.

Although useful, rubidium is limited in half-life and chemistry. The other more studied PET metal isotopes are $^{68}$Ga ($t_{1/2} = 68$ minutes) and $^{64}$Cu ($t_{1/2} = 12.7$ hours). $^{68}$Ga can be produced on site by a generator. $^{64}$Cu production requires a cyclotron, but its relatively long half-life allows it to be shipped from a central source. Most in vivo work with $^{68}$Ga has involved targeting the somatostatin receptor with a peptide–gallium conjugate (33). Often a DOTA derivative is used to chelate the $^{68}$Ga, but the hexadentate NOTA ligand (Fig. ??) also offers high stability.

The bisomocarbazol bismethoxytrifluorophosphine $^{64}$CuTFSM and $^{64}$CuU$^{+}$TSM, Fig. ??, have been used to image perfusion and hypoxia, respectively (36). It is believed that these compounds freely diffuse into cells. Reduction from Cu$^{(II)}$ to Cu$^{(I)}$ inside the cell causes the complex to fall apart, and the copper is trapped inside the cell. The redox potential of Cu$^{(II)}$TFSM is lower than that of Cu$^{(II)}$TSM, and this redox potential is believed to be the root of the selectivity of Cu$^{(II)}$TSM for hypoxic tissue. Copper complexes also are used as bifunctional chelators (37), however, transmetalation remains an issue. The Cu$^{(II)}$DOTA complex is known to lose copper in vivo. Increasing the macrocycle size by two carbons (cyclocyclam) to give TETA results in a more stable/inert complex. Cross-bridged macrocycles also have been used to give more inert complexes. However, it is important to have facile incorporation of the copper into the chelate, and some ligands giving more stable complexes do not exhibit efficient/mild labeling kinetics. The coordination chemistry here still is evolving (37). Several in vivo studies have been reported that involve bifunctional $^{64}$Cu compounds that are targeted, for example, to vascular endothelial growth factor (VEGF) receptor (38), to malignant melanoma (39), or to the $\alpha$V$\beta$3 integrin (40) for tumor detection or for monitoring cancer therapy.
Conclusions

Metal-based imaging agents comprise a broad and growing field with compounds ranging from simple aqua ions to small molecules to supramolecular assemblies and nanoparticles. The imaging modality and chemical approach to the agent are best chosen based on the question that needs to be addressed by imaging. The imaging field continues to embrace advances in chemical biology to meet the challenges of in vivo imaging: the diagnosis and etiology of disease, the characterization of the disease at the molecular level, understanding the metabolic pathways, and monitoring the response to therapy.

References

Metals in Medicine, Imaging Agents


Further Reading

The thematic Chemical Reviews issue on metals in medicine from 1999 contains several thorough reviews on X-ray, MRI, and nuclear imaging probes. These reviews are cited here (Refs. 5, 6, 29, 36) and provide an excellent background on these topics and a good point from which to appreciate current literature.

Werner Krause has edited three recent volumes (221, 222, 252) in the Topics in Current Chemistry series that covers imaging agents in all modalities.


See Also

Imaging Techniques: Overview of Applications in Chemical Biology
Metals in Medicine: Therapeutic Agents
Natural Products as Anticancer Agents

David G. I. Kingston, Department of Chemistry, Virginia Polytechnic Institute and State University, Blacksburg, Virginia
David J. Newman, Natural Products Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Frederick, Maryland
doi: 10.1002/9780470048672.wecb369

Natural products have provided some of the most effective drugs for the treatment of cancer, including such well-known drugs as paclitaxel (Taxol™; Bristol-Myers Squibb) adriamycin, vinblastine, and vincristine. Natural products have also provided many compounds that have led to the discovery of new biochemical mechanisms. This review summarizes the major natural products in clinical use today and introduces several new ones on the cusp of entering clinical practice. The review is organized by mechanism of action, with compounds that interact with proteins discussed first, followed by compounds that interact with RNA or DNA.

Natural products were the original source of almost all the drugs used by mankind before 1900, and they continue to be a major source of new drugs and drug leads (1, 2). The reasons for the continued importance of natural products are not hard to discover. In the first place, a high correlation exists between the properties of drugs and those of natural products (3, 4). In addition, natural products usually have built-in chirality, and they are thus uniquely suited to bind to complex proteins and other biologic receptors. Finally, natural products have been enormously successful as drugs and drug leads, not only in the anticancer area but also in many other pharmaceutical areas (5). It is thus unsurprising that several authors have gone on record as advocating an increase in the drug discovery effort assigned to natural products (4, 6, 7).

This review summarizes the contributions of natural products to the discovery and development of anticancer agents. It includes information on many natural products and natural product analogs that are in clinical use as anticancer drugs, and it describes some natural product drugs in late-stage clinical trials. Because of space limitations, it is does not provide a comprehensive listing of all natural product and natural product-derived anticancer agents; readers interested in such a listing should consult a recent review (8).

Overview of Natural Products as Anticancer Agents

As of the time of writing, 178 drugs are approved worldwide for the treatment of cancer in all of its manifestations, and 172 of these are listed together with their classifications as to source in a recent review (8). Of the 178 approved antitumor agents, 25 (14%) are natural products, 48 (27%) are modified natural products, and 20 (11%) are synthetic compounds derived from a natural product pharmacophore. Natural products have thus led to 52% of the approved drugs against the collection of diseases that go by the collective name of cancer. Another 20 drugs (11%) are biologics, with the remainder being synthetic compounds. The highly significant contribution of natural products to anticancer drug discovery is clear from these figures.

The term "natural product" describes a broad class of anticancer agents, which range from complex compounds like paclitaxel and vinblastine to relatively simple compounds such as combretastatin-A4. A associated with these different structures are several different mechanisms of action, some of which were only discovered when the corresponding natural product was investigated. The following sections are divided on the basis of the mechanism of action of the drugs rather than their source, and consequently, any given section may include plant, microbial, or marine-derived agents. The broadest division is between those compounds that act by targeting proteins in some way and those compounds that act by direct interactions with DNA or RNA. Thus, these two broad areas provide the two major sections of this review.

Compounds that Target Proteins

The mammalian cell cycle is a complex and carefully regulated biologic process that leads to cell division, and faulty regulation...
of this cycle is one feature of most cancers. The cell cycle thus offers several targets for therapeutic intervention, and several of the proteins involved either directly or indirectly in controlling this cycle are the targets of some important anticancer agents. The most important targets, in terms of the number of drugs that target them, are the proteins tubulin, topoisomerase I, and topoisomerase II, but other proteins such as the checkpoint kinases chk1 and chk2 and the heat shock protein Hsp90 are also the targets of some drugs.

The protein tubulin is an interesting and important target. It exists in both α and β forms, and during a normal cell cycle, these two monomeric proteins polymerize into microtubules by noncovalent interaction; these microtubules are then involved in the reorganization of the chromosomes into the nuclei of the mother and daughter cells. After mitosis, the microtubules dissociate to α and β tubulin monomers. This process is dynamic, and even very small perturbations in assembly and/or disassembly of the monomers dimers may lead to cell death, either by inhibition of the assembly of tubulin monomers into microtubules or by inhibition of the dissociation of microtubules to α and β tubulin monomers. These two monomeric proteins polymerize into microtubules by interaction with the proteasome.

Compounds that inhibit tubulin assembly

The vinca alkaloids

The antitumor alkaloids vinblastine (1) and vincristine (2) were the first natural products to be used on a large scale as anticancer agents, and they thus blazed the trail for others that came afterward. Vinblastine (as vincaloloblastine) was isolated from Catharanthus roseus (L.) G. Don, which was formerly known as Vinca rosea L., by two independent teams during the 1950s (9, 10), and vincristine (as leurocristine) was isolated and structurally characterized by Svoboda (11) and Neuss et al. (10) in 1961 and early 1962, respectively. These alkaloids inhibit the polymerization of tubulin into microtubules. Vinblastine is used in combination with other agents for treatment of Hodgkin's disease and bladder and breast cancers, whereas vincristine is used for treatment of acute lymphocytic leukemias and lymphomas.

The semisynthetic analogs vindesine (3) and vinorelbine (4) have been developed more recently. Vindesine, which was first developed in the 1970s, is in clinical use; it seems to be more active than vincristine against non-small-cell lung cancer, but it also has a higher hematological toxicity than vincristine, so its utility is still being evaluated (12). Vinorelbine has been approved for treatment of non-small-cell lung cancer, and the fluorinated analog vinflunine (structure not shown) has entered clinical trials (13). For a recent general review of the vinca alkaloids and their analogs, see Reference 14.

Combretastatin

The first member of this class of compounds, (-)-combretastatin (5), was isolated from Combretum caffrum in 1982 (15). Subsequent studies led to the isolation of many additional combrestatins, including combretastatins A1 (6) and A2 (7), which are two of the most active members of this class, with potent activity as inhibitors of tubulin assembly. Additional development by the Petit group led to the design of combretastatin A4 phosphate (CA4-P) (8) as a promising drug candidate. This compound is actually a prodrug, which only functions as an inhibitor of tubulin assembly after hydrolysis of the phosphate group to yield the active compound (9). It was subsequently found to be an important member of a new class of compounds known as tumor-specific vascular targeting agents. CA4-P operates by binding to endothelial cell tubulin and causing changes in the morphology of the endothelial cells lining the microvessels feeding the tumors. This process causes disruption of the blood flow, which makes the microvessels unable to deliver oxygen to the tumor and leads ultimately to tumor necrosis. It received orphan drug approval by the Food and Drug Administration (FDA) in 2003 for thyroid cancer, and the FDA has approved it for a “fast track” Phase II clinical trial against anaplastic thyroid cancer (16).

Eribulin (E7389)

The natural product analog E7389 is not yet in clinical use, but it is in Phase III clinical trials and hopefully will enter
Natural Products as Anticancer Agents

Clinical use within the next few years. The natural product on which it is based, halichondrin B (9), is a member of a relatively large family of congeners with a polycyclic macrolide structure, which is reminiscent of ionophores. Many years of synthetic work by Kishi’s group at Harvard and a group at the Eisai Research Institute in the United States led to the selection of the two lead compounds E7389 (10) and E7390 (11) (17). These compounds were compared in both in vitro and in vivo assays at the National Cancer Institute (NCI) with the natural product obtained by deep-water dredging of a producing sponge. In an example of the skill of synthetic chemists when given what initially seemed to be an almost impossible task, E7389 was prepared in large quantity and entered clinical trials in 2001 as an inhibitor of tubulin assembly. It has a mechanism of action different from that of other tubulin interactive agents; it inhibits microtubule growth, but not shortening, and sequesters tubulin into aggregates (18). It showed good activity against refractory breast carcinoma in Phase II studies and entered Phase III studies for the same indication in late August 2006. Very recently, Hamel’s group at NCI have reported additional investigations on the mechanism of binding of both halichondrin B and E7389 to tubulin, which indicates that these agents may well form small, highly unstable aberrant tubulin polymers rather than the conventional massive stable structures found with vinca alkaloids and the antimitotic peptides (19).

Dolastatin

The dolastatins are a class of bioactive peptides isolated by the Pettit group from the Indian Ocean nudibranch, Dolabella auricularia. A total of 18 compounds were isolated over a 20-year period, with structures varying from relatively simple linear peptides to cyclic peptidolactones with nonpeptide components. Dolastatin 10 (12) was the most potent compound, with cytotoxicity in the subnanomolar range. It was shown to be a potent antimitotic agent, binding strongly to the β-subunit of tubulin (20), but it could not be isolated in sufficient quantity because of the scarcity of the source and of the low levels of secondary metabolites in the nudibranch. Pettit and colleagues thus devised many synthetic schemes, which led to the production of enough dolastatin 10 to go into human clinical trials as a tubulin interactive agent. Although it progressed to Phase II, it did not continue further because of a lack of activity and toxicity. However, the base structure led to the synthesis and biologic evaluation by various groups of a large number of related compounds, and synthadotin (also known as tasidotin, 13) emerged as a lead compound (21, 22). Synthadotin is an orally available synthetic derivative of dolastatin 15, and it is in Phase II trials as a tubulin interactive agent. Its discovery provides another example of the skill of synthetic chemists and the potential of novel natural products to be developed into drugs.

Compounds that promote tubulin assembly

Paclitaxel (Taxol)

Paclitaxel (Taxol™; Bristol Myers Squibb 14) and its semisynthetic analog docetaxel (15) are two of the most important anticancer agents of the last 25 years. Paclitaxel was isolated originally by Wall and Wani from Taxus brevifolia (23) and named taxol; this name was later trademarked by Bristol-Myers Squibb. Paclitaxel (Taxol™) is a water-soluble polyether that forms nanometer-sized aggregates in aqueous solution. These aggregates then bind to tubulin, which results in the formation of microtubules that are resistant to depolymerization. Paclitaxel is approved for the treatment of ovarian cancer, breast cancer, and non-small-cell lung cancer.

Docetaxel (Taxotere®)

Docetaxel (Taxotere®; Sanofi-Aventis 15) is a semisynthetic derivative of paclitaxel. It is approved for the treatment of breast cancer, non-small-cell lung cancer, and prostate cancer.

Other tubulin interactive agents

Other tubulin interactive agents include vincristine, vinblastine, and colchicine, which are all derived from the vinca alkaloids. These agents inhibit microtubule formation and cause the depolymerization of microtubules, leading to纺丝 segregations and cell death.

References

taxel on tubulin have been proposed, but recent REDOR NMR studies have established T-Taxol as the most probable conformation (36). The synthesis of the highly active bridged analog britaxel-5 (36), which is constrained to a T-Taxol conformation, confirmed this hypothesis (37).

Epothilones

The epothilones A-D were isolated from the myxobacterium Sorangium cellulosum as antifungal agents (38), but they were found subsequently to have the same mechanism of action as paclitaxel, which promotes the assembly of tubulin into microtubules (39). The epothilones are thus of great interest as potential antitumor agents because of their mechanism of action and because they are also active against some paclitaxel-resistant cell lines. At first glance, they would seem to have a very different shape from paclitaxel, but molecular modeling has shown that some significant common structural features exist in the two base molecules (40). Originally, the epothilones were difficult to obtain in large quantity, and a significant amount of work was performed in academia and industry to synthesize both epothilones A and B and their more active precursors, epothilones C and D (42). However, by using genetic manipulation, Frykman et al. (41) cloned and expressed the polyketide gene cluster that produces epothilones A and B. Subsequent removal of the terminal gene for the P450 enzyme and transfer to a different host enabled them to produce crystalline epothilone D from a large-scale fermentation.

The aza-analog of epothilone B (ixabepilone, 21), which is constrained to a T-Taxol conformation, is in Phase III trials. Epothilone D (KOS-862, 20) and ZK-EPO (23) are in Phase II clinical trials, and the synthetic analog fludelone (22) looks very promising in animal trials (43). However, an evaluation of the epothilones as a class has been limited to taxane-sensitive tumor types (prostate cancer and breast cancer) and does not seem to be distinctly different to the activity of taxanes. Epothilones should certainly not be considered as alternative taxanes, but
whether epothilones are here to stay or will fade away has yet to be determined." (44).

Compounds that inhibit Topoisomerase I

Camptothecin analogs

Camptothecin (24) was isolated from Camptotheca acuminata in 1966 by Wall et al. (45). It had potent anticancer activity in preliminary in vitro and animal assays, but its development was hampered by its extreme insolubility in water. It eventually entered clinical trials in the 1970s as the sodium salt of the carboxylic acid formed by opening the lactone ring, but this proved to have no efficacy and it was dropped from development. Interest in camptothecin was rekindled by the discovery that its primary cellular target was inhibition of topoisomerase I (46). Extensive medicinal chemical studies then led to the development of the two water-soluble derivatives topotecan (Hycamtin; GlaxoSmithKline, Brentford, Middlesex, United Kingdom, 25 and irinotecan (Camptosar; Pfizer, New York, NY, 26). The camptothecins are unique pharmacologically in having topoisomerase I as their only target, and in being able to penetrate mammalian cells readily, and several analogs are in clinical trials (47). Hycamptin and Camptosar are in clinical use for second-line treatment of metastatic ovarian cancer and small-cell lung cancer (topotecan) and for treatment of metastatic colorectal cancer in combination with 5FU/leucovorin (irinotecan) (47, 48).

Rebeccamycin

Rebeccamycin is an indolocarbazole; a comprehensive review of this class has appeared recently (49). Compounds related to rebeccamycin (27) are extremely interesting from a mechanistic standpoint, because relatively simple modifications of the indolocarbazole skeleton generate molecules with enhanced topoisomerase I activity. Thus, active agents can be made by modification of the rebeccamycin skeleton using fluorine substitution, which gives rise to BMS-250749 (28); this compound is headed for Phase I trials as a topoisomerase I inhibitor (47, 50). Second, modification of the base skeleton to include other heterocyclic and carbocyclic rings extends the compounds into previously unexplored chemical space. An example is demonstrated by asymmetric phenyl substitution, which produces compounds such as 29 with significant cytotoxic activity in cell lines, blocking at G2/M or S phase in the cell cycle (51).

Compounds that inhibit Topoisomerase II

Podophyllotoxins

Podophyllotoxin is a major constituent of the rhizome of the American May apple, Podophyllum peltatum. It was shown as early as 1947 to inhibit formation of the mitotic spindle, and its structure (30) was elucidated in 1951 (52). Podophyllotoxin is too toxic for use as an anticancer agent, but medicinal chemical studies led to the development of etoposide (31) and teniposide (32) as podophyllotoxin analogs. They differ chemically from podophyllotoxin in their stereochemistry and glycosylation at C4 as well as being demethylated at C4', but their most significant difference is in their mechanism of action. Unlike the parent compound, etoposide and teniposide act as inhibitors of topoisomerase II rather than as inhibitors of tubulin polymerization. Clinically etoposide (31) is used in combination with cisplatin against small-cell cancer, and it is also effective for the treatment of testicular cancer and non-small-cell lung cancer. Teniposide (32) is used in combination with cisplatin against neuroblastoma, with ara-C against acute lymphoblastic leukemia, and with carboplatin against small-cell
lack of water solubility, and the soluble compound Etopophos (Bristol Myers Squibb 33) was developed to circumvent this problem. It can be administered intravenously and is then rapidly converted to etoposide by plasma phosphatase. For a recent general review of the podophyllotoxins, see Reference 53.

Anthracyclines

From the perspective of the number of patients treated, one of the most important classes of topoisomerase II inhibitors is that of the anthracyclines, with daunorubicin (34) and its derivative doxorubicin (adriamycin) (35) being the best known of these agents currently in clinical use. Adriamycin is still a major component of the treatment regimen for breast cancer, despite its known cardiotoxicity (54). The mechanism of action of these molecules is now known to be inhibition of topoisomerase II (55), although they are also effective DNA binders (56). Both drugs are used for the treatment of acute non-lymphocytic leukemia, Hodgkin and non-Hodgkin lymphomas, and sarcomas, in addition to breast cancer. Derivatives of doxorubicin, such as epirubicin, idarubicin, pirirubicin, and valrubicin, have also been approved for clinical use, and the expansion of the efficacy of doxorubicin is being explored through targeted delivery techniques, including both liposomally encapsulated and monoclonal-linked derivatives.

Compounds that interact with other proteins

Geldanamycin

The first signal transduction modulator to enter clinical trials, other than a formal cyclin-dependent kinase or protein kinase C inhibitor, was the microbial product 17-allylamino-geldanamycin (17-AAG, 36). This modulator entered Phase I trials in 2001 and is currently in Phase II trials in a variety of cancers. Geldanamycin and its derivatives bind at the major ATP-binding site of the protein chaperone Hsp-90. The protein chaperones are emerging as attractive targets for cancer chemotherapy, and the reader is referred to three recent reviews for additional information (57–59).

Staurosporine

The indolocarboxazoles first came into prominence with the identification of staurosporine (37) and its simple derivative UCN-01 (38) as inhibitors of components of the eukaryotic cell cycle and of protein kinase C. Although these compounds are related structurally to rebeccamycin, they have very different mechanisms of action, in that they are highly potent but entirely nonselective inhibitors of protein kinases. UCN-01 has entered Phase I/II clinical trials against a variety of cancers, including leukemias, lymphomas, various solid tumors, melanoma, and small-cell lung cancer. Its clinical development
has been hampered by its high binding to human plasma proteins (60, 61).

Salinosporamide

The marine bacterial metabolite salinosporamide A (39) was isolated from the totally new genus Salinispora that mapped to the Micromonosporaceae, which are found in marine sediments across the tropics. It demonstrated activity as a cytotoxic proteasome inhibitor (62) similar to that observed for the structurally related compound omuralide (40) (63, 64), which resulted from a spontaneous rearrangement of the microbial metabolite lacystin in neutral aqueous media. Salinosporamide has been synthesized (65, 66) and has been fermented in saline media on a large scale under cGMP conditions. It entered Phase I clinical trials in May 2006.

Compounds that target DNA or RNA

The second major class of anticancer agents consists of compounds that act directly on DNA or RNA, either by intercalation, by alkylation, or by cleavage

Actinomycin D

The first microbial-derived agent in clinical use for cancer was actinomycin D (42) (which was systematically named as 9-D-actinomycin C1 and generically named daclomycin) that was introduced in the early 1960s. Despite extensive research into the preparation of analogs, no derivatives have progressed beyond clinical trials (67). Its mechanism of action is inhibition of DNA-dependent RNA synthesis, which in turn depends on the strong intercalation of actinomycin into double-helical DNA (68, 69). It is used clinically in the treatment of trophoblastic tumors in females, in metastatic carcinoma of the testis, and in Wilms’s tumor in children (67). In recent years there have been reports that actinomycin D may also act on the signal transduction cascade(s) at the level of transcription factor(s), and it will be interesting to see whether these activities rejuvenate interest in this class of molecules (70).

Bleomycins

A rather important class is the family of glycopeptolide antibiotics known as bleomycins (e.g., bleomycin A2 and Blenoxane; Nippon Kayaku Co., Ltd., Tokyo, Japan) (42); the bleomycins are structurally related to the plasmodiomics (71, 72). Bleomycin was originally thought to act through DNA cleavage, because it cleaves both DNA and RNA in an oxidative, sequence-selective, metal-dependent fashion in the presence of oxygen. Recent studies, however, suggest that an alternative mechanism of action may be inhibition of t-RNA from experiments reported recently by the Hecht group (73). Bleomycins are used clinically in combination therapy for the treatment of squamous cell carcinoma and malignant lymphomas.

Mitomycins

The mitomycins (mitosanes) were discovered in the late 1950s, and mitomycin C (43) was approved for clinical use in Japan in the 1960s and in the United States in 1974. Its serious bone marrow toxicity has led to extensive synthetic studies aimed at developing a less toxic analog but without significant success; it remains the only clinically used member of this class. It alkylates DNA only after undergoing a one-electron reduction. The current model postulates that mitomycin C alkylates and cross-links DNA by three competing pathways (74). Clinically it is used primarily in combination with other drugs for the treatment of gastric and pancreatic carcinomas (75).

Calicheamicin

Calicheamicin (44) is a member of a large group of antitumor emedine antibiotics. It was isolated from Micromonospora echinospora ssp. calichensis by workers at Lederle Laboratories (Pearl River, NY, now Wyeth) (76, 77); the structures of the related esperamicins were published simultaneously (78, 79). Calicheamicin is one of the most potent biologically active natural products ever discovered. It causes single-stranded and double-stranded DNA cleavage through a unique mechanism that involves reductive cleavage of the trisulfide “trigger” followed by Bergman cyclization to a diradical. It proved to be too potent and too toxic for direct clinical use, but it has been used as the “warhead” in the antibody-targeted chemotherapeutic agent Mylotarg (Wyeth Laboratories, Collegeville, PA), which was approved by the FDA in 2000 for clinical use for the treatment of acute myelogenous leukemia (80). Mylotarg is
Ecteinascidin

Currently, no approved antitumor drugs directly are derived from marine sources, but ecteinascidin (Yondelis; PharmaMar, Madrid, Spain) was submitted to the EMEA in early August 2006 for approval as an antisarcoma agent and was recommended for approval by the EMEA advisory committee in July 2007 (entered in proof) approved in September 2007 for Sarcoma. This compound was isolated originally from the Caribbean tunicate, Ecteinascidia turbinata (81, 82). The original supplies for preclinical studies came from a combination of wild harvesting and aquaculture both in sea and on land. The supply problem was finally overcome by the Spanish company, PharmaMar, which developed a 21-step semisynthetic route from the bacterial product cyanosafacin B (46) that could be carried out under cGMP conditions (83). This route provided an adequate source for advanced clinical trials and an assured supply if the drug is approved for clinical use.

Ecteinascidin 743 has a novel mechanism of action, binding to the DNA minor groove and alkylating the N2 position of guanine. This process strongly inhibits the transcription of specific genes. Ultimately it causes a p53-independent cell-cycle block, which leads to apoptosis. It has shown a clinical benefit rate close to 40% in Phase II studies on sarcomas (84).

Other agents

Mitomycin (47) is an antitumor antibiotic isolated from Streptomyces plicatus. It currently is used to a limited extent for the treatment of embryonal cell carcinoma of the testes and of cancer-related hypercalcemia (85). It is reported to be a specific inhibitor of the Sp1 transcription factor in hematopoietic cells (86).

Streptozotocin (48) is an N-nitroso urea isolated from Streptomyces achromogenes. It acts as a DNA-alkylating agent (87), and it is recommended for use in combination with doxorubicin (35) as the drug of choice for the chemotherapy for patients with malignant neuroendocrine pancreatic tumors (88).

Conclusions

As noted in the Introduction, natural products have served historically as the major source of drugs and lead compounds for the treatment of cancer, and the examples provided in this short review indicate that important discoveries in this area are still being made. Despite this impressive track record,
Natural Products as Anticancer Agents

Figure 13 Structures of ecteinascidin and cyanosafracin B.

Figure 14 Structures of mithramycin and streptozotocin.

many pharmaceutical companies have deemphasized natural product–based drug discovery efforts in favor of approaches such as combinatorial chemistry. Sadly de novo combinatorial chemistry, which was expected to be a panacea for the discovery of small-molecule drug leads over the last 15 or so years, has so far yielded only one drug for antitumor therapy. This drug is the orally active multikinase inhibitor, Sorafenib from Bayer AG (Leverkusen, Germany), which was approved in 2005 (89). The comments of Ortholand and Ganesan (4) are appropriate here.

The early years of combinatorial chemistry suffered from an excess of hype, and a major victim was natural-product screening. Many organizations went through an irreversible shift in policy, and prematurely discontinued their efforts in this area. We are now seeing the backlash from this knee-jerk reaction. The early combinatorial strategies were flawed and unproven, and have yet to deliver any blockbuster drugs. Meanwhile, we have lost the uniqueness of screening natural-product space as a complement to synthetic compounds. If past indicators are any guide, there are undoubtedly many more unique and potent biologically active natural products waiting to be discovered.

The data in this review support this statement and show clearly that natural products continue to provide both tools to probe biologic mechanisms and skeletons upon which to “improve” on the properties of the natural product. Scientists are still discovering or rediscovering the truth that natural products are the best lead structures from which to begin a search for novel mechanisms and novel treatments for a multitude of diseases, not just cancer.

References

11. Svoboda GH. Alkaloids of Vinca rosea

10. Neuss N, Cone NJ, Gorman M, Boaz HE. Vinca alkaloids. II.


Natural Products as Anticancer Agents


Natural Products as Anticancer Agents


Further Reading

See Also
Cancer, Chemical Biology of
Chemistry of Protein Reactive Natural Products
Natural Products in Marine Organisms, Chemical Diversity of
Natural Products in Microbes, Chemical Diversity of
Natural Products in Plants, Chemical Diversity of
Natural Products, Common Biological Targets for
Natural Products, Inhibition of Protein Biosynthesis by
Pharmaceutical Industry, Biocatalysts and Chemocatalysts

David J. Ager, DSM Pharmaceutical Products, Raleigh, North Carolina
Oliver May, DSM Pharmaceutical Products, Advanced Synthesis, Catalysis and Development, Geleen, The Netherlands
doi: 10.1002/9780470048672.wecb651

Catalytic reactions provide the opportunity to perform more environmentally friendly reactions. As the pharmaceutical industry produces a large amount of waste for a relatively small amount of drug product manufactured, the use of catalytic reactions is becoming more important. Catalysts can be biological or chemical in nature and can be used to effect a wide variety of transformations.

The pharmaceutical industry employs a wide variety of chemical transformations to prepare the active components of drugs. Cost and environmental pressures encourage the use of catalytic reactions for both bond-forming reactions and the creation of stereogenic centers. As the pharmaceutical industry generates a large amount of waste in the preparation of a relatively small amount of drug product, catalytic reactions will only increase in importance in this industrial sector (1). The E-factor, which is the amount of waste produced (Kg) to make a Kg of product, is high for the fine chemical and pharmaceutical industries. The use of catalytic methods, rather than stoichiometric ones, can help reduce waste (2). The development of a “green” process, however, has to be weighed against the speed of developing the process to the target molecule.

The purpose of this article is to provide an overview of the different types of chemical and biological catalysis currently available to the pharmaceutical industry in the process area. In other words, these transformations can be performed at scale. The types of catalysts that have been used are given together with systems that show potential for future application. The chemocatalytic area has addressed the synthesis of aromatic and heterocyclic compounds, which are common classes in pharmaceutically active compounds, whereas biocatalyst applications tend to be aimed toward the production of chiral molecules.

The uses of catalysts for asymmetric pharmaceutical synthesis have been reviewed by others (see the Further Reading section).

Types of Catalysts

Catalysts can be classified in many ways. A summary of the methods discussed in this article is given in Table 1. For enzyme catalysts, the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) have been followed (3). Enzymes are classified into six general groups, and the first digit of the enzyme commission number corresponds to the following general categories: 1) oxidoreductases, 2) transferases, 3) hydrolases, 4) lyases, 5) isomerases, and 6) ligases. The number of large-scale applications differs significantly among these enzyme types. Most commercial applications use hydrolases or oxidoreductases, which can be attributed to the broad range of enzymes available in these two classes (4).

Biological Catalysts

In a few cases, biocatalysts have the advantage that no chemo-catalytic alternative exists. It usually occurs when the exquisite stereoselection of a biocatalyst is used to distinguish between two equally reactive groups within a molecule based on stereoechemistry; the stereoselective oxidations of steroids and aromatic compounds are examples (5). A further instance in which biocatalysts are very powerful and no chemocatalyst equivalent is available is for glycosylation reactions and the stereoc-trolled synthesis of polysaccharides. In many areas, however, biological and chemical catalysts compete; examples of this competition include the reduction of ketones (vide infra) and the desymmetrization of cyclic anhydrides (6, 7). In these cases, the choice of which catalyst system to use will depend on accessibility and on process performance in such areas as selectivity, activity, and consumption, as well as cost. These parameters are highly product specific and often are difficult or impossible to predict. For the development of syntheses of new products, the fast screening of highly diverse libraries, be they biocatalytic
**Table 1** Catalysts useful for pharmaceutical applications

<table>
<thead>
<tr>
<th>Type</th>
<th>Class</th>
<th>Catalyst</th>
<th>Example of transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>Living whole cell Enzyme</td>
<td>Many within the cell</td>
<td>Preparation of secondary metabolite</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oxidoreductases</td>
<td>Oxidations and reductions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transferases</td>
<td>Methylation, glycosylation and amino group transfers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrolases</td>
<td>Ester hydrolysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lyases</td>
<td>C-C and C-N bond formations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Isomerases</td>
<td>Racemization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ligases</td>
<td>Coupling reactions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrogenations</td>
<td>A-kene reductions</td>
</tr>
<tr>
<td>Chemical</td>
<td>Transition metal</td>
<td>Asymmetric hydrogenations</td>
<td>Generation of new stereogenic center</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aryl coupling reactions</td>
<td>Preparation of biaryl compounds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coupling reactions</td>
<td>A-nilne preparation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Isomerizations</td>
<td>Chiral imines from alyl amine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metathesis</td>
<td>Ring formation</td>
</tr>
<tr>
<td>Organocatalysis</td>
<td>Carbon–carbon bond formation Oxidations</td>
<td></td>
<td>Aldol reactions, Epoxidations</td>
</tr>
</tbody>
</table>

or chemocatalytic, is, therefore, an important tool to determine the best choice of a catalytic system (8, 9). The use of molecular biological methodologies do allow for highly selective and efficient biocatalysts to be developed in a relatively short period of time (7). Without precedence, the development of a chemocatalyst is a long-term option.

**Enzymes**

Enzymes can be used in different formulations, immobilized or soluble, and with different degrees of purity, such as cell preparations and crude or enriched isolates. Isolation to a purified form takes time and effort and is usually avoided unless absolutely necessary. In many cases, molecular biology allows for an enzyme to be highly enriched (overproduced) in an organism, which reduces the need for purification (10). Such recombinant cells are, therefore, often used as cell preparations except if the cell needs to be treated to make the substrate accessible to the biocatalyst. More than 50 different enzyme subclasses are commercially available and can be used to prepare chiral molecules. A summary (11-14) of the most often used classes of enzymes that have been used in chemical synthesis is given in Table 2 (16-45). Reactions do not have to be performed in totally aqueous media as some enzymes can tolerate organic solvents (15). 

Enzymatic processes are now being applied to a wide range of pharmaceutical product syntheses (46). Examples are given for the preparation of cyanohydrins, which can then be used to prepare α-hydroxy acids and α-amino acids. Cyanohydrins are a very useful class of compounds as they can be transformed into a wide variety of compounds while retaining the stereogenic center (32, 35). Hydroxy nitrilases are available from natural sources (13), which can give access to either enantiomer of the product cyanohydrin (Fig. 1) (47).

**Table 2** Enzymes used in the preparation of pharmaceuticals

<table>
<thead>
<tr>
<th>Enzyme subclass</th>
<th>Substrate</th>
<th>Product</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racemases</td>
<td>α-Hydroxy acids</td>
<td>α-Hydroxy acids</td>
<td>16</td>
</tr>
<tr>
<td>Oxidases</td>
<td>A-Alcohols</td>
<td>A-Amino acids</td>
<td>16, 17</td>
</tr>
<tr>
<td>Dehydrogenases</td>
<td>Carbonyl compounds</td>
<td>Alcohol, hydroxy acids, amino acids</td>
<td>20, 21</td>
</tr>
<tr>
<td>Hydrolases</td>
<td>Esters, amides</td>
<td>Alcohol, carboxylic acids, alcohols</td>
<td>22-26</td>
</tr>
<tr>
<td>Aldolases</td>
<td>Carbonyl compounds</td>
<td>Hydroxy carbonyl compounds</td>
<td>27-30</td>
</tr>
<tr>
<td>Hydroxynitrilase</td>
<td>Carboxyl compounds</td>
<td>Cyanohydrins</td>
<td>27, 31, 32</td>
</tr>
<tr>
<td>Esterases</td>
<td>Esters</td>
<td>Alcohol, carboxylic acids</td>
<td>33, 34</td>
</tr>
<tr>
<td>Nitrilases</td>
<td>Nitriles</td>
<td>Carboxylic acids</td>
<td>35</td>
</tr>
<tr>
<td>N-Acetyl amino acid hydrolase</td>
<td>N-Acetyl amino acids</td>
<td>Amino acids</td>
<td>36, 37</td>
</tr>
<tr>
<td>Amidases</td>
<td>A-Amino acids</td>
<td>Amino acids</td>
<td>38-40</td>
</tr>
<tr>
<td>Hydantoinases</td>
<td>S-Mono-substituted hydantoins</td>
<td>Amino acids</td>
<td>41, 42</td>
</tr>
<tr>
<td>Halohydrin dehalogenases</td>
<td>Halohydrins, epoxides</td>
<td>Diols, epoxides, p-hydroxyamino acids</td>
<td>43</td>
</tr>
<tr>
<td>Ammonia lyases</td>
<td>Cinnamic acid derivatives</td>
<td>Phenylalanine derivatives</td>
<td>44</td>
</tr>
<tr>
<td>Proteases</td>
<td>Amino acids</td>
<td>Peptides</td>
<td>45</td>
</tr>
</tbody>
</table>

© 2008, John Wiley & Sons, Inc.
Pharmaceutical Industry, Biocatalysts and Chemocatalysts

Figure 1  Cyanohydrin formation with hydroxy nitrilases.

Figure 2  Synthesis of L-methionine.

An example of an acylase to perform a resolution is provided by the Degussa process to L-methionine (1). The racemic acetyl methylthionine (2) is prepared by a chemical synthesis. The acylase hydrolyses only the L-isomer (Fig. 2). The D-isomer is racemized by base and put back into the process stream (48).

The most powerful approaches, which can be used with several different enzyme systems, lead to a single enantiomer as the product in high yield and do not rely on a classic resolution approach in which the unwanted enantiomer is discarded. These approaches include dynamic kinetic resolutions, deracemizations, and asymmetric and desymmetrization reactions (49, 50). In some cases, a chemical catalyst may be available to "recycle" the unwanted isomer in the same reactor (vide infra). It is sometimes possible to racemize the unwanted isomer of the substrate and then to perform the reaction again (51).

Whole cells

When chemical transformations were performed by whole cells, such as the reduction of carbonyl compounds by baker's yeast, low asymmetric induction could result as two enzymes are present in the organism that provide the antipodes of the product (52). This result has now been circumvented by the use of genetically modified microorganisms so that the desired enzyme is overproduced (53, 54).

The use of a whole cell allows for a required enzyme cofactor to be regenerated. In other cases, it allows for several enzymes to work in parallel and to perform many complex transformations. An example is provided by the synthesis of D-amino acids from hydantoins (Fig. 3). The carbomylase drives the reaction to completion as carbon dioxide and ammonia are evolved. The same approach has been used with the L-versions of the enzymes to synthesize L-amino acids (14, 42, 55).

Several complex antibiotics are prepared by whole-cell fermentations. Examples are the penicillin antibiotics in which the side chain can be removed and replaced with a synthetic one to enhance activity or stability. Other examples include the macrolide antibiotics, such as avermectin (56) and erythromycin (57), in which the organism uses an enzyme "cassette" to build up the seco-chain before cyclization.

Figure 3  Synthesis of D-amino acids from hydantoins.
In some instances, metabolic engineering of an organism can provide the desired compound. As an example, shikimic acid is used as the starting material in the synthesis of Tamiflu (Roche Laboratories, Inc., Nutley, N.J.), which is an antiviral drug. Bacteria produce shikimic acid as an intermediate on the biosynthetic route to chorismic acid, itself an intermediate for several essential products such as phenylalanine, tyrosine, and ubiquinone that the cell needs to function (58). By knocking out or controlling the genes that develop the enzymes that use shikimic acid as the substrate, the organism can be persuaded to overproduce this valuable starting material (Fig. 4).

Chemists have also taken lessons from nature and often use biomimetic syntheses or approaches to complex molecules; here, reactions used in an organism are mimicked in the laboratory (59, 60). In addition, catalytic transformations can be coupled, and it could be two chemocatalysts (vide infra) (61, 62).

**Chemical Catalysts**

**Transition metal**
Transition metal-based catalysts perform a wide variety of reactions. Many useful reactions can be used to build the carbon–carbon framework of the target molecule or to introduce functional groups into complex molecules. Many achiral methods exist; they often are named after the person who discovered or popularized them (see Table 3 (64-170)). In some instances, the achiral reaction has been adapted to provide an asymmetric method; the latter examples are included in Table 4 (93, 118, 120, 124, 142, 149–202). The use of metal catalysts that act as Lewis acids or bases have been omitted as numerous examples can be described (63).

When implementing a transition metal-catalyzed step at scale, many factors have to be considered, some of which also relate to biological and organocatalytic reactions. The one factor that does not overlap with these other types of systems is the price of the metal. Although cheaper metals such as iron, nickel, and copper can be used for some transformations, often the metal required is precious, such as palladium, platinum, rhodium, iridium, or ruthenium. The use of gold catalysis has recently become an area of intense research (141). These precious metals are expensive; usage needs to be minimal, and they must be recycled either for reuse in the reaction or through recovery. Refining has to be a topic of considerable economic concern.

For some reactions, especially asymmetric transformations, the ligands needed to perform the reaction may be more expensive than the metal! Here, the catalyst has to be extremely efficient to achieve the required cost benefits. The economics of the

<table>
<thead>
<tr>
<th>Reaction type</th>
<th>Substrate</th>
<th>Product</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aryl coupling reactions</td>
<td>Alkene</td>
<td>Arylalkene</td>
<td>64-71</td>
</tr>
<tr>
<td>Heck reaction</td>
<td>Boronic acid or ester</td>
<td>Biaryl</td>
<td>71, 73, 84-86</td>
</tr>
<tr>
<td>Suzuki reaction</td>
<td>Amine</td>
<td>Biaryl</td>
<td>71, 73, 84-86</td>
</tr>
<tr>
<td>Buchwald-Hartwig reaction</td>
<td>Alkene</td>
<td>Aniline</td>
<td>87-89</td>
</tr>
<tr>
<td>Cyclizations</td>
<td>Various</td>
<td>Various</td>
<td>90-93</td>
</tr>
<tr>
<td>Metathesis</td>
<td>Alkene or alkyne</td>
<td>Alkene or alkyne</td>
<td>93-116</td>
</tr>
<tr>
<td>Hydroformylations</td>
<td>Alkenes</td>
<td>Aldehydes</td>
<td>121-123</td>
</tr>
<tr>
<td>Hydroaminations</td>
<td>Alkene</td>
<td>Amine</td>
<td>125-127</td>
</tr>
<tr>
<td>Protecting group removal</td>
<td>Various</td>
<td>Various</td>
<td>128</td>
</tr>
<tr>
<td>Pauson-Khand reaction</td>
<td>Alkenes</td>
<td>Ketones</td>
<td>129-133</td>
</tr>
<tr>
<td>Oxidations</td>
<td>Alcohols</td>
<td>Carbonyl compounds</td>
<td>134-136</td>
</tr>
<tr>
<td></td>
<td>Alkenes</td>
<td>Epoxide</td>
<td>137</td>
</tr>
<tr>
<td></td>
<td>Alkenes</td>
<td>Dial</td>
<td>138, 139</td>
</tr>
</tbody>
</table>

These include couplings such as the Kumada, Sonogashira, Negishi, and Stille reactions.

---

**Figure 4** Biosynthetic access to chorismic and shikimic acids.
Pharmaceutical Industry, Biocatalysts and Chemocatalysts

<table>
<thead>
<tr>
<th>Reaction type</th>
<th>Substrate</th>
<th>Product</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogenation</td>
<td>Enamides</td>
<td>α-Amino acid derivatives</td>
<td>118, 120, 149-153</td>
</tr>
<tr>
<td></td>
<td>α,β-Unsaturated carboxylic acid derivatives</td>
<td>118, 152, 153</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enamines</td>
<td>Amines</td>
<td>118, 152, 153</td>
</tr>
<tr>
<td></td>
<td>Imines</td>
<td>Amines</td>
<td>118, 152, 153</td>
</tr>
<tr>
<td></td>
<td>Ketones</td>
<td>Alcohols</td>
<td>118, 152, 153, 155, 156</td>
</tr>
<tr>
<td></td>
<td>Alkenes</td>
<td>Alkenes</td>
<td>118, 157</td>
</tr>
<tr>
<td>Reductions</td>
<td>Ketones</td>
<td>Alcohols</td>
<td>158</td>
</tr>
<tr>
<td>Hydroamination</td>
<td>Alkene</td>
<td>Amines</td>
<td>159</td>
</tr>
<tr>
<td>Hydroxylation</td>
<td>Alkene</td>
<td>Alkene</td>
<td>124</td>
</tr>
<tr>
<td>Hydroisylations</td>
<td>Carbonyl compounds</td>
<td>Alcohols</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Carbonyl compounds</td>
<td>α-Substituted carbonyl compounds</td>
<td>161</td>
</tr>
<tr>
<td>Reductions</td>
<td>Ketones</td>
<td>Alcohols</td>
<td>158</td>
</tr>
<tr>
<td>Hydroamination</td>
<td>Alkene</td>
<td>Amines</td>
<td>159</td>
</tr>
<tr>
<td>Hydroxylation</td>
<td>Alkene</td>
<td>Alkene</td>
<td>124</td>
</tr>
<tr>
<td>Hydroisylations</td>
<td>Carbonyl compounds</td>
<td>Alcohols</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>Carbonyl compounds</td>
<td>α-Substituted carbonyl compounds</td>
<td>161</td>
</tr>
<tr>
<td>Strecker reaction</td>
<td>Carbonyl compounds</td>
<td>α-Amino nitrites</td>
<td>165</td>
</tr>
<tr>
<td>Cyanohydrin formation</td>
<td>Carbonyl compounds</td>
<td>α-Hydroxy nitrites</td>
<td>166-168</td>
</tr>
<tr>
<td>Allylic alkylations</td>
<td>Allyl esters or similar</td>
<td>Alkene</td>
<td>169-172</td>
</tr>
<tr>
<td>Aldol and related reactions</td>
<td>Carbonyl compounds</td>
<td>β-Hydroxy carbonyl compounds</td>
<td>173</td>
</tr>
<tr>
<td>Conjugate additions</td>
<td>α,β-Unsaturated compounds</td>
<td>β-Substituted compounds</td>
<td>174-176</td>
</tr>
<tr>
<td>Halogenations</td>
<td>Carbonyl compounds</td>
<td>α-Halo compound</td>
<td>177</td>
</tr>
<tr>
<td>Isomerizations</td>
<td>Alkene</td>
<td>Alkene</td>
<td>178, 153</td>
</tr>
<tr>
<td>Hydrolysis</td>
<td>Epoxides</td>
<td>Diols</td>
<td>179, 180, 181</td>
</tr>
<tr>
<td>Oxidations</td>
<td>Alkenes</td>
<td>Epoxides</td>
<td>182-188</td>
</tr>
<tr>
<td></td>
<td>Alkenes</td>
<td>Aziridines</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>Alkenes</td>
<td>Alkene</td>
<td>196-198</td>
</tr>
<tr>
<td>C-H activation</td>
<td>Various</td>
<td>Various</td>
<td>199, 200</td>
</tr>
<tr>
<td>Heck reaction</td>
<td>Alkene</td>
<td>Alkenes</td>
<td>71, 201, 202</td>
</tr>
<tr>
<td>Metathesis</td>
<td>Alkenes</td>
<td>Alkenes</td>
<td>93</td>
</tr>
</tbody>
</table>

transformation not only depend on the cost of the catalyst and how much is used (usually defined by turnover number, which is the number of times the catalyst goes round the catalytic cycle), but also the duration of the reaction. The turnover frequency is the number of times the catalyst completes a catalytic cycle per hour. Reactor time can be expensive, and time needs to be minimized but not at the cost of making the reaction so fast that it becomes unsafe or reagents, such as hydrogen, cannot be delivered at an appropriate rate.

An example of a metal-catalyzed reaction to form a biaryl product is the Suzuki reaction. The coupling can be performed without any phosphorus ligands for the metal and with only a small amount of the metal (0.05 mol %) (Fig. 5). A reaction that has become popular is the preparation of aromatic

![Figure 5](image_url)
amines by a palladium-catalyzed coupling reaction (Fig. 6). This methodology is general (89).

In addition to carbon–carbon bond formation, transition metal catalysts can also generate a stereogenic center. The first reaction of this type in which useful amounts of asymmetric induction were observed was an asymmetric hydrogenation to make phenylalanine and the method has been used for many years to synthesize the anti-Parkinsons drug, l-Dopa (3) (Fig. 7) (142, 143).

This transformation was important as it showed that a chemical catalyst could perform with similar asymmetric integrity to that of a biological system. Today, literally thousands of ligands and catalysts can be used to perform asymmetric hydrogenations as well as other reactions; see Table 4.

Many aspects must be considered in finding a catalyst to perform a step in the synthesis of a drug. The main aspect is the time required to find suitable catalyst systems. If a closely analogous reaction has been reported in the literature, then it may not be a large problem or concern. In most instances, however, this is not the case. In addition to enantioselectivity or diastereoselectivity, the factors necessary to find an efficient achiral catalyst must also be fulfilled.

Stereogenic centers can also be prepared by carbon–carbon bond-forming reactions or reductions of functional groups other than alkenes. Some reactions are also summarized in Table 4 (144). For a comprehensive work on asymmetric catalysts, see Reference 145. In some cases, two stereogenic centers can be created. This result can be achieved either in a single step as
with the asymmetric reduction of a tetrasubstituted alkene, or by coupling two reactions together as with a conjugate addition followed by trapping the resultant enolate with an electrophile (146, 147). An illustration of this strategy is the synthesis of the bicyclic ketone 4 (Fig. 8) (147, 148). The allyl group is a good electrophile and is then converted to the analogous ketone by a Wacker oxidation.

An example of an asymmetric hydrogenation used in the preparation of a pharmaceutical intermediate is provided by a synthesis to carbapenems (5) (178). Reduction of the \( \beta \)-keto ester occurs under equilibrating conditions so that the erythro-product is formed in high yield and selectivity (203). Another catalytic step with ruthenium is used to introduce the acetoxy group (Fig. 9) (153).

An asymmetric oxidation is used in the synthesis of esomeprazole (6), a proton pump inhibitor, which has therapeutic advantages over the racemic mixture omeprazole (Fig. 10) (204).

Chemocatalysts sometimes have an advantage over biological systems. Often the antipode of a ligand is accessible, although if a natural product is used as the source of the stereogenicity, then it may be less abundant and more expensive. As a last resort, and as ligands are relatively small molecules, an achiral synthesis and resolution might be used. This latter option is not available with a biological catalyst.

One of the main concerns of using a transition or heavy metal catalyst, especially toward the end of the synthetic sequence, is the removal of the metal. A wide variety of methods is known
Table 5: Examples of transformations catalyzed by organocatalysts

<table>
<thead>
<tr>
<th>Reaction type</th>
<th>Substrate</th>
<th>Catalyst type</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoxidation</td>
<td>Alkenes</td>
<td>Carbohydrate derivatives</td>
<td>220, 221</td>
</tr>
<tr>
<td>Alkylation</td>
<td>Carbonyl compounds</td>
<td>Alkaloids, amines</td>
<td>222, 223</td>
</tr>
<tr>
<td>Aldol reaction</td>
<td>Carbonyl compounds</td>
<td>Amino acid derivatives</td>
<td>231-240</td>
</tr>
<tr>
<td>Mannich reaction</td>
<td>Carbonyl compounds</td>
<td>Amino acid derivatives</td>
<td>233, 240, 241</td>
</tr>
<tr>
<td>Conjugate additions</td>
<td>Unsaturated carbonyl compounds</td>
<td>Various</td>
<td>242</td>
</tr>
<tr>
<td>Baylis-Hillman reaction</td>
<td>α,β-Unsaturated carbonyl compounds</td>
<td>Nucleophilic</td>
<td>243-245</td>
</tr>
<tr>
<td>Acylations</td>
<td>Alcohol</td>
<td>Ester</td>
<td>246-249</td>
</tr>
<tr>
<td>Hydrosyldations</td>
<td>Carbonyl compounds</td>
<td>Amino acid derivatives</td>
<td>250</td>
</tr>
<tr>
<td>Reductions</td>
<td>Carbonyl compounds</td>
<td>Dihydropyridines</td>
<td>251</td>
</tr>
<tr>
<td>Stetter and benzoin reactions</td>
<td>Aldehydes</td>
<td>Carbenes</td>
<td>249, 252</td>
</tr>
</tbody>
</table>

Organocatalysts

This class of catalysts covers chemocatalysts that do not contain a transition metal. The class has been known for many years, but it is relatively recently that the term "organocatalyst" has been used (209). A wide variety of transformations can be performed, which is currently an area of intense research (209-218). Table 5 (220-252) summarizes some key transformations in which organocatalysis can be useful. Reactions range from the asymmetric epoxidation of alkenes, which need not be conjugated to another functional group, to aldol reactions and other carbon-carbon forming transformations. Some progress has also been made to couple two reactions together (219).

\( \alpha \)-Proline catalyzes the aldol reaction. This approach has been applied to the synthesis of carbohydrate derivatives as illustrated by the glucose derivative 7 (Fig. 11) (237). The three-component Mannich reaction can be used to prepare \( \beta \)-amino and \( \beta \)-amino \( \alpha \)-hydroxy carbonyl compounds in a single step (Fig. 12) (233). As with other types of catalysts, organocatalysts can be immobilized to aid recovery (253).

Systems with Biocatalysis and Chemocatalysis

As enzymes usually only accept one enantiomer or isomer as substrate, many enzymatic reactions are resolutions; the
unaffected isomer is waste. One way to circumvent this problem, which can have significant economical consequences, is to include a racemization or isomerization step with a second catalyst so that the substrate for the desired transformation can be accepted as the correct isomer (254, 255). This method allows dynamic kinetic resolutions to be performed with the desired product isomer being produced in high yield rather than with the 50% maximum available from a classic resolution approach (256).

A chemical catalyst can be used to racemize an alcohol, whereas an enzyme is used to prepare an ester of one of the enantiomers of that alcohol. In this example, reduced pressure was used to remove the isopropanol by-product and drive the reaction to completion whereas the Shvo catalyst was used to racemize the alcohol (Fig. 13) (257).

Future Outlook

Both biological and chemical-based catalysts are useful for a wide variety of reactions that range from carbon-carbon bond formation to the generation of a new stereogenic center. With the increasing awareness of green chemistry and the need to reduce waste in the pharmaceutical industry where this problem has been particularly bad, the use of catalytic reactions will surely continue to increase.

Biocatalysts are being applied widely in the industry, including the preparation of carbon-carbon bonds. Stereoselective oxidation with biocatalysts is an area where chemistry will find it hard to compete. A need still exists for new catalysts to replace stoichiometric reagents, as in the reduction of an amide to an amine, amide formation, and substitution of an alcohol (Mitsunobu reaction) (258). In both arenas of catalysis, the overall goal for green chemistry and stereoselectivity must be carbon-hydrogen bond activation.

References


113. Phillips AJ, Abell AD. Ring-closing metathesis of nitrogen-
112. Grubbs RH, Chang S. Recent advances in olefin metathesis
111. Grubbs RH, Miller SJ, Fu GC. Ring-closing metathesis and
110. Schrock RR, Czekelius C. Recent advances in the syntheses
108. Grubbs RH. Olefin-metathesis catalysts for the preparation of
107. Chauvin Y. Olefin metathesis: the early days (Nobel Lecture
106. Zhang W, Moore JS. Alkyne metathesis: catalysts and synthetic
104. Nicolaou KC, Bulger PG, Sarlah D. Metathesis reactions in total


Pharmaceutical Industry, Biocatalysts and Chemocatalysts


Further Reading

See Also
Enzyme Catalysis, Chemical Strategies for
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Kuo-Hsiung Lee, Hideji Itokawa, Toshiyuki Akiyama and Susan L. Morris-Natschke, Natural Products Research Laboratories, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina
doi: 10.1002/9780470048672.wecb373

Numerous important bioactive compounds have been, and continue to be, isolated worldwide from natural sources. These compounds include both primary and secondary metabolites isolated mainly from plants, as well as from the animal and mineral kingdoms. The recent development of new bioassay methods has facilitated progress in the BDH (bioactivity-directed fractionation and isolation) of many useful bioactive compounds from natural sources (1). These active principles could be developed or additionally modified to enhance the biologic profiles as clinical trials candidates. Many natural pure compounds have become medicines, dietary supplements, and other useful commercial products. This article summarizes research on many different useful compounds isolated or developed from plants with an emphasis on those discovered recently by the laboratories of the authors as antitumor and anti-HIV clinical trial candidates.

From ancient times, many crude herbs have been used as remedies for various diseases. In Asia, these therapies include traditional Chinese medicine (TCM), Japanese-Chinese medicine (Kampo), Korean-Chinese medicine, Jamu (Indonesia), and Ayurvedic medicine (India). In Europe, phytotherapy and homeopathy have found medicinal use. These herbal therapies generally are classified as “alternative medicines” in America. Alternative medicine, which comes mainly from the aforementioned traditional and folk medicines used worldwide, also is being combined with conventional medicine (Western medicine), which results in integrative medicine. In TCM, crude herbal drugs formerly were divided into three categories: upper-, middle-, and lower-class medicines. Upper-class medicines usually are not toxic, have moderate physiologic effects, and often are used to maintain good health. Thus, they sometimes are called therapeutic drugs. Both upper- and middle-class medicines are used as therapeutic drugs, but the former medicines are not as toxic as the latter. Lower-class medicines can contain very toxic substances, which can be used judiciously as medicines. TCM prescriptions usually mix herbs that belong to these three categories, according to a unique principle (2). The experience gained by using TCM for centuries provides a rich source of information for modern research in drug discovery.

Anticancer and Antitumor Compounds

Suffness and Douros (3) suggested the following definitions to avoid confusion of terminology. Cytotoxicity is used when compounds or extracts show in vitro (in cells) activity against tumor cell lines. Antitumor or antineoplastic indicates that the materials are effective in vivo (in animals) in experimental systems. Anticancer refers to compounds that are active clinically against human cancer.

The development of novel, clinically useful anticancer agents is highly dependent on the bio assay screening systems, as well as the sample sources. Two bioassay types have been used: cell-based and mechanism of action (MOA)-based. Initially, cell-based assays used primarily L1210, P388, and nasopharyngeal (KB) cells in preliminary screening for antitumor activity. Screening against a panel of human cancer cell
lines was implemented to discover agents active against different types of cancer. For active agents, the in vitro studies are followed by in vivo xenograft studies to ensure efficacy. New MOA-based bioassay systems aimed at particular molecular targets also have revolutionized the discovery of potential drug candidates. Important anticancer drug targets include tubulin, DNA topoisomerases I and II (topo I and topo II), cyclin-dependent kinases (CDKs), growth and transcription factors, and so forth.

Higher plants have yielded many effective, clinically useful anticancer drugs, including those derived from Catharanthus alkaloids, Taxus diterpenes, Camptotheca alkaloids, and Podophyllum lignans. Research in this area has been reviewed extensively (3–13).

The seminal discoveries of taxol (tubulin-interactive) and camptothecin (topo I-interactive) by Wall and Wani (14–16) represent how natural products have influenced the additional development of natural product-derived and synthetic entities. The following discussion of the discovery and development of current important antineoplastic compounds will be organized by plant species.

**Catharanthus species**

Catharanthus alkaloids, particularly vinblastine (A1) and vincristine (A2), are well-known anticancer drugs, which are used clinically to treat Hodgkin’s lymphoma and acute childhood lymphoblastic leukemia, respectively. These alkaloids interact with tubulin, a protein necessary for cell division, and are inhibitors of mitosis (the process of cell division).

Originally, these compounds were isolated from *Catharanthus roseus* (L.) G. Don [formerly known as *Vinca rosea* (Apocynaceae family)], which is used as folk medicine in Madagascar to inhibit milk secretion and as a hypotensive agent, astringent, and emetic. Moreover, native people in England and the West Indies have used this species to lower blood sugar.

Numerous synthetic analogs have been designed to have activity against other tumor types or to have fewer side effects. Among them, navelbine (vinorelbine) (A3) was developed by Burroughs Wellcome (17) and is used against non-small cell lung and advanced breast cancers. This synthetic analog of A1 has an eight-membered, rather than a nine-membered, C ring, and a dehydrated D ring (18). Eldisine (vindesine) (A4)
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

**Figure 2** Structures of natural and synthetic taxoids.

is another structurally modified analog, which is used against acute lymphoblastic leukemia, breast cancer, and malignant melanoma.

EC145 (A5), a folic acid conjugate of desacetyl vinblastine monohydrate, represents a new generation of receptor specific targeted chemotherapy and is undergoing Phase I anticancer clinical trials (19). Phase II trials for bladder and kidney cancers are underway with vinflunine (A6), a bifluorinated vinorelbine derivative (20, 21).

**Taxus species**

Taxol (paclitaxel) (B1), a taxane diterpene isolated from the bark of the Pacific yew tree Taxus brevifolia Nutt. (Taxaceae family), is used extensively in patients with advanced and metastatic ovarian and breast tumors, particularly tumors that are refractory to standard chemotherapy. Initially, supply problems severely limited the full exploration of its antineoplastic potential. However, the semisynthesis of B1 from 10-deacetylbaccatin III (B2), which is isolated from needles of the European yew tree, Taxus baccata L., provided an alternative renewable resource to resolve the supply problems.

Wall and Wani (22) are the pioneers in taxol discovery. To date, around 400 taxoids have been isolated from the Taxus species. Taxus alkaloids were reviewed recently in the book Taxus, genus Taxus edited by Ikokawa and Lee (23). Biologic
activity and the chemistry of taxoids from the Japanese yew also have been reviewed (24).

\[ B_1 \] interacts with cellular tubulin via promotion of microtubule assembly and inhibits mitosis. It is active against breast, brain, tongue, endometrial, and ovarian, as well as other, cancers (25, 26). The clinically used analog docetaxel (taxotere) (\[ B_3 \]) is synthesized from the more readily available \[ B_2 \]. Taxotere is used particularly against non-small cell lung cancers.

Extensive structure-activity relationship (SAR) studies have led to many related antineoplastic taxane analogs, including ortataxel (\[ B_4 \]), an orally administered taxoid in Phase II clinical trials (27). SAR studies of ring C-seco-taxoids were published recently (28).

In addition, conjugates between taxol and various other compounds, such as 3,17β-estradiol (29), various fatty acids (30), or a biodegradable polymer (poly-L-glutamic acid, paclitaxel polyglumex) (31, 32), seek to improve drug targeting or tissue distribution. The laboratories of the authors have conjugated taxoids with other anticancer agents, including epipodophyllotoxins (\[ B_5 \]) (33) and camptothecin (\[ B_6 \]) (34). A recent review

---

**Figure 3** Structures of natural and synthetic camptothecins.
provides an overview of novel formulations of taxanes, including many in clinical trials, developed to overcome the solubility issues with B1 and B3 (35).

**Camptotheca species**

Camptothecin (CPT, C1) is a potent antitumor pentacyclic alkaloid isolated from Camptotheca acuminata Decne. (Nyssaceae family) and originating in China (36, 37). Interest in CPT was sparked by the discovery that its primary cellular target is DNA topo I (38). 10-Hydroxycamptothecin (C2), which also occurs naturally, has a better therapeutic index and is used in China for treating many cancers.

Poor water solubility of the natural products led to the development of the semisynthetic, more water-soluble analogs topotecan (Hycamptin, C3) and irinotecan (Camptosar, C4), which are used primarily for the treatment of advanced ovarian and metastatic colorectal cancers, respectively (39). Some CPT analogs, such as C5 synthesized in the laboratories of the authors showed greater topo I inhibition than CPT (40). Additional CPT analogs also are of interest in combination regimens as radiation sensitizers (41). Two epipodophyllotoxin-camptothecin conjugates, C6 and C7, from the laboratories of the authors exhibit dual mechanisms of action, being both topoisomerase I/II-inhibitory (42). They have improved in vitro anticancer profiles (42) and are active against etoposide- and camptothecin-resistant KB cells (KB7D and KB/CPT100, respectively).

Clinical application and perspectives on the CPTs have been discussed in several excellent reviews (43–45). New CPT
analogs in anticancer clinical trials include DB-67 (a silatecan or 7-silylcamptothecan, C8 (46) and rubitecan (9-nitrocamptothecin, C9 (47)).

**Podophyllum species**

The genus Podophyllum (Berberidaceae family), including the American *P. peltatum* L. (American mayapple) and Indian or Tibetan *P. emodi* Wall (syn. *P. hexandrum* Royle), has been used for centuries for its medicinal properties. Podophyllin, a resin obtained from an alcoholic extract of Podophyllum rhizome, has been used for a long time as a remedy for warts and was listed in the U.S. Pharmacopoeia from 1820 to 1942, when it was removed because of undesirable toxicity (48).

Podophyllotoxin (D1) is an aryltetralinlactone cyclolignan with a flat, rigid five-ring skeleton; it was isolated in 1880 from rhizomes of *P. peltatum*. It was found to show antineoplastic

**Figure 6** Structures of colchicine and related compounds.

**Figure 7** Structures of tanshinlactones and neo-tanshinlactone.
activity, but it is highly toxic and failed the U.S. National Cancer Institute (NCI) Phase I clinical trials as an antitumor drug in the 1970s. Chemical modification of D1 led to successful development of the clinically useful antitumor drugs etoposide (D2) and teniposide (D3). These compounds inhibit cellular DNA topo II and are used to treat small cell lung and testicular cancers and lymphomas/leukemias. Etoposide (etoposide phosphate, D4) is a clinically used, water-soluble phosphate ester of etoposide. It lessens the chance of precipitation of the drug during intraavenous administration.

Limitations of D2, including myelosuppression, drug resistance development, and poor water solubility, prompted extensive SAR studies. Using drug improvement principles, several series of 4-alkylamino and 4-arylamino epipodophyllotoxin analogs were synthesized and showed increased inhibition of DNA topo II activity and increased cytotoxicity in D2-resistant cell lines (49–51). From the preclinical development in the laboratories of the authors, GL-331 (D5) (52), which contains a p-nitroanilino moiety at the 4-position of D2, emerged as a clinical trials candidate. Compared with D2, GL-331 has advantages of better water solubility, easier manufacturability, and fewer side effects. It also shows cytotoxic activity against D2-resistant cancer cell lines. GL-331 progressed to Phase IIa clinical trials as an anticancer drug (53).

The rational design of improved D2 analogs has made use of several new computational strategies (52, 54–56). In 2004, Goralda et al. (57) reviewed the distribution, sources, application, and new cytotoxic derivatives of D2. More recently, Lee and X Lao (58) also have reviewed podophyllotoxins and related analogs, including GL-331, to demonstrate how plant natural products can lead to successful preclinical drug candidates.

Cephalotaxus species

The genus Cephalotaxus (Cephalotaxaceae family) contains coniferous evergreen trees and shrubs that are indigenous to Asia. Historically, the bark has long been used in TCM for various indications. Powell et al. (59–61) originally isolated the antitumor alkaloids homoharringtonine (E1) and harringtonine (E2). Subsequently, Chinese investigators discovered that alkaloids from C. fortunei Hook. F possess antitumor activity (62). Consequently, E1 was obtained from the Chinese evergreen tree C. harringtonia K. Koch var. harringtonia (63), and other active alkaloids were isolated from various Cephalotaxus species (64, 65). Interestingly, the parent compound, cephalotaxine (E3), is devoid of antitumor activity. E3 has been investigated in anticancer clinical trials at the U.S. National Cancer Institute (NCI), particularly for the treatment of myeloid leukemia (66, 67). However, its side effects still remain an issue. Thus, the authors have continued to study new natural constituents of Cephalotaxus species and to develop new analogs on the basis of SAR studies, as presented in a review by Itokawa et al. (68).

Colchicine species

The alkaloid colchicine (F1) isolated from the medicinal plant Colchicum autumnale L. (Liliaceae family) still is used to treat gout and familial Mediterranean fever. F1 and thiocolchicines (F2) (SCH, rather than OCH3 at C-10), which is more stable and more potent but slightly more toxic, are mitotic inhibitors that inhibit polymerization of tubulin (69). Although they show antileukemic activity, they are too toxic to use as anticancer agents, which prompts the synthesis of new, less toxic analogs. Replacing the C-7 acetamide group on the B ring with various oxygen-containing groups (ketone (F3), thiocolchicoline, hydroxyl (F4), and ester (F5–F6)) led to compounds that were equally or more active in vitro than F2. The C-ring contracted colchinol-7-one bismethyl ether or allo-ketone (F7) was equipotent with the seven-membered ring natural product F1. Three-related, ring-contracted colchinicinoids (F8–F10) showed excellent activity in drug-sensitive and drug-resistant K562 cell lines (71).

The above synthetic colchicinoids have three methylated phenolic groups in the A ring, which are needed for full potency as tubulin/mitotic inhibitors. Removing one or two of the methyl groups reduces potency, and trimethylated colchicines and thiocholicchines (F11–F14) no longer interact with tubulin but rather are a new class of DNA topo II inhibitory agents (72). They show in vitro activity against bone and breast cancers (73).

Salvia species

Salvia officinalis L. (Labiatae family) is native to Europe and America, but the roots and rhizome of S. miltiorrhiza Bunge (called Tanshen) have been used widely in China for treating various cardiac (heart) and vascular (blood vessel) disorders, such as atherosclerosis or blood-clotting abnormalities. This plant exhibits hypotensive effects, causes coronary artery vasodilation, and inhibits platelet aggregation. Accordingly, it should not be used in combination with warfarin. Tanshen also has been applied for hemorrhage, dysmenorrhea, miscarriage, swelling, inflammation, chronic hepatitis, and insomnia (74, 75). Clinically available preparations of a S. miltiorrhiza / Oliberia odorifera mixture may have potential as an anti-anginal drug (76).

Tanshinone diterpenoids, including tanshineone I (G1) and tanshineone IIa (G2), are bioactive compounds from S. miltiorrhiza (77). Sodium tanshinone sulfate (G3) is a water-soluble derivative of G2, used clinically to treat angina pectoris and myocardial infarction, and exhibits strong membrane-stabilizing effects on red blood corpuscles. In addition, novel seco-

Figure 8 Structure of sinococuline.
Other studies have shown that S. miltiorrhiza exerts clear cytotoxic effects and strongly inhibits the proliferation of liver cancer cells (79). Various tanshinones were tested in SAR studies against several human tumor cells, namely nasopharyngeal (KB), cervical (Hela), colon (Colon 205), and laryngeal (Hep-2) cell systems (74, 75). Neo-tanshinlactone (G4), also originally isolated from Tan-shen, has a unique and different structure compared with other compounds from S. miltiorrhiza Bunge. Preliminary studies of this compound showed unique specific activity against the...
in vivo from the stems and rhizomes (82). It also has analgesic, and anti-inflammatory crude drug. Mountains of east Asia, has folkloric uses as a diuretic, an in vitro activity of this compound type (81). The major pigment in Curcuma species (Zingiberaceae family) is the yellow phenolic diarylheptanoid curcumin (J1). Curcumin and its analogs have potent antioxidant and anti-inflammatory effects, cytotoxicity against tumor cells, and antitumor-promoting activity (96). The biologic effects and targets of curcumin, as well as its possible roles in cancer prevention and therapy, have been reviewed recently (97, 98).

Cocculus species

Cocculus trilobus DC. (Menispermaceae family), found in the mountains of east Asia, has folkloric uses as a diuretic, an analgesic, and an anti-inflammatory crude drug. Smicosoline (H1) was isolated as an anti-tumor principle from the stems and rhizomes (82). It also has in vivo activity against P388 leukemia. H2 likely is a general cytotoxic rather than a cell-specific agent (83).

Maytenus species

Maytenus ilicifolia Mart. ex R. E. F. (Celastraceae family), more commonly known as "Cangora," is found in South America where it is used for its analgesic, antipyretic, anti-septic, and anticancer properties and for birth control, particularly in Paraguay. Bioactivity-directed fractionation and isolation by various research groups has led to the isolation of various active principles. Kupchan et al. (84, 85) first identified antileukemic maytansinoids, for example, maytansine (I1), from the African plant Maytenus ovata (later renamed M. serrata (Hochst. ex A. Rich.) R. Wilczek). Although it advanced to Phase II clinical trials, testing then was suspended because of neurotoxicity. The related maytansine (I2), isolated from M. diversifolia (Maxim.) Ding Hou, has been investigated for growth-inhibiting and apoptosis-inducing activities in K562 leukemia cells (86). Cyclotoxic monoterpene indoles include pristimerin (I3) and isolignostane H (I4) from M. ilicifolia (87), as well as the triterpene dimers dihydroisocangorosin A (I5) and cangorosin B (I6) (88, 89). The authors also have identified cytotoxic sesquiterpene pyridine alkaloids, including emargatinones B (I7) and F (I8), from M. emarginata (Willd.) Ding Hou (90, 91).

Recently, new triterpenes were isolated from M. chuchu-huasca R. Hamet and Colas (92), along with sesquiterpenes from the same plant (93). New compounds also were reported from M. cuzzoinea Loes. (95). The major pigment in Curcuma species (Zingiberaceae family) is the yellow phenolic diarylheptanoid curcumin (J1). Curcumin and its analogs have potent antioxidant and anti-inflammatory effects, cytotoxicity against tumor cells, and antitumor-promoting activity (96). The biologic effects and targets of curcumin, as well as its possible roles in cancer prevention and therapy, have been reviewed recently (97, 98).

Several synthetic curcumin analogs, including J2, showed potent antiandrogenic activities against two human prostate cancer cell lines, PC-3 and DU-145 (99). In expanded in vitro testing, these synthetic curcumin derivatives showed antiprostate cancer activity superior to that of hydroxyflutamide, the currently available and preferred anti-androgen for treating prostate cancer (100). In continuing work in the laboratories of the authors, dimethoxy-4-ethoxycarbonylethyl-curcumin (J3) showed potent anti-androgenic activity and therefore is
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Figure 11 Structures of Kansui natural products.

K1 Kansuiphorin A, R = CO(CH2)14CH3
K2 Kansuiphorin B
K3 Kansuiphorin C
K4 Kansuiphorin D
K5 DBDI, R = H
K6 Yuexiandajisu D, R1 = β-OH, R2 = α-OH
K7 Yuexiandajisu E, R1 = α-OH, R2 = β-OH
K8 Yuexiandajisu F

Euphorbia species

Euphorbia kansui Liou (Euphorbiaceae family) is distributed widely in northwest China. The dried roots of the plant are known as “kansui” and classified as a lower-class medicine. It is used as an herbal remedy for ascites (abdominal fluid accumulation) and cancer in China. Ingenol diterpenoids are among the bioactive chemical constituents. Kansuiphorins A–D (K1–K4) were isolated as cytotoxic principles of E. kansui by the laboratories of the authors (105, 106). In particular, K1 and K2 demonstrated potent antileukemic activity against P-388 leukemia in mice (105). A related ingenol-type diterpene (DBDI), K5 showed unique suppression of mast cell activation, a process that occurs during inflammation, and thus, might have the potential to treat allergic diseases (107). In other pharmacological studies, two isolated ingenols from an immunoenhancing E. kansui extract increased immune activity in a dose-dependent manner (108).

Three new cytotoxic diterpenoids, Yuexiandajisus D (K6), E (K7), and F (K8), were isolated from the species E. ebracteolata Hayata. (109). In additional studies, K6 showed moderate cytotoxicity against additional (HCT-8 and Bel-7402) cell lines (110).

Anti-HIV Compounds

Acquired immunodeficiency syndrome (AIDS), a degenerative disease of the immune system, is caused by the human immunodeficiency virus (HIV) and results in life-threatening opportunistic infections and malignancies. Antiviral and immunomodulating natural products have been investigated as treatments for AIDS (111).

Lomatium suksdorfii (coumarin derivatives)

Suksdorfin (L1), a dihydroseselin-type angular pyranocoumarin isolated from Lomatium suksdorfii Couttt. Rose (Apiaceae family) was identified as a lead anti-HIV natural product through...
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Suksdorfin

L1

DCK, \( R_1 = R_2 = H \)

L2

4-MethylDCK, \( R_1 = Me, R_2 = H \)

L3

3-Hydroxymethyl-4-methylDCK, \( R_1 = Me, R_2 = CH_2OH \)

L4

DCP

L5

Figure 12 Structures of anti-HIV coumarins.

BDFI (112). Substitution of two camphanoyl esters for the acetate and isovaleroyl groups in the natural product lead to the extremely potent lead compound 3',4',R-di-O-(-)-camphanoyl-khellactone (DCK) (L2) (113). After additional synthetic modification to improve potency, 4-methyl DCK (L3) and then 3-hydroxymethyl-4-methyl DCK (L4) were found. The latter compound was selected as a clinical trial candidate (114).

Furthermore, a positional isomer of DCK, 3',4',R-di-O-(-)-camphanoyl-2',2'dimethylidydropyranoc2,3'-fichrsimeone (DCP) (L5) is even more promising because most DCP analogs are active against drug-resistant HIV strains, although DCK analogs are not. Adding an ethyl group at the 2-position of DCP decreased toxicity to cells compared with DCP, so, to date, the most likely clinical trials candidate in the DCP series is 2-ethyl DCP (L6) (114, 115).

The DCK and DCP compounds exert their antiviral activity by blocking the HIV reverse transcriptase (RT), however, at a later step than that affected by the clinically approved RT inhibitors, such as AZT. Thus, these compounds have a novel mechanism of action compared with current drugs. DCK and DCP compounds could be useful in the treatment of AIDS, although additional investigation is merited and needed (116).

Syzigium claviflorum (triterpene betulinic acid derivatives)

Triterpenes represent a structurally varied class of natural products existing in many plant species. Thousands of triterpenes have been reported with hundreds of new derivatives described each year. Some naturally occurring triterpenes exhibit moderate anti-HIV-1 activity and, therefore, provide good leads for additional drug development because of their unique mode of action and chemical structures. Research has identified anti-HIV triterpenes that block HIV entry, including absorption (glycyrrhizin) and membrane fusion (RPR103611); inhibit viral reverse transcriptase (RT) (minuscopic acid) and protease (ganoderiol, geumonoid); and act during viral maturation (DSB, see more description below) (117).

Two lupane triterpenes, betulinic acid (M3) and platanic acid (M2), from Syzigium claviflorum Wall. (Myrtaceae family) were reported first to reduce HIV IIIB reproduction by 50% in H9 lymphocytes (118). Afterwards, many derivatives were...
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Betulinic acid

Platanic acid

DSB (Bevirimat)

Figure 13 Structures of anti-HIV triterpenes.

Artemisinin

Artemether, R = CH₃

Sodium artesunate, R = COCH₂CH₂COONa

DZ-277 (RBx11160)

Figure 14 Structures of artemisinin and related compounds.

synthesized and studied for anti-HIV activity. Dimethyl succinyl betulinic acid (DSB, M3) was found to be the most useful candidate as an anti-HIV agent (119, 120).

DSB was discovered originally by the Natural Products Research Laboratories (NPRL), University of North Carolina, directed by the author (K.H.L.) (119), and then licensed to Panacos Pharmaceuticals, Inc. (Watertown, MA) for development. Panacos has named the compound Bevirimat and lists it as the lead antiviral product of the company.

During 2004, two Phase I studies and a Phase I/II study of M3 were completed. In the Phase I studies, the drug was well tolerated and showed good anti-HIV levels in the body. In the Phase I/II study, M3 showed activity in HIV-infected patients and significantly reduced viral blood levels (known as viral (load) (122, 123). A nther 2004 milestone was that the U.S. Food and Drug Administration (FDA) granted Fast Track Status for M3.

The Phase IIa study demonstrated the antiviral potency of M3 following once-daily oral dosing for 10 days in HIV-infected subjects not on other antiretroviral therapy. Viral load was reduced significantly compared with placebo. On day 11, following complete dosing, the median reduction at the 200-mg dose was a 91% decrease. In the Phase IIa trial, M3 was well tolerated, all adverse experiences were mild or moderate, and no dose-limiting toxicity was identified (122, 123).

Subsequently, studies have shown that M3 can be administered successfully in a tablet form rather than by an oral solution. A lbis, two drug interaction clinical trials of M3, in combination with the approved HIV drugs ritonavir and atazanavir,
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Protopanaxadiol-type

Protopanaxatriol-type

Ginsenoside $R_b1$ - $\text{glc}(2\text{-}1)\text{glc}$ - $\text{glc}(6\text{-}1)\text{arap}$

Ginsenoside $R_b2$ - $\text{glc}(2\text{-}1)\text{glc}$ - $\text{glc}(6\text{-}1)\text{araf}$

Ginsenoside $R_g3$ - $\text{glc}(2\text{-}1)\text{glc}$ - $\text{H}$

Ginsenoside $R_g5$ - $\text{glc}(2\text{-}1)\text{glc}$ - $\text{H}$

Figure 15 Structures of Panax ginsenosides.

have been completed and showed little likelihood of significant adverse drug-drug interactions when used in combination therapy (124).

In summary, M3 shows potent viral load reduction, a strong safety profile (with no evidence of organ toxicity or clinical intolerance), no evidence of clinically significant drug interactions, and, quite importantly, no evidence of rapid resistance development, which is a primary cause for antiretroviral treatment failure (125-127).

Phase II clinical trials began in 2006 and are still ongoing. One of the trial goals will be to determine an optimal dose of M3. These trials will involve HIV-infected patients who are failing current therapy and will be randomized, blinded, and placebo-controlled (124).

In Phase III clinical trials targeted for 2007/2008, combination therapy studies will be performed in a total of 300 to 500 patients at a commercial dose. The target for New Drug Application (NDA) is 2008/2009 (124). The efficient clinical trials progress of M3 continues to mark it as a leading new treatment for AIDS.

Active Compounds Isolated from Well-Known Folkloric Medicine

In addition to anticancer and anti-HIV agents, various types of active compounds that are active against other diseases and disorders (e.g., malaria, inflammation, and so forth) also have been isolated from natural sources, especially well-known folkloric medicine. These compounds and their plant sources are described below.

**Artemisia annua (qinghao, artemisinin derivatives)**

Qinghao (Sweet Wormwood) is the dried aerial parts of the herb *Artemisia annua* L. (Asteraceae family), which has been used in China for centuries to treat fever and malaria. Artemisinin ($\text{N}_1$) (Qing Hao Su) (128), the active principle, directly kills *Plasmodium falciparum* (malaria parasites) with little toxicity to animals and humans. Thus, it is a clinically effective, safe, and rapid antimalarial agent (129, 130). The novel endo-peroxide link is essential for the antimalarial activity.

Artemether ($\text{N}_2$) and arteether ($\text{N}_3$) are the most well-studied analogs among many synthetic derivatives and are used in malaria-prone regions, particularly India (131). Artemether and sodium artesunate (a hemisuccinate derivative of dihydroartemisinin) ($\text{N}_4$) have been added by the World Health Organization to its Model List of Essential Medicines (132).

The laboratory of the authors has synthesized analogs related to artemisinin (133). Recently, an antimalarial synthetic trioxolane drug development candidate called OZ-277 ($\text{N}_5$, also known as RBx11160) (134) has sparked great interest and has progressed to Phase II clinical trials in India, Thailand, and
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Figure 16 Structures of two different ginsenosides in *P. ginseng* and *P. quinquefolium*.

| Figure 17 | Structures of *P. notoginseng* natural products. |

Africa. Modification and pharmacological studies are ongoing (135–137). A recent review discusses artemisinin and related antimalarials (138).

Ginsengs: Asian, American, Sanchi, and Siberian

Ginseng is the root of *Panax ginseng* C.A. Meyer (Asian ginseng) (Araliaceae family). In Oriental medicine, it has enjoyed a strong reputation since ancient times for being tonic, regenerating, and rejuvenating. The genus name *Panax* is formed from the Greek pan (all) and akos (remedy). This panacea (panakeia) was believed to be the universal remedy. Wild ginseng is scarce and has been replaced by cultivated ginseng or "true" ginseng. Species include American ginseng (*P. quinquefolium*) L., cultivated in North America; Japanese ginseng (*P. japonicus* (Harms.) C.A. Meyer), widely distributed in Japan; San-chi ginseng (*P. notoginseng* (Burk.) F.H. Chen), reputed as a tonic and hemostatic in China; and Siberian ginseng (*Eleutherococcus senticosus* Maxim.).

Asian ginseng (*Panax ginseng* C.A. Meyer) Many compounds have been isolated from the root of Asian ginseng: polysaccharides, glycopeptides (panaxanes), vitamins, sterols, amino acids and peptides, essential oil, and polyalkynes (139–141). About 30 saponins (called ginsenosides) isolated from the root are dammarane triterpenoids, which generally contain three or four hydroxyl (OH) groups (a 3*β*,12*β*,20(S) trihydroxylated-type (protopanaxadiol-type) and a 3α,6α,12β,20(S) tetrahydroxylated-type (protopanaxatriol-type), which are attached to various sugars. The individual saponins (e.g., ginsenosides Rb1, Rc, Rg1, and Rh2) differ in the mono-, di-, or tri-saccharide nature of the two sugars attached at the C-3 or C-6 and C-20 hydroxy groups. In some cases, the C-20 hydroxy group is absent (e.g., in ginsenoside Rg5) or the C-12 hydroxy group also is attached to a sugar.

Traditionally, ginseng is used to restore normal pulse and remedy collapse, to benefit the spleen and liver, to promote production of body fluid, to calm nerves, and to treat diabetes and cancer. Regarding the last two effects, ginsenoside Rh2 has been found to lower plasma blood glucose in streptozotocin-induced diabetic rats (142), and recently, ginseng and its constituents have been studied for cancer prevention and anticarcinogenic effects against chemical carcinogens. Ginsenoside Rg3 and Rg5 were found to reduce significantly lung tumor incidence, and Rh2, Rg3, and Rh2 showed active anticarcinogenic activity (143). In addition, effects of ginseng on quality of life have been discussed (144).
American ginseng (Panax quinquefolium L.)
American ginseng contains almost the same components as Panax ginseng (139). Thus, it could be used for the same medical conditions as Asian ginseng. However, in Chinese theory, some differences exist: American ginseng is “cool” and is used mainly to reduce the internal heat and promote the secretion of body fluids, whereas Asian ginseng is “warm.” Correspondingly, differences in biologic activity also exist (140): American ginseng stimulates the production of human lymphocytes, whereas Asian ginseng does not have a significant effect, and Siberian ginseng enhances production.

The main chemical difference between Asian and American ginseng is in the presence or absence of ginsenoside Rf (146) and 24(R)-pseudoginsenoside F₁₁ (148). Ginsenoside Rf is found in Asian ginseng but not in American ginseng, and although 24(R)-pseudoginsenoside F₁₁ is abundant in American ginseng, only trace amounts are present in Asian ginseng (141, 145–147).

Sanchi ginseng [Panax notoginseng (Burk.) F.H. Chen]
This ginseng exerts a major effect on the cardiovascular system. It dilates the coronary vessels and reduces vascular resistance, which results in increased coronary flow and decreased blood pressure. Chinese traditional medicine used this ginseng to arrest bleeding, reduce blood stasis, and relieve pain. Recent studies have shown that, in the treatment of angina pectoris, this herb can produce a 95.5% improvement in symptoms. The herb usually can stop bleeding in cases of respiratory bleeding or vomiting of blood.

P. notoginseng contains saponins similar to those found in P. ginseng, both ginsenosides and notoginsenosides. Certain saponins of both types, including notoginsenosides K, R₁, and U (144, 145), showed immunological adjuvant activity (stimulated immune response against antigens) (148, 149). Another important bioactive constituent of P. notoginseng is the non-protein amino acid dencichine (146), which can increase platelets and stop bleeding (150).

Siberian ginseng (Eleutherococcus senticosus maxim.)
Siberian ginseng is harvested from its natural habitat in Russia and northeast China and has been used in China for over 2000 years. It is not a true ginseng like Panax ginseng or P. quinquefolia, but it does belong to the same Araliaceous family. Siberian ginseng has its own bioactive ingredients with unique and proven medicinal values. It possesses significant adaptogenic action (antistress and antifatigue) and is recommended as a general tonic. Because of its nontropic mechanisms of action, Siberian ginseng has a broad range of clinical applications. The root contains polysaccharides, phenolics (coumarins, lignans, and phenylpropanoids), and glycosides. Some members (e.g., eleutheroside K, O₁₃) of the latter group are specifically triterpenoid in nature, whereas others, including isofraxoside (eleutheroside B₁, O₂₀), glycosides of syringaresinol (e.g., eleutheroside E₁, O₂₀), and the ethyl ether of galactose (eleutheroside C, O₂₁), belong in a miscellaneous series. Eleutherosides E (O₂₀) and B (O₂₂, also called syringin) are two major glycosides and typically are used as marker...
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Figure 19 Structures of Ganoderma triterpenes.

compounds associated with bioactivity, particularly antifatigue action (151). A new lignan glycoside eleutheroside E₂ (O23) was isolated recently from E. senticosus (152).

Ganoderma lucidum (fungus)

Chinese people consider *Ganoderma lucidum* (Polyporaceae family) as the “Miraculous King of Herbs.” It is highly regarded for its medicinal properties, which include promoting the healing ability of the human body, strengthening the immune system, and increasing longevity. Accordingly, *Ganoderma* works in the treatment of cancer because it helps cleanse the body from toxins and strengthen the immune system. It also enhances liver detoxification, thus improving liver function and stimulating the regeneration of liver cells.

The chemical composition includes polysaccharides and lanosteroid triterpenes. In the former class, many polysaccharides have been linked to immune-stimulating effects (153–155). The latter class contains over 130 different triterpenoids with diverse pharmacological activities (156). Examples of the polyoxygenated triterpene structures are ganoderic acids B (P1), C₁ (P2), and D (P3), as well as ganoderiol F (P4), ganodermanon- diol (P5), and ganodermanontriol (P6).

Because of its widespread use as a health food as well as for medicinal purposes, the quality control of *G. lucidum* is quite important, and modern analytical methods are being established for the quantitative determination of major triterpenoids, including P₁–P₃ as marker compounds (157).

Cordyceps sinensis (Tung Chung Hsia Tsao)

Vegetable caterpillars are called “Tung Chung Hsia Tsao” in Chinese, which translates as “winter worm and summer grass.” They result from the infection of large underground caterpillars by a fungus. An entomopathogenic fungus grows in the larva of the sphinx moth in autumn. Although the larva hibernates underground through the winter, the fungus kills the infected host and grows throughout the cadaver. In summer, a rod-like stroma of the fungus grows out from the mummified shell of the dead host, looking like a sprouting, dark blue to black grass. The mycelia of *Cordyceps sinensis* (Berk.) Saccar. (Clavicipitaceae family) that colonize the larvae of *Hepialus armoricans* (Hepialidae family) are representative vegetable caterpillars and are highly valued in the Chinese traditional medicinal system.
Initial records of using vegetable caterpillars as medicine date back to the Ming dynasty in China and appear in "Pen-Tsao-Kang-Mu" (Compilation of Materia Medica) in 1596. This fungus is regarded as a popular and effective medicine for treating numerous illnesses, promoting longevity, relieving exhaustion, and increasing athletic power. More recently, their physiologic activities, including immunostimulating and antitumor effects, have encouraged their medicinal use.

In traditional medicinal practices, wild-harvested mycelia are considered to have higher therapeutic benefits and, therefore, command higher prices than cultivated fungus. Native occurrence of the fungus is confined to the highlands of the Himalayan region, in addition to some provinces of China with cold, arid environments. Several mycelial formation products grown in artificial media are available commercially as health food supplements in the United States and Canada.

A comprehensive review (158) recently discussed the chemical constituents and pharmacological actions of the Cordyceps species. Chemical constituents include cordycepin (Q1), 3′-deoxyadenosine and other adenosine analogs, ergosterol derivatives (Q2–Q4), and peptides, including cyclic peptides such as cordycepeptide A (Q5), and other adenosine analogs, ergosterol derivatives (Q2–Q4), and peptides, including cyclic peptides such as cordycepeptide A (Q5), and other adenosine analogs, ergosterol derivatives (Q2–Q4), and peptides, including cyclic peptides such as cordycepeptide A (Q5), and other adenosine analogs, ergosterol derivatives (Q2–Q4), and peptides, including cyclic peptides such as cordycepeptide A (Q5). These compounds likely contribute to the antitumor, antibacterial, antifungal, and anti-inflammatory activities. Regarding the latter activity, Q1 has been found to inhibit platelet aggregation (160). C. sinensis also contains polysaccharides, which account for anti-inflammatory, antioxidant, antitumor, antinflammatory, immunomodulatory, hypoglycemic, and hypolipidemic effects (158, 159). New diphenyl ethers (cordyols A–C, Q6–Q8) were discovered recently in Cordyceps (161). Q6 was associated with significant anti-HSV (herpes simplex virus) activity (161).

Conclusion

Based on the above successful examples, new drugs derived from natural product leads will be discovered continuously, and modern medicinal chemistry-based molecular modification will play an important role in developing the new leads into useful drug candidates. Highly efficient bioactivity-directed fractionation and isolation, characterization, analog synthesis, and mechanistic studies are prerequisites for the development of new plant-derived compounds as clinical candidates for world-class new medicines. Drug discovery also will benefit from the development of new biological targets and the continual improvement of bioassay technology (162–169). The development of new, effective, and safe world-class drugs from medicinal plants, which have been appreciated for centuries for treating illness, will be long lasting, as they are the best and most effective source for generating new medicines by use of modern scientific technology.
Acknowledgment

This investigation was supported by grants from the National Cancer Institute, the National Institutes of Health (NIH) (CA-17625), and the National Institute of Allergy and Infectious Diseases, NIH (AI-33066) that were awarded to K.H. Lee. This article is No. 257 in the series “Anti-AIDS Agents.” article is No. 71 in the series “Anti-AIDS Agents.”

References


---

18. John Wiley & Sons, Inc.
Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY

Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research

124. Data provided by Panacos, Inc.


Recent Progress in Drug Discovery Based on Plant-Derived Natural Products Research


See Also
Drug Discovery
Natural Products as Anticancer Agents
Natural Products in Plants, Chemical Diversity of Natural Products, Common Biological Targets of Natural Products: An Overview
Platinum Anticancer Drugs, Chemical Biology of

Jenny Zhang, School of Chemistry, The University of Sydney
Renee Whan, The Electron Microscopy Unit, The University of Sydney
Trevor Hambley, School of Chemistry, The University of Sydney
doi: 10.1002/9780470048672.wecb657

In the past, studies into the distribution of platinum anti-cancer agents inside cancer cells were experimentally challenging because of a lack of readily available and sensitive platinum mapping techniques. Recent advancements in the area of platinum mapping in cells have meant that the biologic fate of cisplatin and its derivatives can be studied in several new ways. In particular, electron microscopy, fluorescence microscopy, and synchrotron radiation induced X-ray emission have made significant contributions in the study of cisplatin in vitro. The applications of these three techniques and the insights they provide into the intracellular distributions and interactions platinum complexes are reviewed here. The strengths and weaknesses of each technique and their potentials for additional use in this field of research have been evaluated.

Introduction

Platinum-based anticancer drugs including cisplatin, carboplatin, and oxaliplatin are the most widely used and effective chemotherapeutic agents in the current arsenal (1, 2). They are believed to exert their action by binding to nuclear DNA that leads ultimately to the affected cell undergoing apoptosis (3). To reach the nucleus, the complex must traverse the cell membrane, the cytoplasm, and the nuclear membrane to avoid the detoxification mechanisms of the cell along the way. Less than 1% of the platinum that enters the cell successfully makes this journey and, consequently, the great bulk of the platinum suffers other fates (4, 5). Determining how the cell deals with platinum is important for understanding and for overcoming unwanted toxicities and inherent and acquired resistance, but is experimentally challenging.

Biologic Background

Although the cytotoxicity of cisplatin (and its derivatives) is accepted widely to be a result of its interactions with DNA, much is still unknown about the interactions of cisplatin on a cellular level. To begin with, the intracellular uptake of cisplatin is still the subject of much debate. Some studies have suggested that cisplatin diffuses into cells passively, as uptake has been observed to be not saturable, not dependent on pH, and not inhibited by coadministration of structural analogs (6, 7). Other studies have argued that some component of cisplatin uptake is carrier mediated, as cisplatin accumulation is modulated by ATP, membrane potential, ion concentrations, and several transportomes, which includes the copper influx transporter CTR1 and the ATP dependent copper efflux transporter ATP7B (8–15). Indeed, it has been observed that cisplatin accumulation is reduced in many cisplatin-resistant cell lines (16–18). Hence, the mode of platinum influx and efflux in cells plays an important part in drug resistance, and it needs to be understood better.

Another area that requires additional understanding is the role of cytoplasmic organelles in the fate of platinum drugs. Platinum has been shown to be localized within the Golgi apparatus, endoplasmic reticulum, mitochondria, and lysosomes, the cytosol, and the nucleus. However, only the consequences of nuclear localization have been defined satisfactorily (19–21). It has been observed that the Golgi apparatus and the endoplasmic reticulum change in morphology after cisplatin treatment. One explanation for this change is that protein translation has been inhibited after cisplatin localization in these organelles (22, 23). Another suggestion is that these organelles contain enzymes and proteins that are linked directly to mechanisms of cisplatin induced resistance, including glutathione-S-transferase and metallothins (21, 24, 25). Hence, the changes observed in these organelles may be caused by the participation of these enzymes in the deactivation of cisplatin.

It is well documented that the aquated form of cisplatin localizes within the mitochondria, and that the inhibition of respiration and spontaneous release of calcium ions into the...
Reduced X-ray emission. Through the use of three important imaging techniques: electron view and evaluate recent developments in the study of the fate of resistance, with several alterations that have been identified in cisplatin-resistant cell lines. These alterations include a decrease in the expression of proteins involved in the mechanisms of drug-resistance. In a study by Yang et al. (26), the morphology of the mitochondria was found to be changed significantly after one hour incubations with 50 µM cisplatin, suggesting that the damage done to the mitochondria is subsequent to lysosomal and nuclear localization, which suggests that this organelle has only a secondary role in cisplatin-mediated cytotoxicity (23, 26–30). The involvement of lysosomes in the action of cisplatin is also of interest. Cells that have been treated with cisplatin have an increased number and size of lysosomes, suggesting that they may play an important role in detoxification of the drug (8). In other studies, this organelle has been implicated in the mechanisms of resistance, with several alterations that have been identified in cisplatin-resistant cell lines. These alterations include a decreased number of the organelles, a more alkaline lumen, and altered recycling of the luminal contents between the endosomal and lysosomal compartments (31, 32). Interestingly, lysosomes have also been implicated as a potential site of induced cytotoxicity. Endogenous and/or exogenous stress can cause the membrane of the lysosome to permeabilize and release hydrolases. If this release is uncontrolled, it can lead to necrosis and selective release of two classes of hydrolases (sphingolipidases and cathepsins) that can trigger the so-called “extrinsic” pathway to apoptosis. A hallmark in lysosomal cathepsins (found in many cancer types) can degrade the extracellular matrix of the cell and lead to tumor metastasis (33–37).

Thus, the mode of platinum entry into cells and its subsequent intracellular pathways and interactions with subcellular organelles are defined poorly. Addressing where and how platinum drugs are processed in cells could greatly help elucidate the cause of unwanted toxicities as well. The phenomena of drug-resistance, ultimately improving drug design. In the past several techniques such as cell fractionation, atomic absorption spectroscopy, and light and electron microscopy were employed to study the distribution of cisplatin in vitro (8, 19, 21, 30). A lack of readily available, accurate, and sensitive platinum-imaging techniques limited the research in this area. In the past decade, new and improved techniques have emerged for the detection of platinum in cells. Here, we review and evaluate recent developments in the study of the fate and distribution of platinum complexes in cancer cells, achieved through the use of three important imaging techniques: electron microscopy, fluorescence imaging, and synchrotron radiation-induced X-ray emission.

Platinum Imaging Techniques

Transmission electron microscopy

Transmission electron microscopy (TEM) has been used extensively in biology for direct visualization of ultrastructural details and platinum deposits in cells. The underlying physics behind TEM is similar to that of ordinary light microscopy; however, the resolutions achieved by TEM can be some 400-fold greater than that of light microscopy (38). Briefly, the mechanics behind TEM involves an illuminating source, the electron gun, that sends a beam of electrons through a vacuum and onto the specimen of interest. An image is formed according to the differential absorption of electrons by various parts of the specimen; this image is amplified via a series of electromagnetic or electrostatic lenses and is projected onto a fluorescent screen. All manufactured transmission electron microscopes guarantee a resolution of 0.344 nm, TEM is particularly useful in cellular biology, where resolutions of around 2 nm (the width of one third of a mitochondrial membrane) are needed (38, 39).

The preparation of biologic samples for TEM involves several steps. First, the sample is fixed, usually chemically, to preserve the structural integrity and the spatial information of the cells while simultaneously terminating any biologic activity. After fixation, the sample is dehydrated and embedded in an epoxy resin, then sectioned into thin slices (that range from 80 to 800 nm) and stationed onto a grid, ready for examination by the electron microscope. For specific staining of cellular components, immunolabeling can be employed; the contrast between cellular components or the background can be heightened by staining with heavy metal salts, such as those of uranium, lead, and osmium, all of which have sufficient mass and electron density to scatter the electron beam (40).

The high resolution attainable by TEM has led to its use in biologic platinum studies in two ways; to reveal ultrastructural morphologic changes in drug treated biological samples and to map directly the platinum moieties in cells. Several studies have used TEM to examine morphologic features of organ tissues, sarcoma, and tumour cells, before and after cisplatin treatment to gain insights into its toxicity and mode of action (41, 42). Among the more recent findings from these fine-structure studies was the observation that the morphology of rat kidney tissue changes after cisplatin treatment in vivo (8). It was observed that cisplatin induced a large increase in the number and the size of lysosomes in the cells of the kidney, which suggests that lysosomal sequestration is a mechanism used by the kidney for the removal of platinum and other heavy metals (8, 43). Other similar fine-structure studies revealed significant changes in mouse kidney morphology after cisplatin incubation, with observed features such as necrotic proximal tubules, damaged tubular basement membrane, and short, irregular microvilli (44).

In in vivo studies, Yang et al. (26) found that the morphology of the mitochondria to be changed significantly after one and 4 hour incubations with 50 µM cisplatin on the head and neck squamous cell carcinoma cell line (HNSCC). The mitochondria showed swelling, distortion of the cristae as well as the inner and outer membranes, whereas the nucleus and nucleoli remained unchanged. This finding led to the authors to challenge the notion that the site of action of cisplatin is solely with nuclear DNA, which suggests that cisplatin could act directly on the mitochondria to cause the release of cytochrome c trigger in the apoptotic cascade (26). In another study conducted by Mejra et al. (45) on human small cell lung carcinoma cancer cells (GLC-C), a high nuclear to cytoplasmic ratio was observed after 4 hour incubations with 333 µM cisplatin. Furthermore, similar cisplatin treatments on the cisplatin resistant sub-line, GLC-CDDP, were shown to cause swelling of the Golgi apparatus and an increase in the numbers of mitochondria (45).
Early attempts to map the distribution of the electron-rich platinum moieties inside cells employing TEM have met with some success, with platinum(II)-pyrimidine complexes being amongst the first platinum species to be mapped (46). In this study, it was observed that cells stained with platinum(II)-thymineuracil complexes, prepared as 1% solutions, showed platinum staining in the chromatin, nucleolus, and ribosomes with high selectivity. However, it should be noted that the concentration present in a 1% platinum complex solution is much greater than those in normal experimental dosages. In a separate study by Kh and Sadler (47), gray electron dense areas were also observed to be localized primarily within the nucleus and nucleolus of human epithelial carcinoma (HeLa) cells after 4 hour incubations with high concentrations of cisplatin (200 µM). These gray areas were later confirmed by electron-probe X-ray analysis to contain platinum (47).

In later TEM studies by Ghadially et al. (48), a lower concentration of cisplatin (30 µM) was incubated in HeLa and human lymphoblastoid (RPMI 6410) cells for periods that ranged from 1 hour to 4 days. No intracellular platinum was detected in either cell line. However, similar incubations with platinum(II)-uracil resulted in development of lysosome-like bodies in the cytosol, which are referred to as "platinosomes," that contain electron dense species identified by X-ray analysis as platinum (48). In a related study, similar concentrations of cisplatin were injected into rabbit knee joints and incubated for 30 minutes and observed electron dense spots, identified by the employment of a GPt antibody stain. This stain could be performed by counting the gold labels per cm² of area of interest on the images of three or more individual cells. It was found that the distribution of Pt-DNA adducts was the same for both GLC4 and GLC4-CDDP cell lines, with adducts detected in the nucleus (preferentially at loci with high-density chromatin), and in the mitochondria DNA. Platinations were found to be high-est in apoptotic and dividing cells (45). It should be noted that again, high cisplatin incubation concentrations were required to acquire accurate quantitative Pt-adduct formation data.

Overall, TEM has proven to be a valuable tool for biologic platinum studies. It can provide qualitative information, with excellent resolution, regarding the intracellular sites of platinum interactions and accumulations. Quantitative information regarding certain structures, such as Pt-adducts, can now be acquired. Despite these advantages, limitations exist in the use of TEM as a mapping technique for platinum. TEM cannot attain visual information of living samples, and the process of fixation, dehydration, and resin embedding can significantly affect the distribution of the platinum species inside cells. Furthermore, TEM lacks the sensitivity to detect single platinum moieties accurately, mainly because of the damage caused to the ultrastructural sample sections by the high-energy beams (53). Hence, TEM is limited to the detection of clusters or deposits of platinum metals; sparse distributions of platinum remain undetected. Additionally, the cisplatin incubation concentrations required for the detection of platinum deposits and Pt-adducts in cells are much greater than normal experimental dosages. This finding could produce misleading results in relation to the intracellular distribution of platinum.

Fluorescence imaging

Currently, fluorescence microscopy is one of the most useful techniques for the examination of drug localization in fixed and live cells. However, to use this technique, the drug must be fluorescent, or labeled with a fluorescent tag. Two forms of fluorescence microscopy are observed in the literature: wide-field epifluorescence microscopy and later scanning confocal microscopy. Briefly, the mechanics behind both techniques involves irradiation of the sample with high-energy photons, which excite appropriate fluorophores to high-energy states. After excitation, the fluorophores undergo relaxation in the forms of nonradiative transitions between excited states, and this energy can be lost as heat to the solvent. Fluorescence results as the fluorophore relaxes even more back to its ground state that emits a photon with a longer wavelength than the original excitation photon. In both techniques, this process is performed on a light microscope platform. The two fluorescent techniques differ in how the resultant fluorescence is collected and initially excited. In wide-field epifluorescence microscopy, the sample is illuminated using a high-energy light source, such as a Hg or Xe lamp, and the emitting photons are collected and amplified using a charged coupled device camera. The resulting images can be in "red color;" however, this technique collects emitted fluorescence across a large depth of field, including light outside the focal plane, which limits axial resolution to ~1 µm. In confocal microscopy, the sample is excited by scanning with a laser, and the emitting photons are collected and amplified using photon multiplier tubes. The resulting images are in "gray scale;" however, the additional employment of a pinhole positioned before the detector eliminates out of focus light, which enables resolutions of around 0.1–0.2 µm in the XY plane. This method allows the sample to be "optically sectioned" without
physical sectioning, which enables three-dimensional images to be created (52, 53).

Several advantages exist in using fluorescence microscopy in drug distribution studies. Unlike TEM, fluorescence microscopy is a noninvasive technique that can acquire three-dimensional images of living cells in real time with minimal preparation. Additionally, the spatial distribution of fluorescence within cells can be detected with high sensitivity and signal specificity; hence, both qualitative and quantitative data can be acquired at relatively low drug incubation concentrations. Furthermore, intracellular organelles can be labeled with fluorescent organelle-specific stains to identify the sites of drug accumulation in co-localization experiments (52, 54).

Recently, platinum agents have been labeled with various fluorescent tags for fluorescence microscopy studies. The earliest work of this kind was conducted by Molenaar et al. (55), who tagged a cisplatin derivative with a carboxyfluorescein diacetate (CFDA) moiety. The CFDA ligand is a nonfluorescent derivative of fluorescein and the CFDA-Pt complex becomes fluorescent following cellular uptake and acetate hydrolysis. In this study, human osteosarcoma cells (U2-OS) were incubated for 30 minutes in 10 µM CFDA-Pt. Using wide-field epifluorescence microscopy, the complex was observed to distribute initially throughout the entire cell, and then accumulate in the nucleus to form fine granular patterns at 3 hours after incubation. After 8 and 72-hour incubations with the complex, colocalization experiments employing a Golgi fluorescent probe. After 24 hours, no significant fluorescence was observed in the nuclei of the cells, which suggests that the fluorophore had undergone efflux from the cell and the Golgi. The fluorescence distribution of the free CFDA fluorophore proved to be very different from the complex, which suggests that the platinum complex is not mimicking the behavior of the fluorophore (55). A later study by the same group (56) found that the treatment of biologically active dinuclear platinum complexes labeled with the same CFDA moiety in U2-OS cell lines shows very similar localization to the mononuclear CFDA-Pt complex (56). Therefore, fluorescence microscopy has allowed for the intracellular pathway of CFDA tagged platinum complexes to be deduced and subsequently identified as a significant role in the cellular processing of these types of complexes.

In a similar study, Safaei et al. (57) studied the distribution of a fluorescein-labeled cisplatin analog (FDDP) using epifluorescence microscopy in fixed human ovarian carcinoma cells (2008). This study showed the complex was localized within the cytoplasm after a 5 minute incubation with 2 µM F--DDP. After 10 to 30 minutes, fluorescence was observed in the nuclear regions. Like the Pt-CFDA complex, F-DDP also displayed fine granular patterns in the nucleus, and it was distributed in discrete vesicular structures, not localized diffusely throughout the cytoplasm. Staining with organelle-specific markers revealed that F-DDP was sequestered into the lysosomes, Golgi, and secretory compartments of the cells. The roles of these subcellular organelles in the processing of the complex were investigated by the use of various pathway inhibitors. These biologic tools in conjunction with fluorescence microscopy revealed that the secretory pathways, involving the Golgi and lysosomes played an important role in intracellular trafficking of these tagged complexes, particularly with regard to drug efflux (57).

In another fluorescence microscopy study, the trafficking pathways of a fluorescently tagged Alexa Fluor-platinum complex in a cisplatin-sensitive human KB epidermoid carcinoma cell line (KB-3-1) and its cisplatin-resistant subline (KB-CP-r) were compared. During the first 30 minutes of the incubation, the fluorescence from the complex moved from the membrane and into the Golgi apparatus; after 2 hours, the complex was distributed throughout the cell, including nuclear compartments. Internalization of the fluorescent complexes was very different in the cisplatin resistant subline, where the overall accumulation of the complex was lower, as determined by decrease of cellular fluorescence by flow cytometry. Only punctate cytoplasmic staining was observed after incubations, with little fluorescence in the cell nucleus. The altered distribution suggests that cellular resistance could be caused by modified trafficking, possibly involving the failure of cell membrane proteins that bind cisplatin (58).

From these studies, it can be observed that fluorescence microscopy can indeed be a powerful technique for studies into the biologic fate of platinum anticancer agents. It not only provides qualitative and quantitative distribution data, but also offers the option of real-time imaging, which reveals cellular processes in action. Despite the usefulness of fluorescence imaging in distribution studies, inherent disadvantages exist with this technique. For example, it is difficult to ascertain whether the fluorescence observed is from the fluorophore-platinum complex or the fluorophore alone (should cleavage occur in the cell and the distribution will differ from the properties of the platinum complex alone, which affects such factors as cellular uptake (59, 60). Furthermore, a highly conjugated, aromatic fluorophore may act as an intercalator, which generates different reaction profiles with DNA. Hence, caution should be used when evaluating the distributions of fluorescence in cells.

The disadvantage associated with altered distribution because of the fluorescence tag does not apply when the drugs concerned are intrinsically fluorescent platinum species, such as those of the platinum-anthraquinone complexes. Platinum-anthraquinone complexes were designed to increase the DNA affinity of the platinum moiety by its chemical linkage to a bio-carrier, in this case an intrinsically fluorescent DNA intercalator (61, 62). In a series of time-lapse fluorescence microscopy studies conducted by Reedsjö et al. (63, 64), the cellular distribution of a group of dinuclear cationic platinum-anthraquinone complexes was observed in live parental A2780 cells. At the low incubation concentration of 3 µM, the complexes were observed to enter the cells slowly, with fluorescence predominantly being localized within the cytosol after 2 hours. After 24 hours, the complexes were localized in the lysosomes whereas the free ligand was found to be additionally localized in the nucleus. In contrast, studies using the same complexes in the cisplatin resistant A2780 subline show that the complexes were encapsulated by the lysosomes at the beginning of their incubation.
which suggests that lysosomal sequestration plays an important role in the resistance profile of A2780 cells. However, additional DNA titration studies found that the fluorescence of the anthraquinone was quenched in the presence of DNA; hence, any nuclear localization of the drug may not have been detected (63, 64).

Additional studies using the same complexes in both cisplatin sensitive and resistant forms of U2-OS cells also showed differences in the cellular processing of the drugs; both the free ligands and platinum complexes were rapidly found to accumulate in the nucleus, after uptake into the Golgi, which was hypothesized to be involved in transport of the platinum complexes out of the cells (56). In this set of studies, fluorescence imaging helped to illustrate the differences in the intracellular interactions of dinuclear platinum-anthraquinone complexes in different cell lines and in the different resistance profiles of A2780 and U2-OS resistant sub lines.

In a separate fluorescence microscopy study, the distribution of a mononuclear neutral platinum-anthraquinone complex, Pt1C3 (1), was examined in human colon adenocarcinoma cells (DLD-1) and A2780 cells (see Fig. 1). In both cell lines, fluorescence was observed to be predominantly localized within lysosomes after 1-, 4-, and 24-hour incubations with the complex, with no colocalization with mitochondrial and nucleic acid fluorescent probes. These results were in conflict with the high cytotoxicity profile observed for the platinum complex and cellular uptake studies that showed high platinum content within the nucleus after 4 hour incubations (65). Several possible reasons exist for the lack of fluorescence detected in the nucleus.
The amount of DNA binding may be too low to detect by confocal techniques. Alternatively, the platinum cytotoxin could be cleaved from the fluorescent anthraquinone part of the complex and is still entering the nucleus. Last, quenching of the molecules fluorescence could have occurred through reaction with an endogenous molecule such as DNA. Hence, although the bio-distribution of these intrinsically fluorescent drugs will not be altered because fluorescence tagging is unnecessary, limitations still exist in the application of fluorescence imaging in the mapping of these intrinsically fluorescent platinum complexes in cells.

X-ray fluorescence microscopy

A recently embraced technique in the monitoring of platinum in cells is that of synchrotron radiation induced X-ray emission (SRIXE). SRIXE is essentially synchrotron-based X-ray fluorescence microscopy and allows for the accurate two-dimensional analysis of elemental distribution within biologic samples (provided that the element is heavier than fluorine). The high intensity of X-rays derived from third generation synchrotron sources gives rise to the high sensitivity of platinum detection, with detection possible down to 5–10 ppm, and the high spatial resolution of about 0.15 × 0.15 µm² (66–68). At this level of elemental sensitivity, SRIXE has advantages over electron microscopy because the detection of platinum is possible at much lower drug concentrations. Furthermore, the location of the unmodified platinum complex can be monitored directly using SRIXE, and it is not merely inferred indirectly by the imaging of chemically modified analogs of the drug as done in fluorescence microscopy studies. Additionally, it is not necessary to fix or to section samples for SRIXE because of the highly penetrating nature of hard X-rays (photons with energy > 1 keV), and the samples can be studied relatively close to their natural hydrated state (with some cryogenic treatments) (69).

Briefly, the principle behind SRIXE involves the focusing of a beam of hard X-rays from a synchrotron source to a submicron area section on a sample. When absorbed by core electrons of the elements in the sample, these photons result in ejection of these electrons. Electrons from a higher energy orbital move down to fill the vacancies left, which results in the release of photons with energy that is equivalent to the difference in the binding energies of the two shells. This X-ray fluorescence is detected and converted into an X-ray spectrum, conveying characteristic peaks (see Fig. 2) (66). The identity of the elements can be deduced from the energies of the peaks, and the integration of the peaks correlates to the quantities of the elements present. By this method, many elements can be identified concurrently, which allows maps of the distribution of these elements to be generated at submicron resolution (66–68).

So far, studies that use SRIXE to map the distribution of platinum inside cells have been limited, and all studies have been conducted at low resolution. The first of these studies was conducted by Ilinski et al. (68), who used SRIXE to compare the uptake of platinum(II) complexes in the cisplatin-sensitive 2008 cell line and its cisplatin-resistant subline. In this study, incident X-ray energy of 11.7 keV (just above the platinum 2p3/2 absorption edge) was used, and Lβ fluorescence lines were collected with resolutions of 1 × 0.2 µm². It was observed that cisplatin accumulation in the cisplatin-resistant subline was half that of the cisplatin-sensitive cell line after 24-hour incubation with 10 µM cisplatin. In contrast, Pt103, a cisplatin derivative that displayed high activity against cisplatin-resistant tumors, showed a higher accumulation after the same incubation period. This suggests that increase of Pt103 cellular uptake was responsible for its bypass of cisplatin resistance mechanisms in cisplatin-resistant 2008 cells. However, the accuracy of this experiment was compromised because of the overlap of the fluorescence peaks between Pt Lα and ZnKβ when 11.7 keV is used as the incident energy (68). Regardless, the potential of SRIXE to map the distribution of platinum quantitatively within individual cells was demonstrated.
Following this, Hall et al. (60, 70) used SRIXE to examine elemental distribution within A2780 cells that had been treated with cisplatin and platinum(IV) complexes. This study employed incident energy of 12.6 keV (above the Zn Kα absorption edge), and emissions from both Kα and Lα fluorescence were collected at a resolution of 0.3 × 0.3 µm². It was observed that cisplatin and the platinum(IV) analogs accumulated in the nucleus of the cells after 24-hour incubations with 20 µM solutions (note that zinc elemental maps were used to identify the cell nucleus; see Fig. 3). This finding suggests that most platinum(IV) species had been reduced, and it was confirmed qualitatively by collecting micro-XANES spectra at points of high platinum concentration within the cells. The observation was also confirmed quantitatively by obtaining XANES spectra of bulk cell samples and quantifying the percentage of remaining platinum(IV) (60, 70). Hence, this study confirmed that SRIXE can be used for qualitative and quantitative studies of platinum accumulation inside cells, and it showed that the coupling of SRIXE with XANES affords a powerful technique in which both the distribution and oxidation state of a drug within a biologic system can be determined.

In later studies, Hall et al. (71) employed the same method as above to study the intracellular distribution of platinum complexes that carry bromine labels (2, 3) or anthraquinones (1). Again, after 24 hours, all complexes were found to be localized within the nucleus, with little platinum being in the cytosol. In the case of the platinum(II)-3-bromopyridine complex (2), the bromine was also observed to colocalize with the platinum. This finding is consistent with the amine moieties that remain bound to platinum(II) complexes inside cells. In another case of the platinum(IV)-bromacetate complex (3), the distribution of the bromine was much more diffuse when compared with that of the platinum, which is consistent with the reduction of the platinum(IV) complex. However, whether the reduction occurred before or after cell entry is still unclear (see Fig. 4) (71). Hence, by this method of labeling, the bio-transformation of platinum(IV) agents can be monitored via SRIXE.

SRIXE has also been recently employed to investigate cisplatin resistance. Shimura et al. (72) has measured the quantitative difference in cisplatin accumulation between lung carcinoma cells (PC-9) and its cisplatin-resistant subline. Using an incident energy of 15 keV and a resolution of 1.5 × 0.75 µm, it was found that at 12 hour incubations using 1 µM cisplatin, the intensity of platinum in the resistant subline was 2.6 fold less than observed for the sensitive cell line (72). By using the same method as Ilinski et al. (67), Chen et al. (73) mapped the distribution of cisplatin in intrinsically cisplatin-resistant human melanoma cells (MNT-1) and in human epidermoid carcinoma cells (KB-3-1). This study found that cisplatin was sequestered into the melanosomes, and nuclear accumulation was reduced.

Although SRIXE is relatively new to cellular platinum studies, it has proven to be an excellent and reliable technique for the detection of platinum drugs in cells. Currently, success of elemental distribution studies using SRIXE is limited by the comparatively poor resolution. In the above studies, organelles, such as mitochondria, Golgi, endoplasmic reticulum, and lysosomes were not resolved because many of these subcellular organelles and structures are smaller than 0.1 µm (52). Hence, SRIXE in platinum studies currently lacks the high resolution found in TEM and fluorescence microscopy studies.

**Conclusion and Future Applications**

Advancements in imaging techniques, as illustrated above, has facilitated the significant progress made by studies of the intracellular distribution of platinum anticancer agents, providing deeper insight into the cellular fate of these drugs. All techniques mentioned have the potential to make even more valuable contributions biologic studies of platinum complexes. In terms of imaging, TEM provides the highest level of two-dimensional morphologic spatial resolutions. Additional developments into immunoelectron microscopy may allow for the intracellular mapping of more platinum–protein structures.
Figure 4  SRIXE maps showing the distribution of Br (left column) and Pt (right column) in A2780 whole cancer cells treated with cis-(PtCl2(3-Brpyr)(NH3)) (top row) or cis, trans, cis-(PtCl2(OAcBr)2(NH3)2) (bottom row). Figure reproduced with permission from Reference 71 with kind permission from Springer Science and Business Media.

Fluorescence microscopy uniquely allows for real time, three-dimensional tracking of drugs in live cells at low incubation concentrations. When employed in conjunction with tools such as pathway inhibitors, this technique can elucidate drug trafficking pathways and reveal the roles of intracellular components in the drug action.

SRIXE is a powerful technique that can map the distribution of platinum with very high sensitivity and at low incubation concentrations of the unmodified platinum drug. With future improvements to its resolving power, this technique has the potential to obtain unprecedented information on the distribution of both endogenous and exogenous elements at a sub-cellular level.

In addition to the advancement of the mentioned techniques, the future of the imaging of cellular platinum may also benefit from alternative techniques, such as that of secondary-ion mass spectrometry and autoradiography. However, these techniques have yet to be used in investigations of the in vitro distributions of platinum complexes.

References
16. Kalayda GV, Jukunen BA, McLean CA, Weisfeld P, Tate HJ, Reidel J. Dinuclear platinum complexes with N,N’-bis(2-aminophenoxy)tetradecammine-alkyl-1,4-diaminanatoruanones as linking ligands. Part II. Cellular processing in A2780 cisplatin-resistant human ovarian


Platinum Anticancer Drugs. Chemical Biology of


Programmed Cell Death as a Therapeutic Approach

Kurt Deshayes, Genentech, Inc., Department of Protein Engineering, South San Francisco, California
doi: 10.1002/9780470048672.wecb472

Although turnover rate varies greatly with cell type, the survival of multicellular organisms requires the constant renewal of healthy cells and the removal of pathogen-infected or damaged cells. A critical component of homeostasis is the proper regulation of programmed cell death or apoptosis. Both oversensitivity of crucial cells and resistance of malfunctioning cells to apoptotic stimuli threaten survival. A lack of death response permits mutant cells to survive normal death stimuli and potentially proliferate. Conversely, excess apoptosis removes cells required for organism survival. Restoring homeostasis by controlling apoptosis is a potentially powerful technique for alleviating disease.

Biologic Background

Programmed cell death allows the removal of overpopulated cells, pathogen-infected cells, malformed cells, or cells that have sustained genetic damage (1). Several disease states are linked either to a lack of apoptotic response or to unwanted cell death (2) (see Table 1).

When activated by cellular stress (e.g., hypoxia, DNA damage, or nutrient withdrawal), programmed cell suicide is initiated by the intrinsic apoptotic pathway (Fig. 1). A key step, the release of cytochrome c through changes in the mitochondrial outer membrane permeabilization, is regulated by the Bcl-2 protein family (3). Cytochrome c in combination with the Apaf-1 protein catalyzes the processing of procaspase-9 (cysteinyl, aspartate-specific protease) within the apoptosome (4). Caspase-9 initiates hydrolysis of zymogen procaspase-3 and procaspase-7, which, when activated, lead to the degradation of cellular targets and eventual cellular suicide (5). Caspase activity is regulated by Inhibitor of apoptosis (IAP) proteins, which are negative (prosurvival) apoptosis regulators (6). Conversely, the IAP proteins are antagonized by a second mitochondrial activator of caspases (Smac) protein and released from the mitochondria alongside cytochrome c, which serves as a positive (proapoptotic) apoptosis regulator (7).

Another important mechanism for promoting programmed cell death is the binding of ligands to the death receptors, which occurs in the extrinsic pathway (8) (Fig. 1). The death receptors recruit and activate caspase-8, which in turn regulates effector caspase-3 and caspase-7. Caspase-8 processes the Bcl-2 family member Bid, which collaborates with other members of the Bcl-2 family to induce cytochrome c release from the mitochondria and thereby activates the downstream intrinsic pathway (9).

Changes in the balance between the proapoptotic and prosurvival signals modulate the response to apoptotic stimuli. Cancer cells are a prominent example of how overexpression of prosurvival factors leads to a harmful repression of programmed cell death. In contrast, neurodegenerative disorders (Parkinson’s and Alzheimer’s diseases) have been suggested to result from excess cell death in slowly regenerating neurons, which leads to a loss of function (2).

Presumably, targeting the apoptotic pathway can be a means to treat cancer by antagonizing the prosurvival components or by agonizing the proapoptotic components. Recent positive results with agents that target the apoptosis pathway have increased interest in cancer therapies directed at this pathway. It is quite possible that antagonizing selected proteins within the apoptosis pathways may not be sufficient always for broadly applicable therapy. However, the connectivity between the pathways can be exploited by the judicious application of multiple agents that act in concert, perhaps synergistically, on the different components of the pathway to give broad therapeutic benefit.

When thinking about proapoptotic therapies, agents can be designed that trigger a cascade of enzymes, such that once the proteins have been activated the process will continue without the help of the therapeutic agent. This is strikingly different from a therapy that relies on enzyme inhibition for their therapeutic effect. The latter often requires chronic dosing to maintain continued enzyme inhibition. Apoptosis-inducing reagents may require only doses sufficient to overcome initial apoptotic roadblocks. Furthermore, fast-growing solid tumors often are hypoxic and nutrient-starved, conditions that would lead a normal cell into apoptosis. To avoid programmed cell death, cancer cells keep apoptosis in check through the overexpression of prosurvival factors. The studies discussed in the following sections
Programmed Cell Death as a Therapeutic Approach

Table 1

Diseases Associated with Programmed Cell Death

<table>
<thead>
<tr>
<th>Diseases Associated with Downregulated Apoptosis</th>
<th>Diseases Associated with Upregulated Apoptosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cancer</td>
<td>1. Neurodegenerative disorders</td>
</tr>
<tr>
<td>A. Hormone-dependent tumors</td>
<td>A. Alzheimer’s disease</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>B. Parkinson’s disease</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>C. Cerebellar degeneration</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>D. Amyotrophic lateral sclerosis</td>
</tr>
<tr>
<td>B. Follicular lymphomas</td>
<td>E. Retinoid pigmentation</td>
</tr>
<tr>
<td>C. PS3 mutant carcinomas</td>
<td></td>
</tr>
</tbody>
</table>

2. Autoimmune diseases                         | 2. AIDS                                        |
| A. Lupus                                      |                                               |
| B. Glomerulonephritis                         |                                               |

3. Viral Infections                             | 3. Ischemic injury                             |
| A. Herpesviruses                               | A. Stroke                                     |
| B. Poxviruses                                  | B. Myocardial infarction                      |
| C. Adenoviruses                                | C. Reperfusion Injury                         |

4. Neurodegenerative disorders                 | 4. Toxin-induced liver disease                |
| A. Alzheimer’s disease                         |                                               |
| B. Parkinson’s disease                         |                                               |
| C. Cerebellar degeneration                     |                                               |
| D. Amyotrophic lateral sclerosis               |                                               |
| E. Retinoid pigmentation                       |                                               |

5. Aplastic anemia                              | 5. Aplastic anemia                             |

*Adapted from Reference 2.

suggest that cancer cells are primed for programmed cell death and, therefore, are more sensitive to the induction of apoptosis.

Targeting the Bcl-2 Family of Proteins

It was first observed in type 2 B-cell lymphoma that the translocation of the Bcl-2 gene leads to radical overexpression of Bcl-2 (10). The overproduction of Bcl-2 has been shown to be transforming (10). Members of the Bcl-2 protein family contain as many as four characteristically helical Bcl-2 homology motifs (BH1–BH4). Bcl-2 proteins are divided into classes, as pro-survival proteins (Bcl-2, Bcl-xL, Mcl-1, A-1, and Bcl-w), proapoptotic proteins (Bak and Bax), and proapoptotic proteins that contain only the BH3 motif (Bim, Bid, Puma, and Noxa) (3, 11). The manner in which the different Bcl-2 family members interact and the mechanisms by which cytochrome c release is initiated are unclear, but the central role of the Bcl-2 family in apoptosis, and the therapeutic potential of regulating these proteins, is well established.

Changes in the ratio of prosurvival to proapoptotic Bcl-2 proteins regulate cellular response to apoptotic stimuli. For example, decreasing the production of prosurvival proteins will sensitize cells to apoptosis (12). Genta, Inc. (Berkeley Heights, NJ) is investigating this method as a potential cancer therapy.

Another potential therapeutic approach is to mimic the proapoptotic BH3 only proteins with small molecules that bind to the prosurvival members of the Bcl-2 family and promote apoptosis. This interaction involves a large hydrophobic interface, which is a challenging target for a small-molecule antagonist. Nonetheless, progress in the development of Bcl-2 antagonists has been reported. For example, a natural product, Gossypol, isolated from cottonseed oil, was shown to have spermicidal activity, and tested originally as a potential male contraceptive (15). Gossypol is believed to bind to the prosurvival Bcl-2 proteins (16) (see Fig. 2). The cytotoxic effects of Gossypol have been redirected as a potential anticancer therapeutic. Ascenta (San Diego, CA) has moved the R enantiomer of Gossypol, AT-101, into phase II human clinical trials. In these trials, the pan-specific Bcl-2 inhibitor has been combined with Rituxan (Genentech, Inc., South San Francisco, CA) in the treatment of relapsed CLL (chronic lymphocytic leukemia) (17). Recent results with MEFs (mouse embryo fibroblasts) that express neither Bax nor Bak show that Gossypol can kill cells independent of the proapoptotic Bcl-2 family proteins, which suggests that another mechanism is responsible for Gossypol-induced cell killing (18).

Other potential therapeutic approaches involve the use of small-molecule BH3 mimetics. Such agents are believed to antagonize the prosurvival Bcl-2 members by mimicking the action of the BH3 domain. The Abbott small molecule (ABT-737) was developed using the SAR (structure activity relationship) by NMR (nuclear magnetic resonance) method of Dr. Stephen Fesik (19). This molecule binds with subnanomolar potency to members of the prosurvival Bcl-2 family, Bcl-2, Bcl-xL, and Bcl-w, but does not bind to the other widely expressed member of the family, Mcl-1. In vitro and in vivo studies show ABT-737 has excellent...
Programmed Cell Death as a Therapeutic Approach

Figure 1  The extrinsic apoptotic pathway is triggered when death receptors are engaged by their cognate ligands, which results in recruitment of the adaptor protein FADD and the apical caspase, caspase-8. This recruitment leads to the activation of caspase-8 and subsequent activation of the effector caspases 3 and caspase-7. The intrinsic apoptotic pathway is triggered by stimuli, such as irradiation, chemotherapeutic agents, or growth factor withdrawal. Activation of proapoptotic Bcl-2 family members by the p53 pathway neutralizes the antiapoptotic proteins Bcl-2, Bcl-xL, and Mcl-1, which leads to the disruption of the mitochondrial membrane potential and to the release of cytochrome c and Smac into the cytoplasm. These events result in Apaf-1-mediated activation of caspase-9 and subsequent activation of the effector caspase 3 and caspase-7 and then culminate in cell suicide. XIAP is the last line of defense against cellular suicide and acts by inhibiting caspases. Other IAP proteins (ML-IAP, cIAP1, and cIAP2) do not inhibit caspases directly but sequester the proapoptotic molecule Smac, which prevents it from blocking the action of XIAP. Smac mimetics bind to IAP proteins and block their inhibitory activity by antagonizing the critical IAP–caspase and IAP–Smac interactions.

Targeting the Inhibitor of Apoptosis Proteins

IAP proteins prevent cell death through interactions between their BIR (baculoviral IAP repeat) domains and the proteases that are critical for the initiation and execution of apoptosis, caspase-3, caspase-7 and/or caspase-9 (23). X-chromosome-linked IAP (XIAP) is a ubiquitously expressed IAP protein and a potent inhibitor of caspases that plays a critical role in resistance to chemotherapeutic agents and other proapoptotic stimuli. Although no direct genetic mutation defines XIAP as an oncogene, XIAP overexpression is common in cancer and has been linked to poor patient prognosis (24). In another class of IAP proteins, c-IAP1 and c-IAP2 are unique among IAP proteins for their ability to interact with TRAF1 and 2 (tumor necrosis factor receptor-associated factors 1 and 2) (25, 26). Importantly, c-IAP1 and c-IAP2 are also targets of genetic amplification, which correlates with resistance to chemotherapy and radiotherapy (27). The (melanoma inhibitor of apoptosis protein) (ML-IAP) is upregulated in melanomas but not expressed in most normal adult tissues (28). Changes in IAP expression modulate apoptotic response to cellular stress. For example, an antisense oligonucleotide developed by Aegera Therapeutics, Inc. (Montreal, Quebec, Canada), AEG35156, reduces expression of XIAP. Treatment with AEG3516 increases apoptosis in studies with myeloid leukemia cells, and the antisense approach currently is in phase I/II clinical trials for refractory AML (acute myeloid leukemia) in combination with chemotherapy (29). The strategy of up-regulating IAP action to inhibit apoptosis has been validated...
Programmed Cell Death as a Therapeutic Approach

Figure 2: Small-molecule antagonists of prosurvival Bcl-2 proteins: (a) ABT-737 (19); (b) GeminX (21); (c) Yale (20); (d) Gossypol (16); and (e) the University of Michigan (22).

by transfection of a gene-encoding XIAP in a rat glaucoma model (30).

The Smac protein promotes apoptosis by neutralizing the IAP proteins through interactions of the Smac N-terminal region with the BIR domains of the IAP (31, 32). The interaction between Smac and IAP proteins disrupts the IAP inhibition of caspases, which allows caspase activation and eventual cell suicide. In vivo and in vitro studies reveal that cell killing via either the intrinsic or the extrinsic pathways can be enhanced by Smac mimetics (33–36). Collectively, these results suggest that Smac mimetics that bind to the BIR domains of the IAPs extricate the caspases in a manner analogous to Smac. Significantly, interactions between Smac and the BIR3 domain of XIAP are localized to the four N-terminal residues of mature processed Smac (31, 32), suggesting that a small-molecule mimetic might be sensible. The immediate challenge in developing a small-molecule Smac mimic is rescaffolding the key binding determinants of the tetrapeptide onto a platform that exhibits drug-like properties. This rescaffolding has become an active area of research with several groups producing nonpeptide Smac mimics. Examples of small-molecule Smac mimetics under preclinical investigation are shown in Fig. 3, with significant progress reported by groups from Abbott (37), University of Michigan (38), Genentech (39), Novartis (Basel, Switzerland) (40), Princeton University (Princeton, NJ) (41), and Texas Southwestern Medical Center (Dallas, TX) (25). Based on these reports, it seems that an effective Smac mimic will require a free amino group and a small hydrophobic amino acid at the
N-terminus, a proline derivative in the third position, and an aromatic group at the fourth position. The studies mentioned above demonstrate Smac mimetics are effectors of both the extrinsic and the intrinsic apoptosis pathways.

Targeting the P53 Pathway

The transcription factor p53 responds to cellular stress by promoting the production of growth suppressors and proapoptotic factors that lead to cell cycle arrest, senescence, or apoptosis (42). Activation of the p53 pathway is known to target more than 16 genes that regulate apoptosis (43). Significantly, when detecting DNA damage, p53 initiates the expression of proapoptotic BH3-only proteins Puma (p53-upregulated modulator of apoptosis) and Noxa (43) (see the “Targeting the Bcl-2 Family of Proteins” section above). Approximately 50% of cancers show mutations or deletions in the p53 gene, which silences a critical mechanism for limiting the propagation of damaged cells. Most alterations that lead to loss of p53 function consist of single amino acid mutations that destabilize the protein core (44). Function has been rescued in these mutants by binding of peptides derived from the C-terminal domain (44) or small molecules (45). Activity of C-terminal peptides is believed to be at least partially caused by antagonizing the binding of the p53C-terminus to the DNA-binding motif of p53. The small-molecule activators developed by Pfizer (New York, NY) shown in Fig. 4, CP-31398 and CP-257042, restore p53 function by stabilizing the protein core (45). The potency of these molecules requires a hydrophobic group within a fixed distance from an ionizable group, which suggests a specific interaction with the protein. Studies with CP-31398 show inhibition of tumor growth in melanoma and colon cancer xenograft models.

Regulation of p53 activity is complex, including numerous posttranslational modifications. One such modification is ubiquitination by the E3 ligase MDM2 (murine double minute 2); the interaction between MDM2 and p53 has been identified as a potential target for therapeutic intervention. MDM2 binds to p53 in the transcription activation domain, blocking p53 function...
Programmed Cell Death as a Therapeutic Approach

Small molecules that stabilize mutant p53, (a) CP-31398 and (b) CP-257042 (45). Small molecule antagonists of MDM2 binding, (c) RITA (51); (d) Nutlin-1; (e) Nutlin-2; and (f) Nutlin-3 (52).

while simultaneously targeting p53 for ubiquitination and degradation. Cancers with functional p53 often silence p53 by the overexpression of MDM2 (46). Studies that use either blocking antibodies and peptides or reduction of MDM2 levels employing antisense oligonucleotides demonstrate that disruption of the p53–MDM2 interaction activates the p53 pathway, which blocks proliferation and promotes apoptosis (47, 48). Examination of a crystal structure of a p53-derived peptide bound to MDM2 reveals that three amino acids (Phe19, Trp23, and Leu26) project deep into a hydrophobic-binding pocket on the MDM2 surface (46, 49). The localized nature of the interface suggests MDM2 as a potential small-molecule target. The Hamilton group used an approach analogous to the Bcl-2 work mentioned above to obtain a helix mimic that binds to MDM2 (50). By screening for molecules that only show activity in cells with wild-type p53, researchers at the Karonlinska Institutet (Stockholm, Sweden) discovered a small molecule they named RITA, which shows in vivo p53-dependent antitumor activity (51) (see Fig. 4). A major advance in this area is the work reported by the Hoffman-La Roche team (Nutley, NJ) in their development of a series of potent MDM2 antagonists they named the nutlins (Fig. 4). This class of MDM2-binding molecules was identified initially in a high-throughput screen and followed by extensive structural optimization that yielded potency in the 100–300-nM range (52). Studies using an osteosarcoma SJS/A-1 xenograft tumor model demonstrated that a 20-day course of treatment with racemic nutlin-3 resulted in a 90% inhibition of tumor growth as compared with vehicle treatment. Resolution of the racemate to give the active nutlin-3 enantiomer increased potency two fold, and studies with enantiomerically pure compound evidenced 100% growth inhibition in the SJS/A-1 model as well as in an LNCaP prostate xenograft model. Both models showed significant tumor regression, which suggests that the disruption of the MDM2–p53 interaction is antiproliferative and proapoptotic and that antagonizing MDM2 suppression of p53 is a promising therapeutic approach for cancers that retain wild-type p53 function (49).

Targeting the Death Receptors

One of the most promising advances in apoptosis-cancer therapies is the activation of the extrinsic pathway through the death receptors (53). This is an extremely active area of biology research, but with little potential for a small-molecule approach. An attractive feature of the extrinsic pathway is that its unique activation mechanism provides an avenue for promoting cell death in cancers that resist current therapies (8). Another feature of the extrinsic pathway is that many cancer cells highly express death receptors but do not express the decoy receptors that are found on normal cells, which increases the susceptibility of cancer cells to death receptor activation (54). As mentioned earlier, the potential for agents that target multiple components of apoptosis holds promise for effective treatment of resistant cancers. An exciting example of this potential is illustrated in a study in which a ligand that activates death receptors, Apo2L/TRAIL, was combined with a Smac-derived peptide (36); complete regression was observed in a mouse glioma model, one of the most difficult cancers to treat and one for which no current viable therapy exists. This result suggests that activation of the extrinsic pathway in combination with apoptosis-sensitizing agents is an effective strategy for treating cancers resistant to conventional therapies.

References


Further Reading

See Also
Cell Death: Apoptosis and Necrosis
Chemical Tools to Dissect the Apoptotic Pathways
The Apoptosome and Caspase Activation, Chemistry of DNA Damage
Gene Therapy and Cell Therapy
Mitochondria: Topics in Chemical Biology
RNA Interference to Treat Human Diseases, Applications of

Ge Shan and Peng Jin, Department of Human Genetics, Emory University
School of Medicine, Atlanta, Georgia
doi: 10.1002/9780470048672.wecb518

Discovery of the RNA interference (RNAi) pathway has been revolutionary to biology and biologically based therapeutics. The messenger RNA (mRNA) targets once believed to be “undruggable” by conventional methods are now being enthusiastically pursued with RNAi-based therapeutics. Fundamental research into the RNAi pathway and biotechnological research on the application of RNAi may one day make a new class of drugs a reality by targeting the mRNA of disease-causing or disease-promoting genes directly. The whole process begins with the design and validation of short-interference RNA (siRNA) using bioinformatics and cell culture systems. siRNA that has high specificity and potency (knockdown efficiency) then could be applied to animal models of human diseases for more evaluation, with the ultimate hope of advancing to clinical trials.

Numerous chemical modifications of siRNA are currently being investigated in an effort to enhance their stability, to extend their stay in the body, and, perhaps, to avoid immune stimulation. Different delivery methods are developed depending on the diseases, tissues, or organs being targeted. In short, this relatively new biotechnology is advancing quickly as we learn from its failures and successes. In the current article, we will explore the background and current state of research as well as discuss some hurdles and future directions in the application of RNAi to treat human diseases.

In 1998, Andrew Fire and Craig Mello (1) published their seminal work on RNA interference (RNAi) with their discovery of double-stranded RNA (dsRNA) as the trigger of posttranscriptional silencing in Caenorhabditis elegans. The phenomenon of posttranscriptional gene silencing was first observed in plants (2–5). In the two decades since, many components in the RNAi pathway have been identified and characterized (see also RNA Interference, Mechanisms and Proteins Involved in). Another phenomenon of posttranscriptional silencing comes from microRNA (miRNA). In 1993, Lee et al. (6) cloned a short noncoding RNA (later referred to collectively as microRNA), lin-4. In 2000, Reinhart et al. (7) cloned another miRNA called let-7 that functions similarly to lin-4. Both lin-4 and let-7 regulate translation via partially complementary sequences in the 3′ untranslated region (UTR) of mRNA they target. Advances in research eventually merged the RNAi pathway with the miRNA pathway by showing that they closely share core components (see also RNA Interference, Mechanisms and Proteins Involved in).

In principle, dsRNA and short-hairpin RNA (shRNA) are cut by Dicer and its associated proteins, such as TRBP, into ∼21-nucleotide (nt) short interference RNA (siRNA) or mature miRNA. One strand (the guide strand) will be integrated into the RNA-induced silencing complex (RISC), which has AGO2 as its core component, whereas the other strand (the passenger strand) will be either degraded (siRNA) or released (miRNA). In the current model, siRNA, by perfect complementarity with the targeted mRNA, triggers the degradation of mRNA in the RISC, whereas miRNA generally suppresses the translation of the target mRNA by incomplete complementarity with the 3′ UTR region (see also RNA Interference, Mechanisms and Proteins Involved in). The discovery that dsRNA, particularly siRNA, can be introduced exogenously into mammalian cells to knock down target mRNAs in a sequence-specific manner generated much enthusiasm for exploring siRNA duplexes as gene-specific therapeutics (8). The subsequent development of siRNAs as drugs has been rapid. Currently, three Phase I studies...
RNA Interference to Treat Human Diseases, Applications of

Table 1  Examples of in vivo application of RNAi-based therapeutics

<table>
<thead>
<tr>
<th>Gene targeted</th>
<th>Disease model</th>
<th>Delivery strategy and route</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEGF</td>
<td>Laser-induced CNV</td>
<td>Lipoplex/Subretinal</td>
<td>9</td>
</tr>
<tr>
<td>VEGF</td>
<td>Laser-induced CNV</td>
<td>Saline/Intravitreal</td>
<td>10</td>
</tr>
<tr>
<td>VEGF R1</td>
<td>Subcutaneous tumor xenograft</td>
<td>A telocollagen/intravenous</td>
<td>11</td>
</tr>
<tr>
<td>VEGF R2</td>
<td>Subcutaneous tumor xenograft</td>
<td>Cholesteryl-R9/intratumoral</td>
<td>12</td>
</tr>
<tr>
<td>c-myc, MDM2, VEGF</td>
<td>Subcutaneous tumor xenograft</td>
<td>RDG peptide/PEI/Intravenous</td>
<td>14</td>
</tr>
<tr>
<td>TGFr R2</td>
<td>Subconjunctival scarring</td>
<td>Lipoplex/Subconjunctival</td>
<td>16</td>
</tr>
<tr>
<td>EWS-FL1</td>
<td>Ewing’s sarcoma model</td>
<td>Transferin-targeted</td>
<td>17</td>
</tr>
<tr>
<td>PLK1, BCL2</td>
<td>Subcutaneous tumor xenograft</td>
<td>PSMA aptamer/intratumoral</td>
<td>18</td>
</tr>
<tr>
<td>HPV-E6</td>
<td>Subcutaneous tumor xenograft</td>
<td>Lipoplex/Intratumoral or peritoneal</td>
<td>19</td>
</tr>
<tr>
<td>Plectropinophil</td>
<td>Subcutaneous and glioblastoma</td>
<td>PEI/intrapertioneal or subcutaneous</td>
<td>20</td>
</tr>
<tr>
<td>HER2</td>
<td>Subcutaneous tumor xenograft</td>
<td>PEI/intrapertioneal</td>
<td>21</td>
</tr>
<tr>
<td>RhoA</td>
<td>Subcutaneous tumor xenograft</td>
<td>Chitosan nanoparticle/intravenous</td>
<td>22</td>
</tr>
<tr>
<td>SARS</td>
<td>SARs infection</td>
<td>DSW or surfactant/intranasal</td>
<td>24</td>
</tr>
<tr>
<td>Influenza A-NP, PA</td>
<td>Influenza virus infection</td>
<td>PEI/intravenous</td>
<td>25</td>
</tr>
<tr>
<td>JEV/WNV-envelope</td>
<td>JEV/WNV-induced encephalitis</td>
<td>Lipoplex/intracranial</td>
<td>26</td>
</tr>
<tr>
<td>HSV-2 UL27, UL-29</td>
<td>Herpes simplex 2 virus infection</td>
<td>Lipoplex/intravaginal</td>
<td>27</td>
</tr>
<tr>
<td>ApoE</td>
<td>Dyslipidemia</td>
<td>Cholesterol conjugated/intravenous</td>
<td>28</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>Dyslipidemia</td>
<td>Liposomen/intravenous</td>
<td>29</td>
</tr>
<tr>
<td>TNF</td>
<td>Collegen-induced arthritis</td>
<td>Lipoplex/intravenous</td>
<td>30</td>
</tr>
<tr>
<td>HMOX1</td>
<td>Dextran sodium sulfate-induced colitis</td>
<td>Lipoplex/intracranial</td>
<td>31</td>
</tr>
<tr>
<td>KC, MIP2, Fas</td>
<td>Septic acute lung injury</td>
<td>Saline/intranasal</td>
<td>32</td>
</tr>
<tr>
<td>ANGPT2</td>
<td>Hyperoxic lung injury</td>
<td>Saline/intranasal</td>
<td>33-34</td>
</tr>
<tr>
<td>DDR1</td>
<td>Bleomycin-induced fibrosis</td>
<td>Saline/intranasal</td>
<td>35</td>
</tr>
<tr>
<td>Cavertinin-1</td>
<td>Vascular permebility</td>
<td>Liposomen/intravenous</td>
<td>36</td>
</tr>
<tr>
<td>AGRP</td>
<td>Metabolic alterations</td>
<td>Saline/Steroidastic injection</td>
<td>37</td>
</tr>
<tr>
<td>Dopamine Transporter</td>
<td>Hyperoculumotor response</td>
<td>Saline/intraventricular</td>
<td>38</td>
</tr>
<tr>
<td>Serotonin Transporter</td>
<td>Behavior response</td>
<td>Saline/intraventricular</td>
<td>39</td>
</tr>
<tr>
<td>P2X3</td>
<td>Chronic neuropathic pain</td>
<td>Lipoplex/intrathecal</td>
<td>40</td>
</tr>
<tr>
<td>DDR1</td>
<td>Delt-induced nociception</td>
<td>Lipoplex/intrathecal</td>
<td>41</td>
</tr>
<tr>
<td>NMADR (subunit 2B)</td>
<td>Formalin-induced nociception</td>
<td>PEI/intrathecal</td>
<td>42</td>
</tr>
<tr>
<td>PrP(C)</td>
<td>Scapie infection</td>
<td>Lentvector/intracranial</td>
<td>43</td>
</tr>
<tr>
<td>mir-21</td>
<td>Murine glioma model</td>
<td>LNA-anti-mir-21</td>
<td>44</td>
</tr>
</tbody>
</table>

without unacceptable toxicity are already complete (Table 1) (9-45).

RNAi-based therapeutics begin with the design and identification of siRNAs that show high specificity and knockdown potency in the lab. Once several siRNAs have been selected, they are applied to animal models with the diseases of interest. Before delivery, siRNA may also be modified to enhance its stability. siRNA duplexes can be delivered via different routes and even with different strategies. We will now summarize and discuss the entire process, from the initial design of a siRNA duplex to its application in clinical trials.

Design and Evaluation of siRNA

Currently, two methods are available for the endogenous RNAi pathway for therapeutic purposes: either by introducing a viral vector to express shRNA that mimics a miRNA precursor, which then would be processed by Dicer into siRNA, or by delivering siRNA that mimics a Dicer cleavage product into the cytoplasm. Because siRNA has to be processed into siRNA in vivo to be functional, the design of a shRNA vector follows the same rules as for siRNA. For either method, the first consideration is to ensure that the siRNA specifically targets the mRNA of interest (specificity), at the same time, the siRNA
should have a minimum desirable (if not the highest possible) knockdown efficiency (potency).

**Specificity**

The initial siRNA design starts with a bioinformatics-aided search for "targetable" sequences —21 nt long in the mRNA of interest (46). Because a perfect complement with the targeted mRNA triggers degradation and an imperfect complement triggers translational suppression, the chances of generating a 21-nt siRNA with an off-target effect is high, without careful attention to design (47, 48). The principle here is to avoid sequences in the siRNA "seed region" complementary to untranslated mRNAs (Fig. 1). The "seed region" is at positions 2-7 or 2-8 of the guide strand of the miRNA or siRNA duplex. For miRNA function, the seed region usually complements the 3′ UTR of the target mRNA perfectly. Off-targeting of siRNA is found to be associated with perfect complementarity between the seed region and the 3′ UTR of unwanted mRNA (49). It is highly possible that targeting to the coding region of nontranslated mRNA would also contribute to off-target effects. Several software and Internet search programs can be helpful in the selection of siRNA sequences to minimize off-target effects (50, 51). These effects should also be checked with a microarray assay in cell culture systems because currently no ideal strategy to design (47, 48). The principle here is to avoid sequences in the siRNA "seed region" complementary to untranslated mRNAs (Fig. 1). The "seed region" is at positions 2-7 or 2-8 of the guide strand of the miRNA or siRNA duplex. For miRNA function, the seed region usually complements the 3′ UTR of the target mRNA perfectly. Off-targeting of siRNA is found to be associated with perfect complementarity between the seed region and the 3′ UTR of unwanted mRNA (49). It is highly possible that targeting to the coding region of nontranslated mRNA would also contribute to off-target effects. Several software and Internet search programs can be helpful in the selection of siRNA sequences to minimize off-target effects (50, 51). These effects should also be checked with a microarray assay in cell culture systems because currently no ideal strategy completely eliminates off-target effects.

**Potency**

Although our current understanding of RNAi activity cannot provide us with a precise prediction of individual siRNA potency, algorithms based on common features of empirically tested high-potency siRNAs are available (46, 52). Once the bioinformatic part is complete, a candidate siRNA can be synthesized and tested in cell culture systems for knockdown efficiency. Knockdown efficiency is commonly assessed with real-time polymerase chain reaction (PCR) to check for changes in the mRNA level and with western blot to check for decreases in the protein level. The final goal of this stage is to identify several siRNAs that show high knockdown efficiency and minimal off-target effects at nanomolar or lower concentrations. To be prepared for possible underachievers at any point in the long process from the bench to the bedside, it is better to have more than one siRNA to begin with.

Theoretically, an alternative method is available for selecting siRNA with high specificity and potency: A library of tiling siRNAs covering the whole mRNA sequence of a disease-causing or relevant gene of interest can be screened with cell culture systems, and individual siRNAs with stringent specificity and high efficacy then can be identified from this library. Drawbacks of this method are that it is more expensive and more labor intensive, although the knowledge gained from this approach can assist in the future design of therapeutic siRNAs.

When considering specificity and potency, it is helpful to bear in mind sequence conservation across species as well. Therapeutic siRNAs usually have to undergo tests that involve nonhuman cell cultures and animal disease models, so it is more reasonable to start with siRNAs that target conserved sequences in disease-related mRNA. One should remember that homologs in different organisms may function differently, and each siRNA could show different knockdown efficiency in different organisms.

Several practices could enhance the specificity and potency of siRNA once its sequence is decided. For example, because the loading of siRNA into RISC is asymmetric (53, 54), the first nucleotide of the 5′ end of the guide strand is efficient at suppressing off-targeting (57).

**Chemical Modification of siRNA**

Some chemical modifications can enhance the specificity of siRNA even more. Chemical modifications of riboses in the guide strand have been found to suppress off-target effects without affecting potency (55, 56). In fact, 2′-O-methyl modification at nucleotide 2 of the guide strand is efficient at suppressing off-targeting (57).

**Chemical modification to increase siRNA stability**

The main purpose of modifying siRNA chemically is to increase its stability so that it remains active in the animal and human body. Unmodified siRNA gets degraded quickly in human plasma, with a half-life of only minutes (58, 59). To convert siRNA into drugs, several ways of modifying siRNA to prolong its half-life have been investigated (Fig. 1). Knowledge of...
how to avoid nucleic degradation has come to us from previous therapeutics research that use antisense oligonucleotides and aptamers, and this knowledge has been beneficial in the chemical modification of siRNA. Nuclease degradation is prevented mainly via two types of modification.

The first type is phosphodiester modification: Replacing one of the two nonbridging oxygen atoms with a sulfur atom (P-S) or an isophosphorane borane (BH3) moiety is found to protect siRNA from exonuclease degradation (58-62; Fig. 1). Whereas moderate P-S modifications are well tolerated in terms of knockdown potency and toxicity, extensive P-S modifications should be avoided because of increased binding of the modified siRNA duplex to serum proteins, which may result in cytotoxic knockdown. Although BH3 modification has not been investigated extensively, current research shows it has some advantages over P-S modification (65); however, one situation to avoid is having a BH3 modification at the center of the guide strand, which could reduce the efficacy of the siRNA.

The second type is 2'-sugar modifications: Modifications at the 2'-position of the ribose ring protect siRNA from endonuclease degradation. These modifications mainly include 2'-O-methyl (2'-OMe), 2'-deoxy-2'-fluoro (2'-F) modifications, and locked nucleic acid (LNA) (57, 60, 66-68; Fig. 1). A proprietary 2'-OMe modifications not only increase plasma stability but also ameliorate off-target effects and enhance the in vivo potency of siRNA (55, 57). One important point to remember is that 2'-OMe modifications could impair the cleavage of passenger strand and target mRNA, so it is crucial to keep this modification away from the cleavage site in the passenger strand (69, 70). 2'-F modifications dramatically increase the stability of siRNA in human plasma without adverse effects on specificity and potency; these modifications do not enhance the performance of siRNA, unlike 2'-OMe modifications (58). The most commonly used LNA in the siRNA duplex contains a methylene bridge to connect the 2'-oxygen with the 4'-carbon of the ribose ring (Fig. 1). Besides increasing stability, LNA modifications may also help to reduce off-target effects and enhance potency. Moderate LNA is generally well tolerated in siRNA, with several notable exceptions. LNA modification should be avoided at the 5'-end of the passenger strand because this may affect the asymmetric loading of the guide strand (71). Both extensive LNA and LNA at certain locations (depending on the specific siRNA) may block the RNAi activity of the modified siRNA (72).

### Chemical modification to avoid innate immune response

Nucleic acids can trigger innate immune responses (73). dsRNA longer than 30 base pairs (bp) can efficiently trigger serine/threonine protein kinase (PKR). Although siRNA is smaller, at higher concentrations it may trigger this pathway, which leads to global blockade of translation and ultimately cell death (74, 75). Another issue of perhaps greater concern in RNAi therapeutics is that siRNA could activate Toll-like receptors (TLRs), especially the dsRNA receptor TLR7 in plasmacytoid dendritic cells, which triggers the production of type I interferons and proinflammatory cytokines and thus induces nuclear factor-B activation (76). Some siRNAs have a greater tendency to activate TLRs, and, for this reason, they should properly be called immunostimulatory RNA (isRNA) (72, 77). 3’ blunt ends and GU-rich sequences are strong stimulants to PKR; therefore, they should be avoided when designing the siRNA (78). It is important to investigate whether a siRNA duplex is an siRNA using in vitro plasmacytoid dendritic cell culture before in vivo application to animal models (72). Chemical modifications at the 2’-sugar can be beneficial because they help avoid immunostimulation (Table 2).

Different chemical modifications may be used in combination to attain the accumulated benefits of each individual modification, although research in this area has not been exhaustive.

#### In Vivo Delivery

siRNA duplexes have to be delivered effectively to treat diseases. Efficient and cell type-specific delivery may be the biggest obstacle to the development of RNAi therapeutics. For use in animal models and clinical trials, different delivery strategies have been developed to meet the requirements for different diseases and target tissues (or organs).

#### Direct application of naked siRNA

This delivery method applies siRNA dissolved in saline or 5% dextrose (D5 W) directly to the targeted tissues, with many successes reported. Most siRNA-saline solutions target specific organs, such as the eye (intravitreal injection), lung (intranasal or intratracheal instillation), or central nervous system (intraventricular, thecal, or parenchymal infusion) (Table 1). Because these organs allow direct administration of siRNA, they were a natural focus of initial RNAi therapeutics. Certain cell types can take up naked siRNA efficiently via unknown mechanisms, whereas many other cell types are refractory to naked siRNA. For this reason, siRNA duplexes have to be conjugated or formulated for efficient delivery in most cases.

#### Delivery via liposomes and lipoplexes

Liposomes are vesicles enclosed by a phospholipid bilayer, which can fuse with a cell membrane and deliver the enclosed contents into the cytoplasm (Fig. 2). Liposomes have already delivered many drugs with decreased toxicity and increased pharmacokinetics. In particular, stable nucleic acid lipid particles (SNALPs) that consist of a mixture of cationic and fusogenic lipid bilayers were used for intravenous and intraportal delivery (29, 79). These SNALPs can be coated with a diffusible poly(ethylene glycol)-lipid (PEGylated lipid) conjugate, which stabilizes the liposomes during formation and prevents rapid systemic clearance from the circulating blood. Liposomes have successfully delivered siRNA that target liver diseases like hepatitis B virus (HBV) infection (79, 80; Table 1). siRNAs are commonly transfected into cultured cells with reagents such as Lipofectamine 2000 (Invitrogen Corporation, Carlsbad, CA) or TransIT-TKO (Mirus Bio Corporation, Madison, WI). Complexes formed by siRNA and most commercial
Table 2 Chemical modifications of therapeutic siRNA

<table>
<thead>
<tr>
<th>Chemical modifications</th>
<th>Advantages</th>
<th>Negative aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>2’-O-methyl</td>
<td>Suppress off-targeting (especially at the position 2 of strand, the guide strand) Enhance in vivo potency Increase plasma stability</td>
<td>A void from cleavage site in the passenger; may interfere with cleavage, if not.</td>
</tr>
<tr>
<td>2’-deoxy-2’-fluoro</td>
<td>Increase plasma stability</td>
<td></td>
</tr>
<tr>
<td>Locked nucleic acid</td>
<td>Increase stability; Reduce off-targeting; Enhance potency.</td>
<td>Avoid from the 5 end of the passenger strand; A void extensive modification.</td>
</tr>
<tr>
<td>All 2’sugar modifications also help to avoid the immunostimulation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phosphodiester modifications

- P=S modification Protect from exonuclease degradation. Extensive modification results in cytotoxic cell death. |
- BH3 modification Protect from exonuclease degradation. A void from the center of the guide strand, may reduce efficacy, if not. |

Figure 2 Examples of therapeutic siRNA delivery strategies. a. RNA aptamer conjugation. b. Cholesterol conjugation. c. PEG-liposome. d. Antibody-protamine conjugation. Different chemical modifications and delivery strategies can also be used in combination.
transfection reagents are called lipoplexes. Lipoplexes are relatively easy to form, and they have been widely used for siRNA delivery. Successful applications have been reported via almost all the delivery routes, including subconjunctival, intrathecal, intracranial, intratumoral, intraperitoneal, intravenous, intravaginal, and intraretinal administration (Table 1).

### Chemical conjugation to small molecules

The idea of conjugating cationically small molecules, such as cholesterol and siRNA, was borrowed from previous research based on antisense oligonucleotide therapeutics (81). Both the 5′ and 3′ ends of the passenger strand tolerate conjugations well. 5′-end cholesterol conjugation of the passenger strand has been shown to efficiently deliver siRNA intravenously for targeting the ApoB gene in the liver and jejunum (82).

RNA aptamers can also be conjugated to siRNA for the purpose of cell-specific delivery. Aptamers are artificial DNA or RNA molecules that bind to specific molecular targets (83). Theoretically, RNA aptamer binding to cell-specific receptors or antigens can be linked to siRNA for the purpose of cell type-specific RNAi. One report using a siRNA conjugated with an RNA aptamer that has a high affinity for prostate-specific membrane receptors showed that it silenced survival genes in prostate cancer cells with a high cell specificity and efficacy (18).

#### Conjugation with peptides and antibodies

When it comes to cell-specific delivery, peptides or antibodies can also be conjugated noncovalently with siRNA. The extreme specificity of an antibody to recognize and bind to a cell-specific antigen makes antibody conjugation a very attractive approach for delivering siRNA. A fusion protein with a specific antibody and a protamine fragment, which is arginine-rich and thus positively charged, can bind to siRNA (negatively charged) for delivery. Some successful in vitro and in vivo applications include the delivery of siRNA to B16 melanoma expressing HIV envelope protein or HIV-infected primary CD4+ T cells using a fusion protein of protamine and Fab fragment of antibody to HIV envelope protein (Tables 1 and 3).

The simplest peptide conjugation uses cholesterol oligo-D-arginine (Chol-R9, chosen because of the positive charge of arginines). Noncovariant formation of a complex of siRNA with Chol-R9 efficiently delivered siRNA targeting VEGF into cells. Moreover, in a mouse model bearing a subcutaneous tumor, this complex led to regression of the tumor (12). Other peptides that have been investigated include MPG, derived from the fusion peptide domain of HIV-1 gp41 protein and the nuclear localization sequence (NLS) of SV40 large T antigen. To be used for siRNA delivery, a mutation that affects the NLS of MPG was generated to prevent the nuclear entry of siRNA. In cell culture, the peptide enables rapid delivery of the siRNA into the cytoplasm, which results in robust downregulation of target mRNA (84). Similar to MPG, Penetratin (Qbiogene Inc., Irvine, CA), which is a peptide derived from the homeodomain of the Drosophila protein Antennapedia, is known to deliver cargo into cells. Indeed, Penetratin 1 can be conjugated to siRNA for rapid and efficient delivery into cultured primary mammalian hippocampal and sympathetic neurons (85). For all these peptides, only one report of successful in vivo delivery has been made to date, in an animal model with chol-R9.

Some peptides can be recognized by cell-specific receptors and thus can be conjugated with siRNA for cell-specific delivery. Recently, a 29-amino-acid peptide derived from rabies virus glycoprotein (RVG) was shown to be capable of delivering conjugated siRNA (via a nonamer arginine fusion, RVG-R9) transvascularly to the brain, which resulted in efficient gene silencing. This peptide was found to bind specifically to the acetylcholine receptor expressed by neuronal cells (82).

#### Delivery with nanopolymers

Nanotechnology involves the manipulation of atoms and molecules to construct structures on the nanometer scale (often 100 nm or smaller). Liposomal vesicles, lipoplexes, as well as antibody- and some peptide-conjugated siRNA complexes discussed above are actually nanoparticles. Nanoscale technologies are changing the foundations of therapeutics; in fact, this branch of medicine is now referred to as "nanomedicine" by the National Institutes of Health. Nanoparticles have special advantages in drug delivery. First, therapeutic agents can be encapsulated and hence protected from degradation, clearance, and nonspecific binding. Second, the release speed and location (tissue or even subcellular localization) of drugs can be well controlled by manipulating the composition of nanoparticles. And finally, the pharmacokinetics of the drugs can be optimized (98).

Currently, the most widely used nanoparticles for siRNA delivery are nanopolymers formed with polyethyleneimine (PEI). PEI polymers are synthetic, are highly cationic charged, and can be used directly to form complexes with siRNA duplexes. The resulting PEI-siRNA polyplexes are thought to enter cells via endosomes, in which PEI disrupts the low endosomal pH, leading to the eventual release of PEI-siRNA complexes into the cytoplasm. Several reports using PEI polymers demonstrated efficient siRNA delivery in animal models of influenza, Ebola virus infection, and tumors (Table 1). Components such as peptides and small compounds (e.g., folate) for cell-specific delivery can also be added into the PEI-siRNA polyplexes. For example, arginine-glycine-aspartic acid-(RGD)-peptide, which is known to bind to both tumors and tumor-endothelial cells in vivo (87), has been successfully used to deliver PEI-siRNA into tumor tissues (14). One drawback of PEI has been its extreme toxicity at high concentrations. Its methylene backbone (-CH(2)CH(2)N(3)-) and high charge density make for poor biodegradability and high toxicity to cells. Much effort has been expended to optimize the PEI structure to expand its safety margin (88-91).

Another material of current interest is chitosan. Chitosan is produced commercially by the deacetylation of chitin, which is the structural element in the exoskeleton of crustaceans. Chitosan is positively charged and is bioadhesive to negatively charged molecules, such as nuclear acids, and surfaces, such as mucosal membranes. Chitosan enhances the transport of polar drugs across epithelial surfaces and is biocompatible and biodegradable. Previously, it was used for DNA delivery in other contexts. However, for siRNA delivery, chitosan appears to have particular advantageous properties. For example, siRNAs are delivered to mammalian hippocampal and sympathetic neurons (85). For all these peptides, only one report of successful in vivo delivery has been made to date, in an animal model with chol-R9.

Some peptides can be recognized by cell-specific receptors and thus can be conjugated with siRNA for cell-specific delivery. Recently, a 29-amino-acid peptide derived from rabies virus glycoprotein (RVG) was shown to be capable of delivering conjugated siRNA (via a nonamer arginine fusion, RVG-R9) transvascularly to the brain, which resulted in efficient gene silencing. This peptide was found to bind specifically to the acetylcholine receptor expressed by neuronal cells (82).
<table>
<thead>
<tr>
<th>Delivery strategy</th>
<th>Method</th>
<th>Advantages</th>
<th>Negative aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naked siRNA</td>
<td>siRNA duplex dissolved in saline or 5% dextrose</td>
<td>Cheap and straightforward</td>
<td>Applicable only to some tissues with easy direct accessibility</td>
</tr>
<tr>
<td>Lipoplexes</td>
<td>siRNA duplex in complexes with transfection reagents (cationic lipids and neutral lipids).</td>
<td>Convenient to make in the lab</td>
<td>Not tissue specific delivery</td>
</tr>
<tr>
<td>Liposomes</td>
<td>siRNA duplex enclosed in a bilayer vesicle of cationic and fusogenic lipids. To stabilize them, or for other purposes, molecules such as cholesterol and PEG are often incorporated in liposomes.</td>
<td>Decrease toxicity and increase pharmacokinetics</td>
<td>Not easy to prepare. Not tissue specific delivery at this stage.</td>
</tr>
<tr>
<td>RNA aptamer conjugation</td>
<td>Chemically conjugate RNA aptamer to the passenger strand of siRNA duplex during RNA synthesis.</td>
<td>Tissue specific delivery; can design all kinds of aptamer for different specificity.</td>
<td>Relatively expensive; large RNA molecule may be immuno-stimulative.</td>
</tr>
<tr>
<td>Cholesterol conjugation</td>
<td>Chemically conjugate cholesterol to the passenger strand of siRNA duplex.</td>
<td>Specific delivery to cells (e.g., liver cells) uptake cholesterol.</td>
<td>Sometimes it may be hard to find a small molecule for a specific cell type.</td>
</tr>
<tr>
<td>Peptide conjugation</td>
<td>Positive charged peptide (e.g., cholesteryl-R9 and RVG-R9) conjugate non-covalently to siRNA.</td>
<td>Tissue specific delivery</td>
<td>Need more in vivo experiments</td>
</tr>
<tr>
<td>Antibody conjugation</td>
<td>Antibody-protamine fusion protein conjugates non-covalently to siRNA duplex.</td>
<td>Cell specific delivery</td>
<td>Need more in vivo experiments</td>
</tr>
<tr>
<td>Nanopolymers</td>
<td>Encapsulate siRNA duplex into nanoparticles formed with PEI or other materials.</td>
<td>All the benefits of using nanoparticles for drug delivery. Maybe next generation of siRNA delivery strategy.</td>
<td>PEI polymers still need optimized to reduce toxicity. Need more in vivo experiments</td>
</tr>
</tbody>
</table>

Disease Models and Clinical Trials

Current RNAi therapeutics are focused on infectious diseases caused by viruses, neurological diseases, and cancers or other overgrowth-related diseases (summary in Table 1).

Infectious diseases caused by viruses

Current viral diseases under consideration for RNAi therapeutics include those that infect the respiratory system, such as respiratory syncytial virus (RSV), parainfluenza virus (PIV), severe acute respiratory syndrome-associated coronavirus (SARS-
RNA interference to Treat Human Diseases, Applications of

Cov5), and influenza viruses; those that infect the central nervous system, such as Japanese encephalitis virus (JEV); viruses that infect the liver, such as HBV and Ebola virus; and those that infect other organs or are tumorigenic, such as herpes simplexes viruses types 2 (HSV-2) and human papillomaviruses (HPV). From these efforts, one Phase I clinical trial has been completed without untoward toxicity for siRNA-based therapeutics for RSV through intranasal administration of naked or lipoplex complexes. Notably, RSV infections are the leading cause of pediatric hospitalization in the United States today.

Tumors and other overgrowth-related diseases

The most investigated target for RNAi therapeutics to treat tumors and overgrowth-related diseases is vascular endothelial growth factor (VEGF) or its receptor VEGFR. siRNAs that target to VEGFR for age-related macular degeneration (AMD) have passed Phase I clinical trials. VEGF is overproduced in AMD, which resulted in the overgrowth of choroidal blood vessels into the subretinal space. AMD is a leading cause of blindness. Choroidal neovascularization (CNV) is the advanced stage of AMD and accounts for ~80% of the vision loss in AMD. Mouse or rat models of laser-induced CNV have been used to test the efficacy, potency, and delivery of siVEGFR1 (Table 1).

Most RNAi therapeutics use subcutaneous tumor xenograft mouse models. siRNA targeting overgrowth-related genes, such as VEGF, VEGFR, HER2, c-Myc, gliotoxin, and RhoA, have been tested with different delivery methods and demonstrated varying degrees of success (Table 1).

Dominant diseases

Hopes of applying RNAi-based therapeutics to treat dominant inheritable diseases continue. siRNA can distinguish a single-nucleotide difference between wild-type and mutant alleles when well designed (53, 94). A nother strategy for applying RNAi therapy to dominant diseases was initiated by Kiang et al. (95) and validated by O'Reilly et al. (96). Their strategy comprises two elements: gene suppression in conjunction with gene replacement. This strategy was tested in a model of retinitis pigmentosa, which is caused by single-site, dominant mutations in the rhodopsin gene. Using recombinant adeno-associated virus (AAV), researchers delivered a siRNA targeting a site independent of the mutation for both mutant and wild-type alleles while they applied in conjunction a codon-modified replacement gene refractory to that siRNA. The strategy proved successful both in vitro and in vivo.

Other diseases

The first report of efficient in vivo RNAi therapeutics involved a mouse model of fulminant hepatitis (acute liver failure) (97). Here, intravenous injection of Fas (also known as Tnfrsf6) siRNA protected hepatocytes from apoptosis. Some chronic diseases, such as dyslipidemias, arthritis, and colitis, are also under investigation for RNAi therapeutics. siNF (for arthritis and colitis) and siApoB (for dyslipidemias) have already been developed in animal models. Diseases of the nervous system are another category of special interest. siRNAs that target neurotransmitters, such as dopamine transporter (hyperlocomotor response), serotonin transporter (behavior response), P2X3 (chronic neuropathic pain), DOR (DELT-induced nociception), AGRP (metabolic alterations), and NR2B (formalin-induced nociception), are among those tested with animal models (Table 1).

RNAi-based therapeutics also holds out new hope for the treatment of neurodegenerative diseases, such as spinocerebellar ataxia, amyotrophic lateral sclerosis, Huntington’s disease, Alzheimer’s disease, and prion diseases (98–100; Table 1). Some neurodegenerative and other inheritable diseases are mediated by dominant alleles with trinucleotide repeats; allele-specific RNAi was developed for this kind of disease, including Huntington’s disease and myotonic dystrophy (101–103). Among preclinical research that is not related to trinucleotide repeats, Pfeifer et al. (44) used lentivectors to deliver PpIC-specific siRNA to both cell cultures and a mouse model of scrapie infection. They achieved efficient and stable suppression of PpIC accumulation in cultured neurons and a significant extension of survival in the mouse model. Hong et al. (104) generated replication-defective herpes simplex virus vectors for Aβ-specific siRNA and neprilysin-specific shRNA. These viral vectors inhibited Aβeta accumulation, both in vitro and in vivo.

Current Hurdles and Future Promises

With three Phase I clinical trials having passed to date, the area of RNAi therapeutics has moved rapidly from the first demonstration of efficient gene knockdown using siRNA duplex in mammalian cell culture in 2001 to many reports of the successful application of siRNA in animal disease models by 2003 (Table 1). More recently, John et al. (105) demonstrated that long-term systemic administration of siRNAs effectively silenced hepatocyte gene expression in rodents and primates, without significant changes in the levels of three hepatocyte-expressed miRNAs (miR-122, miR-16, and let-7a) or any effect on miRNA activity.

Many negative concerns that once surrounded RNAi therapeutics have been eliminated or at least addressed. Side effects that could be caused by off-target effects and immunostimulation from siRNA can now be reduced to acceptable levels through improvements in design, modification, and delivery strategies, and all these have greatly improved in vivo knockdown efficiency.

Despite all the progress, every aspect of RNAi-based therapeutics bears improvement. Algorithms for siRNA design are still far from ideal. In many cases, the required dosage is still high, in the range of 100 micrograms per administration in mice. Chemical modifications and delivery strategies remain to be optimized, with the goal of tissue-specific delivery and high knockdown efficiency.

A nother strategy for RNAi therapeutics, in parallel with siRNA, is shRNA delivered with recombinant viral vectors,
development advance, we are optimistic about the miraculous emergence of the RNAi pathway itself and as techniques and strategies that allow us to function and therefore may cause hazardous mutations, which have already hindered the use of retrovirus in gene therapy. Grimm et al. (107) investigated adenovirus-associated virus type B (AAV8) as a viral vector for shRNA expression in liver. lethality was found to be widespread at higher viral titers because of overexpression of the endogenous miRNA pathway. Thus, the application of adenovirus for RNAi therapy requires caution in optimizing the viral dosage and sequence of the encoded siRNA.

Dramatic changes in the expression of specific miRNAs are found in a variety of diseases, especially cancers (108). Some of these changes contribute to the etiology of certain diseases. It has been proposed that miRNA-based therapeutics could be developed by delivery of either miRNA duplexes to compensate for decreased levels or antisense oligos to block the effect of elevated levels of specific miRNA (109). For example, miR-21 levels were found to be elevated in gliomas, and their knockdown is associated with increased apoptotic activity (45, 110). Recently, Corsten et al. (45) showed that application of (LNA)-anti-miR-21 oligonucleotides has a synergistic effect in vitro and decreasing cell viability in human glioma cells (111). Given the potential to control the level of any miRNA through the RNAi pathway, RNAi therapies hold great promise for the treatment of virtually any disease with an etiology or pathology associated with an elevated or bad mRNA. As a treatment of virtually any disease with an etiology or pathology associated with an elevated or bad mRNA, RNAi therapy holds great promise for patients.

Given the potential to control the level of any miRNA through the RNAi pathway, RNAi therapies hold great promise for the treatment of virtually any disease with an etiology or pathology associated with an elevated or bad mRNA. As a phenomenon that biologists just began to uncover at the end of the last century, RNAi has generated enormous interest among scientists in both basic and medical research. As the secrets of the RNAi pathway itself emerge, and as techniques and strategy development advance, we are optimistic about the miraculous emergence of the RNAi pathway itself and as techniques and strategies that allow us to function and therefore may cause hazardous mutations, which have already hindered the use of retrovirus in gene therapy. Grimm et al. (107) investigated adenovirus-associated virus type B (AAV8) as a viral vector for shRNA expression in liver. lethality was found to be widespread at higher viral titers because of overexpression of the endogenous miRNA pathway. Thus, the application of adenovirus for RNAi therapy requires caution in optimizing the viral dosage and sequence of the encoded siRNA.

Dramatic changes in the expression of specific miRNAs are found in a variety of diseases, especially cancers (108). Some of these changes contribute to the etiology of certain diseases. It has been proposed that miRNA-based therapeutics could be developed by delivery of either miRNA duplexes to compensate for decreased levels or antisense oligos to block the effect of elevated levels of specific miRNA (109). For example, miR-21 levels were found to be elevated in gliomas, and their knockdown is associated with increased apoptotic activity (45, 110). Recently, Corsten et al. (45) showed that application of (LNA)-anti-miR-21 oligonucleotides has a synergistic effect in vitro and decreasing cell viability in human glioma cells (111). Given the potential to control the level of any miRNA through the RNAi pathway, RNAi therapies hold great promise for the treatment of virtually any disease with an etiology or pathology associated with an elevated or bad mRNA. As a treatment of virtually any disease with an etiology or pathology associated with an elevated or bad mRNA, RNAi therapy holds great promise for patients.

References


77. Judge AD, et al. Sequence-dependent stimulation of the mam-


79. Morrissey DV, et al. Potent and persistent in vivo anti-HBV ac-

tivity of chemically modified siRNAs. Nat. Biotechnol. 2005;23:

1002–1007.

80. Shin, D. et al. Immunostimulatory properties and antiviral activ-


Commun. 2007;346:436–442.

81. Eckstein F. The versatility of oligonucleotides as potential ther-


85. Davidson TJ, et al. Highly efficient small interfering RNA deliv-
ery into primary mammalian neurons induces microRNA-like ef-
fects before mRNA degradation. J. Neurosci. 2004;24:

10040–10046.


87. Judge AD, et al. Sequence-dependent stimulation of the mam-


89. Morrissey DV, et al. Potent and persistent in vivo anti-HBV ac-

tivity of chemically modified siRNAs. Nat. Biotechnol. 2005;23:

1002–1007.

90. Shin, D. et al. Immunostimulatory properties and antiviral activ-


Commun. 2007;346:436–442.

91. Eckstein F. The versatility of oligonucleotides as potential ther-


95. Davidson TJ, et al. Highly efficient small interfering RNA deliv-
ery into primary mammalian neurons induces microRNA-like ef-
fects before mRNA degradation. J. Neurosci. 2004;24:

10040–10046.


97. Judge AD, et al. Sequence-dependent stimulation of the mam-


99. Morrissey DV, et al. Potent and persistent in vivo anti-HBV ac-

tivity of chemically modified siRNAs. Nat. Biotechnol. 2005;23:

1002–1007.

100. Shin, D. et al. Immunostimulatory properties and antiviral activ-


Commun. 2007;346:436–442.

101. Eckstein F. The versatility of oligonucleotides as potential ther-


105. Davidson TJ, et al. Highly efficient small interfering RNA deliv-
ery into primary mammalian neurons induces microRNA-like ef-
fects before mRNA degradation. J. Neurosci. 2004;24:

10040–10046.


107. Grimm D, et al. Fatality in mice due to oversaturation of cel-

ular microRNA/short hairpin RNA pathways. Nature 2006;441:

537–541.


111. Persengiev SP, et al. Nonspecific, concentration-dependent stim-


113. Judge AD, et al. Sequence-dependent stimulation of the mam-

Aberrant transcription patterns are associated with most human diseases. Therefore, enormous interest exists in the development of designer molecules that can be used to regulate directly the transcription of predetermined genes for the ultimate treatment of a wide range of disease states. One emerging strategy is to identify molecules that reconstitute one or more functions of the endogenous proteins that upregulate transcription, transcriptional activators. In doing so, they function as either inhibitors or activators of transcription. Using this approach, a variety of protein- and small molecule-based transcriptional regulators have been developed, and at least one has reached clinical trials.

Diseased cells possess different transcription profiles relative to their normal counterparts; therefore, considerable interest exists in the discovery of molecules that correct errant transcription patterns for use as mechanistic tools and as therapeutic agents. One emerging mechanism for accomplishing this task is the use of molecules that mimic key functions of the endogenous proteins that upregulate transcription: transcriptional activators. By doing so, they either inhibit or activate the expression of specifically targeted genes (1). As illustrated in Table 1, exogenous agents that can upregulate or downregulate transcription are being developed for the eventual treatment of such ailments as cancer, inflammation, viral infections, metabolic disorders, and genetic disorders. For example, one method by which apoptosis can be induced in cancer cells is via the modification of the aberrant expression levels of those proteins that regulate cell growth and survival. A molecule that upregulates the proapoptotic bax gene directly could induce apoptosis when introduced into p53-deficient osteosarcoma cells (3). Conversely, a molecule that inhibits transcription of the survivin gene (an inhibitor of apoptosis protein) could induce apoptosis when introduced into lung carcinoma cells (9). In this article, we describe the structure and the function of natural transcriptional activators and outline the most common strategies for designing exogenous molecules that affect their function directly. References for more detailed treatments of the individual topics are provided at the end of the article.

Biological Background

As their name implies, eukaryotic transcriptional activators are responsible for initiating gene-specific transcription. To accomplish this task, activators localize at specific DNA sequences in a signal-responsive manner and facilitate the assembly of the eukaryotic transcriptional machinery (RNA polymerase II and associated factors) (1, 21). This process requires activators to participate in many protein-protein and protein-DNA interactions yet can be accomplished with a fairly simple architecture. Activators are composed minimally of a DNA binding domain (DBD) and a transcriptional activation domain (TAD). The primary function of the DBD is to localize the transcriptional activator to specific sites within genomic DNA. The DBD thus imparts much of the gene-targeting specificity of the activator. In contrast, the TAD participates in many protein-protein interactions that are critical for transcription initiation. By doing so, it dictates the timing and extent of gene activation. The two domains can, in general, function independently. In other words, if the DBD of transcriptional activator A is attached to the TAD of activator B, the new chimeric activator will upregulate transcription of gene A.

The modular character of transcriptional activators facilitates the design and the implementation of non-natural molecules that can affect gene transcription. Designer replacements of each of the two domains can be used individually to inhibit transcription by preventing either activator-DNA interactions or activator-transcriptional machinery interactions (Fig. 1). In contrast, linking a DBD and a TAD either covalently or noncovalently is a common strategy used to create activator artificial transcription factors (ATFs), which are molecules that seek out
<table>
<thead>
<tr>
<th>DBD</th>
<th>Regulatory domain</th>
<th>Gene target</th>
<th>Experimental model</th>
<th>Therapeutic application</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZF</td>
<td>VP64</td>
<td>γ-globin promoter</td>
<td>Cell culture (Endogenous gene)</td>
<td>Sickle cell disease</td>
<td>(2)</td>
</tr>
<tr>
<td></td>
<td>VP16</td>
<td>Bax promoter</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(3)</td>
</tr>
<tr>
<td></td>
<td>VP16, p65</td>
<td>VEGF-A gene</td>
<td>In vivo (Mouse, Rat)</td>
<td>Vascular disorders</td>
<td>(4, 7)</td>
</tr>
<tr>
<td></td>
<td>Gal4, Sp1</td>
<td>Utrophin promoter</td>
<td>Cell culture (Reporter plasmid)</td>
<td>Duchenne muscular dystrophy</td>
<td>(6)</td>
</tr>
<tr>
<td>TFO</td>
<td>None</td>
<td>MCP-1 promoter (Overlapping the Sp1 site)</td>
<td>Cell culture (Endogenous gene)</td>
<td>Inflammation</td>
<td>(7)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Ets2 promoter (Overlapping the Sp1 site)</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(8)</td>
</tr>
<tr>
<td></td>
<td>ATF4, ATF29</td>
<td>Survivin gene</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(9)</td>
</tr>
<tr>
<td>PNA</td>
<td>None</td>
<td>Translocated c-Myc (Eμ) enhancer (Overlapping the Ets-1 and AML-1 site)</td>
<td>Cell culture (Endogenous gene)</td>
<td>Burkitt's Lymphoma</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Kras5 gene (Mutant allele)</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(12)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>HNR-8 and A transcription start sites</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(13)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Gγ-globin 5′ flanking region</td>
<td>Cell culture (Endogenous gene)</td>
<td>Sickle cell disease</td>
<td>(14)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>G80BP-A 5′-AAGGAGGAGA-3′ binding sites</td>
<td>In vitro (Reporter plasmid)</td>
<td>—</td>
<td>(15)</td>
</tr>
<tr>
<td>Polyamide</td>
<td>None</td>
<td>HIV-1 (5′ LTR) promoter (Adjacent to the TBP, LEF-1, and Ets1 site)</td>
<td>Cell culture (HIV virus)</td>
<td>HIV replication</td>
<td>(16)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>VEGF promoter (Overlapping the HRE site)</td>
<td>Cell culture (Endogenous gene)</td>
<td>Cancer</td>
<td>(17)</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>5-TGTATAT-3′ binding sites</td>
<td>In vitro (Reporter plasmid)</td>
<td>—</td>
<td>(18)</td>
</tr>
<tr>
<td></td>
<td>Wrench2006</td>
<td>5-GCACTAT-3′ binding sites</td>
<td>In vitro (Reporter plasmid)</td>
<td>—</td>
<td>(19)</td>
</tr>
<tr>
<td></td>
<td>TBHK6</td>
<td>5-WG WWWWWW-3′ binding sites (IV= A or T)</td>
<td>Cell culture (Endogenous gene)</td>
<td>—</td>
<td>(20)</td>
</tr>
</tbody>
</table>
Transcription-Based Therapeutics

Inhibitors of activator-coactivator interactions: handful of small molecules

Inhibitors of activator-DNA interactions: polyamides, PNA TFOs, small molecules (natural & unnatural products)

Activator ATFs: minimally a DNA binding domain (DBD) coupled to a transcriptional activation domain (TAD)

Figure 1. Strategies for designing molecules that regulate transcriptional activators. Transcriptional activators initiate transcription by binding to DNA sequence specifically and stimulating the assembly of the transcriptional machinery through one or more protein interactions. Molecules that prevent either the DNA binding of the activator or the interaction of activators with their binding partners within the transcriptional machinery can be used to inhibit transcription. Alternatively, gene-specific transcription can be initiated by a molecule that mimics both key functions of an activator, which is an activator artificial transcription factor (activator ATF).

and upregulate the transcription of specific genes (22). The design of inhibitor and activator molecules relies on understanding the structure and the mechanism of the DBD and the TAD of the natural proteins. As illustrated in the subsequent sections, the DBD is more understood than the TAD; therefore, the development of artificial DBDs is far more advanced than TAD replacements.

Activator ATFs

As described above, the function of a transcriptional activator can be reconstituted minimally by an activator ATF that contains a DNA-binding domain and a transcriptional activation domain. The earliest artificial activators were composed of DBDs and TADs taken from naturally occurring proteins (1). Although these activator ATFs are powerful mechanistic tools, their application scope is narrow because they can only target genes that contain the DNA binding sites of the endogenous protein DBDs. In addition, controlling the delivery and the stability of the constructs in vivo can be challenging. The development of non-natural replacements for each of the two key activator ATF domains has been an important goal to address these fundamental limitations.

DBDs

Much work has been done to develop artificial DBDs that can bind with high specificity and affinity to predetermined DNA sequences (22). This binding has been achieved, for example, by mutating amino acids on natural protein scaffolds to recognize novel sequences (zinc fingers), using the hydrogen bonding properties of nucleic acid-like molecules [triplex-forming oligonucleotides (TFOs), peptide-nucleic acids (PNAs)], or tailoring the DNA-binding properties of natural products (polyamides, for example). This section describes only the properties of zinc fingers and polyamides, because TFOs and PNAs are described in depth in related articles. Table 1 summarizes many of the applications of these molecules.

Zinc fingers (ZF)

Protein DNA-binding domains offer several attractive features for activator ATF design; the most important of which is the high affinity and specificity with which they typically recognize their cognate DNA sequence. The Cys2His2 ZF fold has proven to be enormously versatile as a DNA-targeting entity. It is composed of ~30 amino acids folded into a ββα structure that is stabilized by hydrophobic interactions and by the coordination of a zinc ion by two conserved cysteine residues in the antiparallel β sheet and two histidine residues in the α helix. The solid-state structure of the 3-finger protein Zif268 in complex with DNA illustrates that each finger makes its primary sidechain–base interactions to three adjacent nucleotides in the sense strand of the DNA duplex (Fig. 2a). It does so by inserting its α helix into the major groove of DNA, on which amino acids at positions −1, 3, and 6 of the helix contact the 3′, middle, and 5′-nucleotides of the 3-bp subsite, respectively (23). Also, in some ZFs an aspartic acid at position 2 of the helix interacts with a cytosine or adenine base in the antisense strand of the adjacent triplet, which makes these domains seem to recognize a 4-bp subsite instead (23, 24).

One simple method for creating a ZF protein capable of binding to a predetermined DNA sequence is through the “modular assembly” approach. In this approach, pre-existing, single-finger ‘modules’ with known specificities are assembled into a multifinger array. To facilitate this, three archives of known zinc finger modules have been created by the Barbas laboratory, Sangamo BioSciences Inc., Richmond, CA and ToolGen Inc. Seoul National University, South Korea. The Barbas modules were developed using a combination of phage display and rational design methods under the assumption that ZF domains...
function with position independence. They are capable of recognizing all GNN triplets, most ANN and CNN triplets, and a few TNN triplets (N = any base). The Sangamo modules were also developed by phage display but under the assumption of position dependence, and are capable of recognizing all GNN triplets and a smaller number of non-GNN triplets. And finally, the ToolGen modules are naturally occurring human zinc fingers whose nucleotide triplet sequences were identified through a yeast one-hybrid assay. And in collaboration with the Zinc Finger Consortium, these archives are now available through a web-based server called ZiFiT (Zinc Finger Targeter) that facilitates the design of multifinger arrays that bind to your desired DNA sequence (24).

In practice, ZF proteins composed of 3–6 fingers with apparent dissociation constants in the picomolar to nanomolar range have been attached to proteinaceous transcriptional activation domains and used successfully to upregulate endogenous genes in mammalian cell culture (Table 1). One such activator ATF that contains a six-finger DBD that binds within the \( \gamma \)-globin promoter can increase fetal hemoglobin levels 7–16 fold in human erythroleukemia cells (2). Additionally, ZF-based activator ATFs that target the VEGF-A (vascular endothelial growth factor) gene have even been shown to function in animal models (4, 5); despite this success, stable delivery remains a challenge because they must be administered by viral vectors (23).

Polyamides

Small molecule DBDs represent an attractive choice for activator ATF construction because they may circumvent the delivery limitations of proteins. Toward this end, considerable progress has occurred in developing programmable small molecules that can be designed readily to target a wide range of DNA sequences (22). In particular, the polyamide class of DBDs has been used successfully for the construction of several activator ATFs (25). The inspiration for the polyamides developed from the minor groove-binding natural products distamycin and netropsin, which exhibit greatly enhanced DNA binding affinities relative to distamycin and netropsin, with dissociation constants in the picomolar to nanomolar range, and they have been shown in several applications to traffic to the nucleus and to interact with their cognate DNA sites. Enhancing their use even more, Dervan et al. (25) have developed a set of “pairing rules” that can be used to design molecules to target specific DNA sequences. In addition, the molecules can be prepared by solid phase synthesis, which makes them accessible to many users.

Polyamides have been used as the basis for several different activator ATF constructs that function in cell-free and in cellular systems. In contrast to protein DBDs, they are synthesized easily to contain both peptidic and nonpeptidic TADs (19–20, 26). However, they often require special modifications to enhance cellular permeability and typically target shorter DNA sequences (6-8 base pairs) relative to proteins (27).

TADs

The most common TADs used in the construction of activator ATFs are derived from the activation domains of natural proteins. For example, sequences taken from the amphipathic class of activators are composed of aromatic amino acids linked through amide bonds; they bind to A/T-rich tracts of DNA in the minor groove with moderate affinity through a combination of hydrogen bonds between the amide bonds and the minor groove functional groups, hydrophobic contacts, and electrostatic interactions with the phosphate backbone. Polyamides consist not only of pyrrole amino acids, but also of imidazole, pyrrole, and other heterocyclic amino acids that enable recognition of A•T, T•A, C•G, and G•C base pairs through the formation of specific hydrogen bonds with minor groove functionality, although overall specificity varies with sequence context (Fig. 2b). Although several different polyamide structural motifs exist, the hairpin polyamide in which a flexible amino acid tether connects two polyamide arms is used most commonly. As the name suggests, this molecule folds into a hairpin-like structure in the minor groove such that the arms are side-by-side, which maximizes hydrophobic interactions with the walls of the minor groove and facilitates the formation of polyamide-DNA hydrogen bonds. The molecules exhibit greatly enhanced DNA binding affinities relative to distamycin and netropsin, with dissociation constants in the picomolar to nanomolar range, and they have been shown in several applications to traffic to the nucleus and to interact with their cognate DNA sites. Enhancing their use even more, Dervan et al. (25) have developed a set of “pairing rules” that can be used to design molecules to target specific DNA sequences. In addition, the molecules can be prepared by solid phase synthesis, which makes them accessible to many users.
Transcription-Based Therapeutics

Interspersed with polar ones and typically possess robust activity across organisms (1, 22). For example, activating sequences from the viral protein VP16, the yeast activator Gal4, and the p65 subunit of the human activator NF-κB have all been attached to ZF proteins and function as activator ATFs in mammalian cell culture (2–4, 6). However, activator ATFs that contain nonprotein DBDs typically use much smaller sequences to minimize the overall size of the construct. For instance, a monomeric or dimeric repeat of eight residues of VP16 (VP1 and VP2) could upregulate a reporter gene in vitro using yeast nuclear extracts when attached to a polyamide (18), whereas a monomeric or dimeric repeat of an 11 residue sequence taken from VP16 (ATF14 and ATF29, respectively) could upregulate a reporter gene in mammalian cell culture when attached to a TFO (10).

Several strategies have been employed to develop novel peptidic TADs that function similarly to natural TADs. For example, a 20 amino acid peptide sequence designed rationally to form an amphipathic helix (AH) (Fig. 3a) was successful in upregulating a reporter gene in yeast when fused to the Gal4 DBD (28), as well as in vitro with yeast nuclear extracts when fused to a polyamide DBD (18). In addition, another successful approach has been to use phage display peptide libraries to select against the protein targets of natural TADs. For instance, a selection performed against the KIX domain of the mammalian coactivator p300/CBP yielded an 8-amino acid peptide named KBP 2.20 that is capable of upregulating a reporter gene 40-fold in mammalian cells when attached to the Gal4 DBD (29). Additionally, a selection performed against the masking protein (Gal80) of the yeast activator Gal4 yielded a 20-amino acid peptide named G80BP-A that is capable of activating transcription of a reporter gene in both yeast (30) and mammalian (15) cell culture when attached to the Gal4 DBD. This TAD functioned only weakly when a PNA was used as a DBD, however, because of the distortion of the promoter on binding (15).

Finally, nonpeptidic TADs have been developed recently that possess the advantage of increased stability toward proteolytic degradation. Often, these molecules are considerably smaller than protein-derived TADs and thus may exhibit advantageous cell permeability properties. The first small molecule TAD to be reported was an amphipathic isoxazolidine 1 (Fig. 3b). Its ability to upregulate transcription was demonstrated initially in a cell-free assay where it proved to be as active as a larger peptidic activator (MW 290 vs 1674) (31). This amphipathic isoxazolidine also functions in mammalian cell culture in a dose-dependent manner with up to 80-fold activation at 1µM and an EC_{50} of 31 nM (32). In addition, a hydrophobic molecule named wrenchnool designed originally as an inhibitor of the ESX-Sur2 interaction is also capable of functioning as an activation domain on localization to DNA. When conjugated to a polyamide, this molecule upregulated transcription in a cell-free system 3.5-fold over background. This synthetic activator ATP was, however, inactive in mammalian cell culture, which was possibly caused by limited nuclear localization (19). Finally, peptoids (oligo-N-substituted glycines) are emerging as an effective alternative to peptidic TADs. KBP02, which is a peptoid that was identified from a library screen against the KIX domain of CBP, is capable of activating robustly (up to 1000-fold) transcription of a reporter gene in mammalian cells.
DNA–protein interactions

One method by which transcription can be modulated passively is by targeting the promoter of a gene with artificial DBDs. Using this method, inhibition of transcription can be achieved by competing for the binding sites of endogenous transcriptional activators (22). For example, TFOs designed to target a sequence that encompasses the Ets-1 and AML-1 sites in the hypoxic response element (HRE) in the VEGF promoter can block RNA polymerase II during elongation. For example, a PNA that was designed to target a region encompassing a point mutation in codon 12 of the K-RAS proto-oncogene showed a concentration-dependent ability to downregulate expression of the mutant allele over the wild type in pancreatic cancer cells (12). In part because of their lower binding affinities, polyamides must be conjugated to DNA-modifying agents to arrest a transcribing polymerase. For example, a polyamide conjugated to the DNA alkylation chlorambucil exhibited cytostatic activity in colon carcinoma cells; this effect was attributed to the downregulation of the H4c gene—2-fold, most likely because of the alkylation of a G residue located two bases downstream of a target site present in its coding region. This alkylating polyamide was even effective at suppressing tumor growth in a dose-dependent manner in a mouse model with no obvious toxicity (38). On the other hand, unconjugated polyamides have been used successfully to alleviate blockage of an elongating RNA polymerase by targeted expanded intronic repeat sequences and by preventing formation of non-B DNA structures (39). Finally, in addition to inhibiting elongation, PNA s also can block RNA polymerase II at the initiation step of transcription. This action can be achieved by targeting ssDNA sequences in the open complex located at the transcription start site, and it has been applied successfully with PNA s to downregulate expression of the human progesterone receptor (hPR) isofoms A and B in breast cancer cells (33).

Protein–protein interactions

Another method by which transcription can be modulated passively is by inhibiting the interaction of natural activators with their protein targets (Fig. 4). Transcriptional inhibition can be achieved by blocking activator-coactivator interactions. For example, the histone acetyltransferase CBP/p300 is a global coactivator capable of binding to a diverse group of activators including CREB, c-Myb, Jun, and HIF-1α (40). To identify a molecule that can disrupt the interaction of HIF-1α with p300, a high-throughput competition binding screen was performed against the minimal protein complex, and a small-molecule fungal metabolite named chetomin was identified (Fig. 4). Chetomin not only reduced endogenous HIF-1α/p300 complex formation in cells, as demonstrated by coimmunoprecipitation experiments, but also reduced the activation of hypoxia-responsive reporter genes both in cell culture and in vivo. On probing the specificity of this inhibitor against other p300-dependent activators such as RAR, SREBP2, SRC-1, and STAT2, only the activity of STAT2 was significantly attenuated, and this is most likely because it is the only one that, along with HIF-1α, targets the CH1 domain on p300 (41). Overall, this example illustrates the challenges of inhibiting activator function selectively by this mechanism. For instance, although the molecule KG-561, which inhibits the IKK/NF-κB interaction within the CREB/CBP complex, could inhibit significantly transcription of CREB-dependent genes in human cells, preliminary studies indicate that it also impedes activation by another CBP-dependent activator NF-κB (42). In addition, although inhibiting complex formation between the coactivator p300 and the activator Tcf4 with the fungal metabolites PKF113-584 and CP949091 leads to antiproliferative effects in colon cancer...
Inhibitors of activator–coactivator and activator–masking protein interactions. Structures of small molecules that compete effectively with transcriptional activators for binding sites within coactivators and within masking proteins.

Cells, these molecules also can prevent complex formation between β-catenin and the tumor suppressor protein APC, which impacts their effectiveness (43). Taken together, these results demonstrate that inhibitors of activator–coactivator interactions have enormous potential and are in need of additional evaluation.

Inhibiting protein interactions can also lead to the activation of transcription. Activators are often regulated by masking proteins that bind to the TAD, which prevents it from contacting the transcriptional machinery until upregulation is required. Typically, TAD-masking proteins are of higher affinity and specificity than TAD-transcriptional machinery interactions and are therefore more straightforward to target with small molecules.

For example, the p53 TAD binds as a helix to a relatively deep hydrophobic cleft in the protein MDM2, and a small molecule library screen for inhibitors of this interaction yielded a series of cis-imidazoline analogs termed Nutlins. The active enantiomer of Nutlin-3 inhibits recombinant p53/MDM2 complex formation in vitro with an IC50 value of 0.09 µM, and activates p53 (thereby inducing apoptosis) in cancer cells that contain wild-type p53. When evaluated in vivo, Nutlin-3 treatment of osteosarcoma xenografts established in nude mice results in a 90% inhibition of tumor growth (44). In addition to the Nutlins, many other inhibitors of the p53/MDM2 interaction that exhibit biologic activities have been reported (40). Terphenyl 2 increases p53 activity by 10-fold in colon cancer cells at a concentration of 40 nM (45). In tumor cells that overexpress MDM2, sulfonamide 3 (IC50 value of 32 µM) increases p53 activity by 20% (46). Benzodiazepine 4 binds to MDM2 with an IC50 value of 30 nM (47). Finally, an isindolinone with an IC50 value of 5 µM (5) induces p53-dependent transcription in a dose-dependent manner in MDM2 overexpressing human sarcoma cells (48).

Other Approaches

Additional strategies are available to alter gene expression patterns for therapeutic purposes, many of which are discussed in other articles of this volume. In most cases, these approaches target processes that are upstream or downstream of transcription and can thus influence a wide array of genes both positively and negatively. One strategy involves targeting the...
ligand-dependent class of transcription factors known as nuclear receptors with small molecule agonists or antagonists to promote their interaction with coactivator or corepressor proteins, respectively. For example, metabolic disorders caused by malfunctioning PPAR nuclear receptors can be alleviated by targeting the γ isoform with an agonist class of molecules called thiazolidinediones for the treatment of diabetes (49). Another strategy involves targeting protein kinases with small molecule inhibitors to intervene at some stage in the signaling cascade that precedes the process of transcription. For instance, in the treatment of inflammation, one pathway by which the non-steroidal anti-inflammatory drug aspirin and sodium salicylate exert their effects is by inhibiting the binding of ATP to the kinase IKK-γ, which in turn prevents the phosphorylation (and ultimately the degradation) of the cytoplasmic sequencin protein of the transcription factors NF-κB (50). And finally, the HDAC inhibitor SAHA was approved by the FDA in October 2006 for the treatment of cutaneous T-cell lymphoma, and at least 10 more compounds are currently in clinical trials in hopes of finding inhibitors for market against every major tumor type (51).

Targeting the transcript of a particular gene using the related methods of RNAi and antisense is another powerful strategy for manipulating expression. Currently, two siRNAs that target the VEGF transcript are in clinical trials for the treatment of macular degeneration, as well as one for the treatment of respiratory syncytial virus infection. Although many antisense oligonucleotides are in various phases of clinical trials, to date only fomiviren has been approved for the treatment of cytomegalovirus retinitis (52). One disadvantage of this strategy is that by targeting the transcript rather than a sequence within the genome, these methods usually require higher concentrations to produce an efficient outcome (23).

Future Directions

Tremendous progress has been made toward developing activa
tor ATFs as therapeutic agents for the treatment of a variety of disorders. Indeed, ZF-based ATFs are currently in clinical tri-
als (http://www.sangamo.com/index.php) and a purely synthetic activator that upregulates the transcript levels of endogenous genes in cell culture has been developed. The next frontier is the incorporation of additional functionality into activator ATFs that would confer properties such as temporal control and tissue specificity analogous to endogenous activators. The three-hybrid approach in which the TAD and the DBD associate noncova-
sently only in the presence of a small molecule was the first demonstration that temporal control could be engineered in an ATF, and this pioneering approach has been used in a variety of contexts (53). For entirely synthetic ATFs, one recently reported strategy exploits conformational entropy to create protein-DNA dimers that are inactivated at elevated temperatures but func-
tion well at a lowered temperature (54). In all examples, delivery of the molecules to the appropriate tissues and to the nuclei of those tissues remains an open challenge. Creation of a fully functional activator ATF will thus stimulate advances on a va-
riety of scientific fronts.

References

2. Grassl T, Li K, Magnent L, Popkov M, Barbac CI. III. Exp-
loiting strategies for the design of artificial transcription factors: targeting sites proximal to known regulatory regions for the in-
duction of gamma-globin expression and the treatment of sickle
3. Falke D, Fisher M, Ye D, Juliano RL. Design of artificial tran-
scription factors to selectively regulate the pro-apoptotic bag. Nuc-
6. Corbi N, Libri V, Fanucilli M, Timsend J, Davies KE, Pas-
7. Marchand F, Rus K, Radake W. Selective inhibition of mono-
cyte chemokine receptor protein-1 gene expression in human embry-
15. Liu B, Han Y, Fanbous A, Casey DR, Kodak D. Transcrip
Transcription-Based Therapeutics


See Also

- Gene Therapy and Cell Therapy
- Peptide Nucleic Acids
- Protein-Nucleic Acid Interactions
- Protein-Protein Interactions
- Small Molecules to Disrupt Protein-Protein Interactions
- Transcription Factors
- Transcription, Activators and Repressors of Transcription, Initiation of

Further Reading

Chemistry of Neoglycoproteins

Yuan C. Lee, Biology Department, Johns Hopkins University, Baltimore, Maryland
doi: 10.1002/9780470048672.wecb379

Natural glycoproteins often contain heterogeneous oligosaccharide chains, which make it difficult to elucidate the structure–activity relationship of the carbohydrates of these glycoproteins. Neoglycoproteins were developed to overcome such a problem. Neoglycoproteins are proteins modified with carbohydrates of defined structures. The modification can be random or specific and can use chemical or enzymatic methods. The protein functional groups most often used are $\alpha/\epsilon$-amino groups, $\beta/\gamma$-carboxyl groups, $\gamma$-carboxylamides, sulfhydrils, and phenols. In some cases, existing glycans on glycoproteins can be modified to make the glycan structures homogeneous. Alternatively, glycopeptides or glycoproteins can be synthesized totally de novo, without using any natural proteins, peptides, or oligosaccharides. In addition to providing homogeneous carbohydrate structures on a single protein, neoglycoproteins also provide a multivalency effect (glycoside clustering effect). The use of neoglycoproteins greatly promoted the understanding of the roles of carbohydrates in biologic systems, which led to some clinically important development.

Glycosylation is one of the most complex posttranslational modification of proteins (1). It is well known that glycans attached to glycoproteins are heterogeneous, even at a single glycosylation site of a glycoprotein from a single type of cells. Two reasons for this heterogeneity are that multiple possibilities of linking two monomeric units of carbohydrates exist and that carbohydrates have the ability to form branched structure routinely. Branched structures are especially pervasive on the cell surface glycoconjugates. These factors cause great difficulties in elucidating the biologic roles of carbohydrates in glycoproteins and have hampered considerably the progress in this area of research in comparison with proteins and nucleic acids. To partially alleviate such a problem, neoglycoproteins (2) were devised. In neoglycoproteins, carbohydrates attached to proteins are of a known structure and the level of attachment can be varied more broadly than that found in natural glycoproteins. The latter aspect actually is very important in manifesting the "cluster effect" (multivalency effect) (3, 4), for example, a linear increase in valency that produces logarithmically increasing effects. Production of neoglycoproteins needs not start from natural proteins or glycoproteins but can be designed totally de novo. The de novo approach often is used to attain the desired "cluster effect" more precisely than the random attachment of sugars to proteins. Most of methodologies for the preparation of neoglycoproteins are chemical, but incorporation of enzymatic and molecular biologic methods increasingly are gaining popularity. In this article, the emphasis is placed primarily on protein derivatives (neoglycopeptides), but some examples of peptide derivatives and microarrays also are included.

Historical Background

The first neoglycoproteins (without being referred to as such) were prepared by Avery and Goebel, who used sugar-derivatized proteins to raise antibodies directed toward carbohydrates (5, 6). Typically, carbohydrates by themselves are poor antigens. Avery and Goebel demonstrated that by conjugating carbohydrates to proteins, antibodies against carbohydrates can be raised more efficiently than by carbohydrates alone. Avery and Goebel used diazoated $p$-aminophenyl or $p$-aminobenzyl derivatives of sugars for conjugation to tyrosyl side chains of protein (diazo-coupling). Using a similar approach, Monsigny and colleagues (7, 8) prepared (GlcNAc)$_n$–BSA to determine the binding specificity of wheat germ agglutinin. Iyer and Goldstein (9) also used the diazo-coupling technique, but Buss and Goldstein (10) improved the coupling efficiency by first converting the anilino-group of the glycoside aglycon into the isothiocyanate group by reacting it with thiophosgene so that the sugar derivatives will react with the amino groups of proteins rather than with the phenolic groups of proteins. Lee and colleagues (11-14) used thioglycosides that contained various functional groups to conjugate to proteins, and the term "neoglycoprotein" was introduced at this time. Thioglycosides were chosen...
NaCNBH₃ preferentially reduces the Schiff base (aldimine) to 
amiceric amination (15–17). Contrary to sodium borohydride,
principles of various conjugation methods is shown in either 
nonglycosylated or glycosylated. A summary of chemical
Aminophenyl glycosides had been used in a similar fashion to hydrate to conjugate to Tyr groups of globulins and albumins.

The classic method of Avery and Goebel (5, 6) and Tillett via 
tyrosyl groups of proteins
Most neoglycoproteins are prepared using natural proteins—
From Natural Proteins
Preparation of Neoglycoprotein From Natural Proteins
Most neoglycoproteins are prepared using natural proteins—
either nonglycosylated or glycosylated. A summary of chemical principles of various conjugation methods is shown in Fig. 1.

Via tyrosyl groups of proteins
The classic method of Avery and Goebel (5, 6) and Tillett et al. (10) used diazotized aminobenzyl derivatives of carbohydrate to conjugate to Tyr groups of globulins and albumins. Aminophenyl glycosides had been used in a similar fashion to prepare BSA derivatives as mentioned above (8). Such products can be very colorful because of the highly conjugated bonding system. However, usually only a limited number of Tyr groups exist in typical proteins, and the diaco coupling is not totally specific to phenolic side chains. Moreover, the chemical stability of products also is a problem. It should be noted that the modification of Tyr often results in the inactivation of enzymes and other biologically active proteins. The introduction of an aromatic ring for each glycosylation site also considerably increases the hydrophobicity of the product, which results in decreased solubility and increased undesirable nonspecific interaction.

A unique approach of carbohydrate-protein conjugation was reported in the case of chitosan. The target protein first is oxidized with tyrosinase to yield o-quinone groups, which can be conjugated readily to chitosan (22). This approach should be applicable to amino-terminated glycosides also.

Via amino groups of proteins
This approach is by far the most popular approach for the preparation of neoglycoproteins because of the ready surface availability of the ε-aminogroup of Lys side chains and the higher reactivity of ε-amino-groups in most proteins.

Reducing oligosaccharides can be conjugated directly to the amino groups by the reductive alkylation (or reductive amination) method (13) between the carbonyl group of sugars and the amino groups of proteins if a prolonged reaction time is no objection. The slow reaction rate in the reductive amination is because of the extremely low concentration of the acyclic form of sugars in solution that provides the reactive aldehydic group. In the case of oligosaccharides, a compounding effect is a decrease in reactivity as the molecular weight increases. This problem can be overcome by providing the ω-aldehyde group in the alkyron (13, 18), which allows the completion of the reactive alkylation overnight. Another alternative is to reduce the reducing terminus and then use mild periodate oxidation to generate aldehyde group(s) from the acyclic sugar. When oligosaccharide lactones (prepared from bromine oxidation of the reducing oligosaccharides) were used to conjugate to BSA, many days also were required to complete the reaction (22).

Many other reactive derivatives of mono- or oligosaccharides for reactions other than reductive alkylation have been described (Fig. 1). For example, isothiocarbamates was derived from p-aminophenyl glycosides (10). Glycinate was used to conjugate to the amino groups of proteins (23). For more complex reactions, methyl imidate (derived from cyanomethyl thioglycosides), for example, would react with amino groups to form an amindia link, which has an even higher pKa than the ε-aminogroup (12). Various activated form of ω-carboxyl thyglycosides have been used (24–26) to modify amino groups. It should be noted that although amidation converts the amino group into a neutral amide group, reductive alkylation and amidation retain the positive charge of the original amino group, and the protein integrity is better preserved. When the number of amino groups is high (such as in BSA), it is possible to attach a greater number of glycan chains than normally observed in natural glycoproteins.

Via sulfhydryl groups
If a protein contains a surface-oriented free side chain of cysteine, ω-haloacetyl that contains glycoside or glycopeptide can be reacted to form theimide (27). ω-Haloacetylated glycosides or maleimidated glycopeptides (28) also can be used in the same way. However, a free surface-oriented SH in natural proteins/glycosides is rather rare in number, and hence the scope of this type of neoglycoprotein preparation also is limited. If glycosylation is desired at a specific site and if a Cys can be introduced in the peptide sequence by genetic engineering, it may allow a custom-designed neoglycoprotein with site-specific glycosylation. It also is possible to modify the amino groups in protein with haloacetyl or maleimide groups.
that then can react with thiolated carbohydrate derivatives (e.g., 1-thio-oligosaccharides) to form a thioether link. In this case, the number of glycan attachments (on the amino groups) can be far greater compared with the naturally available number of SH side chains.

An alternative to thioether formation is to form a disulfide bond between thioglycoside and the Cys–SH group of protein (29) or by disulfide exchange between disulfide glycoside and Cys–SH (30). The glycan attached via the disulfide bond, however, has a real risk of being detached by a reductive/oxidative cleavage or a sulfide exchange in biologic systems and, thus, is less desirable.

Via special affinity

Chen and Wold (31, 32) used the unique affinity of avidin/streptavidin for biotin to prepare neoglycoproteins. Glycosides that contain a biotin terminus could be bound so tightly to avidin (Kd < 10^{-13} M) that they could be considered virtually covalently bound. The limitation of this method, however, is that the maximum number of glycan attachments is dependent on the valency of avidin, which is four, and the attached glycans face diverse directions. Interestingly, the fourth binding site of avidin turned out to be very difficult to fulfill. The model of biotinylated glycan and avidin also was used to study glycosylation processing (33). It should be noted that avidin itself is a glycoprotein that contains heterogeneous oligosaccharides and that its nonglycosylated recombinant version (NeutralLite) (34) is more suitable for the preparation of definitive neoglycoproteins.

Via enzymatic action

A transglutaminase reaction was found to be useful in the preparation of neoglycoproteins. For example, ω-amino glycosides were used as a substrate for transglutaminase for the addition of carbohydrates to glutamine side chains (35, 36). The success of this reaction, however, seems somewhat unpredictable.

Some endo-β-hexosaminidase were found to be useful in neoglycoprotein preparation. Although the normal function of these enzymes is to cleave the bond between the two GlcNAc residues next to asparagines of the N-glycan attachment site, they can perform an en block transfer of oligosaccharides (e.g., Man9–GlcNAc) to a GlcNAc-terminated acceptor quite effectively under certain conditions (Fig. 2). Endo-A and endo-M have been used to transfer Man9–GlcNAc (endo-A) or a biantennary complex-type oligosaccharide (endo-M) to the GlcNAc–Asn site of glycoproteins, which can be generated by the hydrolytic action of the endo-β-hexosaminidase (37-39). Whereas endo-A is active only on high-mannose-type glycans, endo-M accepts complex-type glycans as a substrate as well, more effectively for biantennary than triantennary structures. Endo-M has been used for modification of eel calcitonin effectively (38, 39). In either case, the acceptor GlcNAc need not be attached to the Asn. GlcNAc itself and its glycosides are known good acceptors.

A very interesting and useful application of endo-A was reported recently by Wang and coworkers (40, 41). Based on the knowledge that some endo-glycosidase effectively can transfer oxazoline derivatives of GlcNAc-oligosaccharides, Wang and coworkers used endo-A on the oxazoline derivative of...
**Chemistry of Neoglycoproteins**

De Novo or Semisynthetic Preparation of Neoglycoprotein

The advancement in the techniques of the chemical synthesis of oligosaccharides and peptides (including molecular engineering) elevated the neoglycoprotein preparation to the *de novo* synthesis level, which opened a wide range of application possibilities.

Via incorporated Cys in the peptide

When a Cys can be incorporated into the peptide sequence by means of bioengineering, thioaldooses can be attached via a disulfide bond with the newly introduced SH group in the protein. In a recent example (29), asparagine (N-297) of IgG–Fc was replaced with Cys and its SH group was used to form a disulfide bond with thio-oligosaccharide (29). When 3-thio-derivative of Man₉GlcNAc₂ was attached to the thio-modified IgG by such a scheme, the Fc-mediated IgG potency increased several fold over that of the IgG that contains the natural Fc. The distance of glycan to the peptide backbone attached by this method is approximately the same distance as in the natural N-glycan.

Via native chemical ligation

The technique of native chemical ligation (NCL) or expressed protein ligation (EPL) (42) tremendously expanded the scope of peptid/protein synthesis. This approach also has been adopted for the construction of neoglycoproteins. For complete synthetic approaches to the preparation of defined glycoproteins, several excellent reviews (19, 42, 44) exist. The basic principle of NCL is illustrated in Fig. 3 (45). In this article, only a few examples are mentioned. Macmillan and Bertozzi (46) used *Escherichia coli*-expressed peptides and chemically synthesized glycopeptides to construct GlyCAM-1, a natural ligand for L-selectin. Similarly, Tolbert et al. reported chemical ligation of synthetic glycopeptides with *E. coli* expressed and TEV protease-cleaved peptides (19).
Chemical synthesis

One prototypic example of chemoenzymatic synthesis is the preparation of “glycotentacles” (36). First, cyclic peptides that contain several glutamine residues were synthesized. Gal-terminated glycan chains were attached to the glutamine residues by transglutaminase reaction, which then was sialylated enzymatically with sialyltransferase to obtain NeuAcα2,3Gal-terminated glycans. This approach provides flexibility in the size of the cyclic peptide and the choice of glycan structures. As mentioned earlier, the transglycosylation by endo-A was facilitated greatly when the reducing terminal GlcNAc was converted chemically to its oxazoline derivative first (41, 47). A synthetic mucin based on a poly (ω-L-glutamic acid) backbone decorated with p-aminophenyl-glycoside of LacNAc followed by sialylation with sialyl transferase was prepared as an anti-influenza agent (48).

Although the use of glycosyltransferases requires more expensive sugar nucleotides, the exquisite specificities offered by these enzymes sometimes override the cost concerns. Glycosylation often are used in terminal modifications such as sialylation and fucosylation.

Other techniques and reagents

The ease of “click chemistry” (49) has attracted many applications in the area of neoglycoconjugates. The principle of “click chemistry” is illustrated in Fig. 4. The conjugation based on click chemistry can be performed between either the ω-alkyn group on a protein or vice versa (50). The popularity of “click chemistry” resides in its ease of operation, its indifference to reaction conditions (solvent, pH), and its high yields. A traceless Staudinger ligation was reported for its high yields. A traceless Staudinger ligation was reported for (47).

A related reaction using the Diels–Adler cycloaddition of a diene-containing glycoside and a maleimide-equipped protein gives neoglycoproteins at ambient temperature in good yield (51). Preparation of “diene” glycosides may be the limiting step in this application. Squaric acid esters (52) sometimes are used to link glycans to proteins. N-Acetyl-chito-oligosaccharides were linked to silk fibroin by reacting solubilized silk fibroin with cyanuric chloride-activated oligosaccharides (Fig. 5a). One prototypic example of chemoenzymatic synthesis is the preparation of “glycotentacles” (36). First, cyclic peptides that contain several glutamine residues were synthesized. Gal-terminated glycan chains were attached to the glutamine residues by transglutaminase reaction, which then was sialylated enzymatically with sialyltransferase to obtain NeuAcα2,3Gal-terminated glycans. This approach provides flexibility in the size of the cyclic peptide and the choice of glycan structures. As mentioned earlier, the transglycosylation by endo-A was facilitated greatly when the reducing terminal GlcNAc was converted chemically to its oxazoline derivative first (41, 47). A synthetic mucin based on a poly (ω-L-glutamic acid) backbone decorated with p-aminophenyl-glycoside of LacNAc followed by sialylation with sialyl transferase was prepared as an anti-influenza agent (48).

Although the use of glycosyltransferases require more expensive sugar nucleotides, the exquisite specificities offered by these enzymes sometimes override the cost concerns. Glycosylation often are used in terminal modifications such as sialylation and fucosylation.

Cluster effect (or multivalency effect)

In addition to providing structurally well-defined glycans on a protein carrier, neoglycoproteins offer a platform for presenting multiple copies of glycans in a limited confine. The first indication of the effectiveness of the clustering of sugars on neoglycoprotein was observed in the study of neoglycoproteins binding by rabbit liver membrane (11). The binding of BSA-based neoglycoproteins that carry several different sugars by rabbit liver membrane revealed that Gal-BSA was bound very effectively but neoglycoproteins of other sugars (GlcNAc, Man, L-Fuc) were totally ineffective, which indicates that the sugar binding is specific for Gal. The more astounding finding was that when the number of Gal on BSA increased linearly, the binding affinity increased logarithmically, which resulted in a subnanomolar binding affinity (as KI) at the Gal content of about 40 residues per molecule of BSA. The monomeric Gal derivative had an affinity of only ca. mM in KI. This effect was demonstrated more dramatically by the synthetic oligosaccharide mimetics of natural Gal-terminated N-glycans of mono-, di-, and tri-antennary structures (58), in which a 2- and 3-fold increase in valency brought about a ca. 1000-fold increase, respectively, in the binding affinity. Other examples of cluster effect are binding intricately synthesized pentamic and decameric clusters of galabiose (Gal–Gal)–containing structures to Shiga-like toxin (59) and Man-containing dendritic...
Chemistry of Neoglycoproteins

Figure 5 Squaric acid (52) and cyanuric acid (53) as linkers in neoglycoproteins.

Figure 6 Preparation of glycoconjugate by one-pot reaction (57).

Vaccines
Avery and Goebel (5, 6) and Tillett et al. (20) constructed their neoglycoproteins for raising vaccines against carbohydrate groups. The attachment of carbohydrates to proteins apparently made the carbohydrates more antigenic. The preparation and the use of neoglycoproteins for antibody generation is a very important endeavor (64). In fact, many vaccines currently in use (Hemophilus b conjugate, meningococcal 4-valent conjugate, and pneumococcal 7-valent conjugate) are neoglycoproteins in one form or another (65, 66). The antigenicity of dextran-chicken serum albumin was studied with respect to glycan size and density (67).
Lectin specificity and glycosyltransferase substrates

The groups of Monsigny and colleagues (7, 8) and Iyer and Goldstein (9) prepared neoglycoproteins for specificity studies of plant lectins (7, 8). The specificity studies later were extended to animal lectins (4, 68–70) with excellent results. Neoglycoproteins also can serve as substrates for glycosyltransferases in the chemo-enzymatic preparation of more complex glycan structures. In addition, neoglycoproteins are useful as substrates in the assay for glycosyltransferases. Examples are β-1,4-galactosyltransferases (71, 72) and α-fucosyltransferase (73).

Isolation Media and Probes

Many animal lectins have been isolated and purified using neoglycoprotein conjugated to agarose beads. C-type lectins are especially amenable to purification using neoglycoprotein-affinity gel because of the ease of elution simply by using EDTA-containing buffers. The examples include the use of GlcNAc-c-BSA-Sepharose for chicken hepatic lectin (74) and Man-BSA-Sepharose for alligator hepatic lectin (70). Neoglycoproteins also are useful as probes for sugar-binding proteins. Fucose-specific adsorptions on gel tubes of Candida albicans were demonstrated with fluorescence-labeled BSA derivatives bearing different sugars (75). More recent and elegant application was the use of de novo-designed glycoproteins to study the ER-quality control process as related to the activity of glycosyltransferase (76). T-antigen bearing neoglycoprotein was found to be useful as probes for breast carcinoma (77).

Glycoarrays

Microarrays of glycans (glycoarrays), which can be regarded as a special form of neoconjugates, are proving to be a powerful tool for glycomic analysis (78–81). Many above-mentioned conjugation methods and some other methods have been used in the construction of glycoarrays (82). For example, amino-terminated glycosides were reacted with carboxyl derivatives on a glass slide (amide formation) (83, 84) and reducing oligosaccharides were conjugated to 1,2-dihexadecylglycerol-3-phosphoethanolamine (DHPE) and the resulting NGL was adsorbed directly on a solid surface (86). A modified version of NGL uses amine derivatives of DHPE (87). In a recent report, sulfation specificities of glycosaminoglycan interactions with growth factors and chemotactic proteins were probed with microarrays (88).

References

catalyzed coupling to create protein–chitosan conjugates. Bio-

22. Anakata T, Adnell G, Katab EA. Immunochemical studies on
dextrins. V. Specificity and cross-reactivity with dextrins of
the antibodies formed in rabbits to isomaltosonic and isoma-

23. Loringen J, Goldstein JJ, Niederhuber JE. Aldonate coupling,
a simple procedure for the preparation of carbohydrate-protein
conjugates for studies of carbohydrate-binding proteins. Arch.

24. Leniveau R, Bunde DR, Baker DA. The properties of a “syn-
thetic” antigen related to the human blood-group Lewis a. J. Am.

Specific antibodies to diphtheria toxin and type A4 pneumococcal
capsular polysaccharide induced by a model of semi-synthetic

phenyl ester coupling reagent for the preparation of neoglyco-

27. Watt GM, Boons GJ. A convergent strategy for the preparation of
N-glycan core di-, tri-, and pentasaccharides triadoses for the
site-specific glycosylation of peptides and proteins bearing free

AJ. UDP-Glc-glycoprotein glucosyltransferase recognizes struc-
tured and solvent accessible hydrophobic patches in molten

Site-specific glycosylation of an aglycosylated human IgG1-Fc gly-
copeptide protein generates neoglycoproteins with enhanced function.

30. Groothuys S, Kuijpers B.H.M, Quaedflieg P.J.L.M, Roelen H.C.P.
On the preparation of carbohydrate-protein conjugates using
70:1213–1222.

31. Chen VJ, Wald F. Neoglycoproteins: preparation of noncovalent
glycoconjugates through high-affinity protein (glyco) ligand com-

32. Chen V.J, Wald F. Neoglycoproteins: preparation and properties
of biotinylated asparagine-oligosaccharides with avidin and

33. Shao MC, Wold F. The use of avidin-biotinylglycan as the model
glycoprotein to prepare non-glycosylated, acidic mutant of chicken
avidin that exhibits complexes of biotinylated asparagine-oligosac-

34. Marttila AT, Laitinen OH, Airenne K.J, Kulik T, Bayer EA,
Bayer LA. Transglycosylation of intact sialo complex-type oligosac-
charides to the N-acetylglucosamine moieties of glycopeptides
by Mucor hiemalis endo-beta-N-acetylglucosaminidase. Carbo-

35. Yan SB, Wold F. Neoglycoproteins: in vitro introduction of gly-
copeptides to proteins bearing free cysteines by squaric acid-

36. Otta T, Miura N, Fujitani N, Nakajima F, Ikukusa K, Sadamoto
R, Goto CT, Suzuki T, Suzuki Y, Monde K, Nishimura S. Gly-
costereities: synthesis of cyclic glycopeptides, toward a tailored
blocker of influenza virus hemagglutinin. Angew Chem Int Ed

S. Synthesis of neoglycoproteins using glycoconjugate-transfer

of caliciviruses derivatives containing N-linked oligosaccharides.

39. Haneda K, Inazu T, Yamamoto K, Kuchigai H, Nakahara Y, Ko-
bata A. Transglycosylation of intact sialo complex-type oligosac-
charides to the N-acetylglucosamine moieties of glycopeptides
by Mucor hiemalis endo-beta-N-acetylglucosaminidase. Carbo-

40. Li B, Song H, Hauser S, Wang L.K. A highly efficient chemenzy-
matic approach toward glycoprotein synthesis. Org. Lett. 2006;8:
3081–3084.

tide synthesis through endo-glycosidase-catalyzed glycoconjugate
transfer of sugar oxazolines: probing substrate structural require-

42. Mills KV, Petie FB. The mechanism of inter-mediating protein
755.

43. Pratt MR, Bertozzi CR. Synthetic glycopeptides and glycoproteins

44. Halt L, Ficht S, Wong CH. Strategies for the preparation of
644.

45. Buskas T, Iopke S, Boons GJ. Glycopeptides as versatile tools

46. Macmillan D, Bertozzi CR. Modular Assembly of Glycoproteins:
Towards the Synthesis of GlyCAM-1 by using expressed protein

47. Granjas C, Boutonner A, Guerero C, Fournier JM, Mu-
land LA. On the preparation of carbohydrate-protein conjugates
70:1213–1222.

48. Tada N, Kubota T, Kuroda T, Murata T, Hidai KI, Suzuki
T; Suzuki Y; Kobayashi K; Ashida H; Yamamoto K; Utsi T.
Chemoselective synthesis and application of glycopolymer con-
taining multivalent sialyloligosaccharides with a polyclutamic acid
backbone for inhibition of infection by influenza viruses.

49. Lewis WG, Green LG, Glynnza F, Radzi Z, Carrier PR, Tay-
lor P, Finn MG, Sharpless K.B. Click chemistry in situ: acetyl-
holinesterase as a reaction vessel for the selective assembly of
a fentomolar inhibitor from an array of building blocks. Ange-

3146–3152.

51. Plougou V, Villeneuve NE, Yergoul A. A method for biocona-
jugation of carbohydrates using Ditho-Alder cyclodaddition. Org.

52. Saksena R, Ma A, Kovac P. One-pot preparation of a series
of glycoconjugates with predetermined antigen-carrier ratio from
oligosaccharides that mimic the O-PS of Vibrio cholerae O:1,

53. Ma X, Sakrete A, Cheymak Y, Kovac P. Neoglycoconjugates
from synthetic tetra- and hexaamidoglycans that mimic the terminus
2003;328:2501–2603.

54. Ma X, Sakrete A, Cheymak Y, Kovac P. Neoglycoconjugates
from synthetic tetra- and hexaamidoglycans that mimic the terminus
2003;1:775–784.

55. Cheymak A, Karavorn A, Ovagya Y, Kovac P. Conjugating
glycoconjugates to proteins by squaric acid diester chemistry:


Further Reading


See Also

Glycan Synthesis, Key Strategies for
Glycan Therapeutics, Engineering of
Glycans, Chemistry of
Glycoengineering
Glycopeptides and Glycoproteins, Synthesis of
Mammalian glycosylation, the process by which a cellular complement of glycans of dazzling complexity is produced, involves a minimum of 1% of the genome (in humans) and is the dominant postsynthetic modification of both proteins and lipids endowing these molecules with an extended range of structure and function. This article provides a brief outline of the metabolic process by which monosaccharides—the raw material for the construction of complex oligosaccharides—are converted into high energy nucleotide sugar “building blocks” that are in turn assembled into four major classes of mammalian carbohydrates: glycoproteins, glycolipids, GPI anchored structures, and polysaccharides. Finally, the article concludes with a brief discussion of modern methods for manipulating glycans in living cells and in animals—using synthetic, molecular biology, and “chemical biology” approaches—as early steps toward developing sugar-based medicines.

As more information is uncovered regarding the human genome and proteome, it becomes increasingly clear that a deeper level of information resides in the great variety of signaling, receptor, and structural molecules that comprise the human body than is predicted from a strict examination of the genetic code. The answer to this apparent disparity lies in the posttranslational modification of proteins that endows the unexpectedly modest number of genes with great product diversity. Although dozens of posttranslational modifications occur, glycosylation—the addition of one or more sugar residues to a protein or lipid to convey additional information, structure, or function—is arguably the most common (1). In recent years, major strides in the development of sensitive and reliable methods of detection and functional analysis for complex carbohydrates have revealed just how significant these posttranslational glycosylation modifications are; yet, many of these structures remain elusive as large-scale sugar-specific technologies have lagged the rapid rate of discovery for genes and proteins. Now, information gleaned from glycomics is accelerating the study of glycosylation (2, 3) and has generated claims that 220–250 genes, roughly 1% of the human genome, are involved in glycan production and modification (4), and that a minimum of 50% of proteins are glycosylated (5). This article outlines the biosynthetic process of mammalian glycans by first examining the basic monosaccharide building blocks and the ways they combine to form oligosaccharide and polysaccharide structures (Fig. 1). An inspection of the various glycoprotein, glycolipid, GPI anchor, and independently functional polysaccharide linkages is then provided along with a brief description of the structure and function of the various classes of enzymes in the respective biochemical pathways. The article concludes with examples of current research efforts in synthetic chemistry, biologic, and “chemical biology” strategies that seek to exploit the flurry of recent advances in understanding mammalian glycans to develop novel sugar-based therapies for human disease.

Monosaccharides—The Building Blocks for Glycosylation

Monosaccharides are obtained from the diet and transported into cells

In mammals, the primary source for the monosaccharides used for glycan biosynthesis is the diet, but many cell types also scavenge sugars released into the bloodstream by other tissues and organs. A typical mammalian diet containing polysaccharides and starches will result in a rich supply of simple sugars after digestion in the gastrointestinal tract (the full names, abbreviations, and chemical structures of mammalian monosaccharides are given in Fig. 2). These sugars, as well as several less abundant monosaccharides such as galactose (Gal), mannose (Man), or glucosamine (GlcN), are absorbed into the bloodstream and taken up by cells throughout the body via transporters located in the plasma membrane.
Glycan Biosynthesis in Mammals

Figure 1 - Overview of mammalian glycan biosynthesis. Monosaccharides, most commonly glucose (shown), are imported into a cell by membrane transporters of the GLUT or SGLT families. Imported sugars can undergo extensive processing in the cytosol, as exemplified by ManNAc conversion to sialic acid (see Fig. 7), or be directly phosphorylated and converted to high energy nucleotide sugars (an NDP-monosaccharide is shown, where N is a nucleoside). Nucleotide sugars may be used in the cytosol to synthesize lipid-linked precursors of N-linked glycans (see Fig. 3), or they may be transported into the lumen of the ER or Golgi by nucleotide sugar transporters (NST) and therein used as donors for oligosaccharide biosynthesis catalyzed by glycosyltransferases (GTs). The NDP released by GT action is converted to NMP by pyrophosphatase (PPase). Upon further processing and transport, glycans are displayed on the cell surface (shown) or exported from the cell (not shown) where they have diverse functions ranging from signaling and metabolic regulation (e.g., glycoprotein hormones) to a structural role (e.g., polysaccharides that are major ECM constituents).

Characterization of the function and structure of monosaccharide transporters found in the plasma membrane led to the identification of the SGLT (sodium-dependent co-transporters from the gene SLC5A) and GLUT (sodium-independent facilitative transporters from the gene SLC2A) families (6, 7). Various members of these two transporter families are localized to different tissue types. For example, GLUT1 is found in erythrocytes; GLUT4, GLUT5, and GLUT12 predominate in skeletal muscle tissue although other transporters are also expressed at lower numbers in this metabolically voracious tissue (8); and GLUT14 is specifically expressed as the predominant transporter in two alternative splice forms of the human (but not mouse) testis (9). Another feature of these proteins is overlapping substrate specificity with many family members capable of transporting multiple monosaccharides, albeit with differing efficiencies (10).

De novo synthesis of high-energy nucleotide sugars

After monosaccharides are delivered into cells, they are subjected to a series of chemical conversions, including epimerization, acetylation, condensation, and phosphorylation reactions, to produce the spectrum of building blocks required for glycan biosynthesis. The intracellular metabolic network is capable of the de novo synthesis of sufficient amounts of glucosamine (GlcN), fructose (Fru), mannose (Man), fucose (Fuc), N-acetyleneuraminic acid (sialic acid, Neu5Ac or Sia), galactose (Gal), N-acetylgalactosamine (GalNAc), N-acetylgalactosamine (GalNAc), xylose (Xyl), and glucuronic acid (GlcA), all of which exist in the α-conformation except for α-fucose. Mammals other than humans also produce the N-glycolylneuraminic acid (Neu5Gc) form of sialic acid (11).

In the cytosol, these monosaccharides can be phosphorylated and subsequently coupled with nucleotides such as uridine diphosphate (e.g., UDP-GlcNAc), guanosine diphosphate (e.g., GDP-mannose), or cytosine monophosphate in the case of sialic acids (Fig. 2b) to create a set of high energy "building blocks" for glycan assembly. In some cases, such as for the initial steps...
Figure 2  Mammalian monosaccharides and nucleotide sugars. (A) Chemical structures, common names, and abbreviations of the 11 monosaccharides found in mammalian glycans as well as ManNAc, the precursor for sialic acids. (B) Examples of the three classes of nucleotide sugars are provided by UDP-GlcNAc (Glc, GlcA, Gal, GalNAc, and Xyl also use UDP), GDP-Man (Fuc also is linked to GDP), and CMP-Neu5Ac/Gc (IdoA is produced by postsynthetic epimerization of GlcA and therefore does not require a nucleotide sugar).
in the synthesis of the dolichol-linked 14-mer used in N-linked glycan biosynthesis (discussed below) or for O-GlcNAc protein modification (12), nucleotide sugars are used in the cytosol; more often they are transported into the lumens of the Golgi apparatus and endoplasmic reticulum (ER) where the bulk of oligosaccharide assembly and processing occurs. In either case, the release of the monosaccharide from its bonded nucleotide phosphate provides the energy currency for the formation of glycosidic bonds found in glycolipids and glycoproteins.

Transport of nucleotide sugars into ER/golgi

The transport of high energy nucleotide sugars from the cytosol into the ER and Golgi lumens occurs by highly specific
membrane proteins of the SLC35 nucleotide sugar transporter family. This class of proteins has at least 17 members, some of which can accept multiple substrates; at the same time, certain nucleotide sugars can be accepted by multiple transporters. These transporters are organellar specific; typically nucleotide sugars are only transported into an organellar compartment endowed with the corresponding glycosyltransferases (13). For example, CMP-Sia, GDP-Fuc, and UDP-Gal are transported solely into the Golgi; UDP-GalNAc, UDP-GlcNAc, UDP-Glc, and UDP-Xyl are transported twice as rapidly into vesicles of Golgi as ER; conversely, UDP-Glc is transported into ER vesicles much more rapidly than into the Golgi (10).

Regulation of the assembly of complex carbohydrates from nucleotide sugars, which is an extremely complex and still poorly understood process, is accomplished by several means, including compartmentalization of glycosyltransferases, the activities of these enzymes, nucleotide sugar-transport rates, and the available concentration of substrates (14). By influencing the latter two parameters, nucleotide sugar transporters play a major role in determining the outcome of glycan structure and significant efforts have been expended to understand the transport mechanism. Briefly, these proteins are anti-porters that exchange the nucleotide sugar for a corresponding nucleotide monophosphate in an equimolar fashion (15). For example, a nucleotide monophosphate (NMP) may be exported from the ER in exchange for the importation of a nucleotide sugar (Fig. 1). It is worth noting that NMP is not a product of the glycosylation process within these organelles; instead NMP is produced by enzymatic dephosphorylation of NDP generated during the glycosyltransferase-catalyzed attachment of a monosaccharide residue to a growing oligosaccharide chain. Consequently, both the specific transporter and the corresponding nucleotide diphosphatase are required within the lumen of a specific organelle for successful transport, and evidence suggests that the transport of these nucleotide sugars into the ER or Golgi apparatus regulates which macromolecules will undergo glycosylation (16). Transport is competitively inhibited by corresponding nucleoside monophosphate or diphosphate in the cytosol, but not by the free sugars, and it does not require an energy source such as ATP. Finally, after glycan assembly, the many posttranslational modifications that glycans undergo (phosphorylation, acetylation, and sulfation) also require active transport of necessary materials into the ER and Golgi; for example, PAPS (3′-phosphoadenosine 5′-phosphosulfate) required for sulfation is imported by the PAPST1 gene product (17).

Glycoconjugate Assembly

Once the required high energy nucleotide sugars and other building blocks have been translocated to the appropriate cellular compartments, the glycosylation of newly synthesized proteins and lipids can begin. In this section we examine the production of the major classes of prevalent mammalian glycan structures; a more thorough discussion, including low abundance glycans not discussed here, can be found in review articles (see the Further Reading list).

Glycoproteins

In proteins there are two major glycan classes, N-linked (Fig. 4) and O-linked (Fig. 5), based on the atom (nitrogen or oxygen) of the specific amino acid residue to which the glycan is tethered. Glycans are also distinguished from one another by significant differences in biosynthesis, structure, and function.

N-linked glycosylation

N-Linked glycosylation is one of the most prevalent protein modifications and serves many invaluable functions, including stabilization of structure, enhanced solubility, immunomodulation, mediation of pathogen interaction, serum clearance rate, protein half-life, and proper folding (18). Dysfunctional N-glycosylation can result in serious detriment to the cells and organism as a whole as exemplified by congenital disorders of glycosylation (CDGs) (19) and adult diseases such as cancer (20).

The term “N-linked” refers to the chemical linkage of the glycan moiety to the nitrogen of the amido group of an asparagine (Asn) residue in the host protein. N-linked glycosylation is a multicompartmental affair, which involves the cytosol and both the ER and the Golgi complexes (Fig. 3). Biosynthesis of glycoproteins begins on the cytosolic face of the ER by the formation of a 95-105 carbon polysaccharide lipid, dolichol phosphate (Dol-P), which acts as a carrier for the nascent glycan structure. Assembly of the core glycan on this carrier begins by addition of a GlcNAc-P (from UDP-GlcNAc) onto Dol-P, forming GlcNAc-P-glycosylphosphoryl dolichol (GlcNAc-P-Dol) through the action of the GlcNAc-1-phosphotransferase DPAGT1. This reaction exemplifies the use of monosaccharide transferases that occur throughout the assembly of the core structure as well as in subsequent elaboration processes. A second GlcNAc and five Man residues are added (from UDP-GlcNAc and GDP-Man, respectively) in sequence to form Man9GlcNAc2-P-Dol, which is then flipped to the luminal side of the ER (A in Fig. 3) (21). Additional Man and Glc residues are added in the lumen of the ER via donors Dol-P-Man and Dol-P-Glc, which results in the primary core structure Glc3Man9GlcNAc2-P-Dol (B in Fig. 3). The terminal α-1,2-linked Glc residue is required for recognition by the oligosaccharide transferase (OST) that attaches this core glycan structure on block to the host protein (C in Fig. 3).

N-Linked glycosylation is considered to be a cotranslational rather than a posttranslational modification because OST searches unfolded polyopeptides emerging from the ER during translation for a universal Asn-X-Ser/Thr consensus sequence, where X is any residue except proline. Proline is disallowed because its rigidity prevents the consensus sequence from forming a loop structure wherein the hydroxyl group of Ser/Thr interacts with the amido group of Asn, thereby making it more nucleophilic and enhancing the installation of the glycan moiety (23). OST binds to Glc3Man9GlcNAc2-P-Dol and catalytically cleaves the phosphoglycosidic bond in the GlcNAc-P moiety, thereby releasing Dol-P during the transfer of Glc3Man9GlcNAc2 to the targeted Asn residue (B in Fig. 3) (24).

Once the transfer of the core Glc3Man9GlcNAc2 14-mer to protein is complete, the terminal Glc residues are removed in...
Figure 4: Overview of mucin-type O-glycan biosynthesis. The production of the eight core structures found in O-glycans is shown along with the enzymes and nucleotide sugars that are involved in each step. Additional information on the enzymes shown can be found in the KEGG databases (see the legend for Fig. 3).
sequence by glucosidase I (the terminal α1-2-linked Glc) and glucosidase II (the α1-3-linked Glc residue). Glcα1-ManGlcNAc2 targets the nascent glycoprotein for entry into the calnexin/calicretin cycle within the ER, which is a major component of the quality control system that assists glycoproteins to fold properly and to achieve their ideal conformation (C in Fig. 3). After folding is completed, the final Glc residue is removed by glucosidase II, and a terminal α1-2-linked Man residue is removed from either of the two other arms of the oligosaccharide by ER mannosidase I or II. The remaining Manα1-GlcNAc2 oligosaccharide structure is transported along with the newly formed protein to the cis-Golgi for further modification (D in Fig. 3). Golgi mannosidases IA and IB subsequently remove three additional α1-2-linked Man residues to form an intermediate Manα1-ManGlcNAc2 glycan structure that is subsequently built into the high-Man, complex, and hybrid subclasses of N-linked glycans. The formation of complex N-linked glycans begins in the medial-Golgi with the addition of a GlcNAc residue by mannosyl-α1-3-glycoprotein-α1-2-N-acetylglucosaminyltransferase (MAGT1) onto the α1-3-linked Man residue of Manα1-GlcNAc2 (26). Manosidase II removes the two remaining terminal Man residues from Manα1-GlcNAc2 and mannosyl-α1-3-glycoprotein-α1-2-N-acetylglucosaminyltransferase (MAGT2) adds a GlcNAc residue to the final remaining terminal Man residue (27). Additional modifications, such as the addition of a Fuc to the proximal GlcNAc, addition of a β1-6-GlcNAc to the α1-6-linked Man residue that already bears β1-2-GlcNAc, or capping with terminal sialic acids, generate a wide variety of diverse structures within the complex N-linked glycan class (10). Biosynthesis of hybrid N-linked glycans begins with the addition of β1-2-GlcNAc to the α1-2-linked Man residue of the intermediate Manα1-GlcNAc2 glycan structure; the removal of the two remaining Man residues by mannosidase II, as occurs in complex N-linked glycan biosynthesis, is prevented by the addition of a β1-4-GlcNAc to the proximal β1-4-Man residue (28). This mannosidase II-protected structure is translocated to the trans-Golgi where additional modifications to the oligosaccharide structure occur, which once again generates a vast array of structures.

O-linked glycosylation

O-Linked glycosylation is a posttranslational modification where the glycan moiety is attached to the hydroxyl group of a serine or threonine amino acid residue of a protein, often in dense clusters of carbohydrate that force the peptide chain into a highly extended, poorly folded conformation. There are several potential O-glycans, including O-linked Fuc and O-glycan linkages to hydroxylsine (in the collagen sequence Gly-X-Hyp-Gly) and hydroxyproline (in plants, but by far the most common form of O-glycosylation is the addition of α1-N-acetylgalactosamine to form O-linked Galα1-3Ser/Thr (known as mucin-type O-glycosylation) to which many subsequent oligosaccharides can be added for varying functionality. Aberrations in O-linked glycosylation are often found in various disease states, including blood disorders, cancer, and diabetes.

Unlike N-glycosylation that always begins with an N-linked Glcα1-ManGlcNAc2 14-mer core structure, mucin-type O-glycosylation begins in the Golgi apparatus with the addition of a single monosaccharide, typically GalNAc, onto a Ser or Thr residue of a protein that has already completed the folding process (Fig. 4). The production of the initial O-linked Galα1-3Ser/Thr structure, known as the Tn-antigen, is facilitated by an enzyme, O-GalNAc transferase, which forms a complex with the protein. Often, this simple glycan moiety is translocated to the trans-Golgi for elongation through the stepwise addition of Gal, GalNAc, or GlcNAc residues that form the basis of eight core structures. These structures can be further modified by sialylation, sulfation, acetylation, fucosylation, or polylactosamine extension (10). Although there is not a specifically defined consensus sequence for mucin-type O-linkages (29), statistical analysis has yielded a rule set to predict sites of O-GalNAc modification. For instance, there are inherent differences in site specificity between tissue types based on different GalNAc transferase expression patterns between cells. Moreover, because O-glycosylation occurs on fully folded proteins, only surface-exposed Ser and Thr residues will be accessible for O-glycosylation. The density patterns of O-linked glycans also suggest that nearby residues can influence transferase activity (29).

A iso of interest in a discussion of O-linked glycans is the addition of a single GlcNAc to Ser or Thr to form a class of cytosolic and nuclear glycosylated proteins (30). O-GalNAc is very common in nuclear and cytosolic proteins, including nuclear pore proteins, transcription factors, and cytoskeletal elements (31). O-GalNAc modification is likened more to phosphorylation than to the other forms of glycosylation because of its transient nature and frequent yin-yang status with phosphorylation at the same sites, particularly during different cell-cycle stages and in development (32).

Glycolipids

A glycolipid is any compound containing one or more monosaccharide residues bound by a glycosidic linkage to a hydrophobic moiety such as an acylglycerol, a sphingoid, or a prenol phosphatide. In mammalian, most glycolipids are glycosphingolipids (GSLs), which is a large and widely varying family of amphipathic lipids based on the ceramide N-acylsphingoid lipid moiety (Fig. 5). GSLs reside in cellular membranes, typically in the plasma membrane where the glycan is almost always oriented outward, exposed to the extracellular space. These molecules play a role in the protective glycocalyx covering of a cell and participate in raft assemblies such as the “glycosynapse” (33). GSLs participate in cell–cell recognition, cell–matrix interactions, and cell surface reception and messaging. GSLs are required for proper development. Biopsythetetic or catabolic defects result in pathologies ranging from liver disease to insulin-resistant diabetes, multiple sclerosis, Tay–Sachs, and Gaucher’s disease.

Glycolipid synthesis begins on the cytosolic face of the ER (34) with the condensation of a serine residue and a palmityl-CoA to form 3-dehydrophosphoinositol, which is hydroxylated at the 4-oxygen, acylated, and unsaturated between C4 and C5 in a trans-fashion to form ceramide (Cer)
Figure 5  Overview of glycolipid biosynthesis. Mammalian glycosphingolipids (GSLs) are synthesized from Cer after the addition of Gal, to form the small GalCer-series, or after the addition of Glc and Gal to form the ubiquitous LacCer class, which is subdivided into gangliosides, globosides, and the neo-lacto-series. Additional information on the enzymes and specific GSLs shown can be found in the KEGG databases (see the Legend for Fig. 3).
Glycan biosynthesis in Mammals

Glycolipids; most commonly, Cer is conjugated with a Gal or one of several modifications that lead to different classes of monosaccharides and phosphoethanolamine linked to the C-terminus of a protein.

Ceramide then crosses the ER membrane and undergoes modifications, whereas the GlcCer core can experience extensive differences; a prime example is 9-O-acetylation of the terminal residue of GD3 that endows this potent inducer of apoptosis with anti-apoptotic properties (38). The N-terminal domains of the promiscuous glycosyltransferases responsible for the construction of the gangliosides specify the distribution of these enzymes within the Golgi stacks, which results in a differential expression pattern (37). A salvage pathway also exists for resynthesizing gangliosides, recycling them from their endosomal breakdown through the Golgi; this recycling pathway dominates in slowly dividing cells, whereas de novo synthesis dominates in highly mitotic cells.

GPI anchors

The discovery that phospholipase C could release alkaline phosphatase from lipid-linked structures on cellular surfaces (39) led to the identification of the glycophasphatidylinositol (GPI) membrane component (40). GPI structures are a synthetic tour de force of nature, combining lipids, carbohydrates, and proteins into a single macromolecule. Certain proteins require GPI anchoring to be functional; for example, L-64E-mediated T-cell activation is critically dependent on its GPI anchor (41), and folate uptake functions efficiently only when its receptors are GPI anchored (42).

The basic structure of the GPI-anchor (Fig. 6) maintained across all species studied thus far) begins with phosphatidylinositol (PS, consisting of diphosphateinositol, which is attached to an inositol via a phosphodiester. An oligosaccharide chain, which is attached to the inositol, consists of GlcN (donated by Dol-P-Man donors), three linear Man residues (provided by Dol-P-Man donors). Finally, phosphoethanolamine (P-EtN) is linked to the terminal Man residue (43).

The GPI anchor is synthesized in the ER lumen and trafficked into the Golgi where it can be modified further by sulfation or additional glycosylation before cell-surface presentation. Unlike GalCer-derived GSLs, structures based on GlcCer are ubiquitous. This GSL forms at the cis-Golgi and is translocated to the lumen of the Golgi via the Golgi stack trafficking process (37) to become a substrate for various glycosyltransferase enzymes and complexes. The addition of a Gal residue to GlcCer results in LacCer, which is the foundation for three additional classes of GSLs. First, the (neo-)lacto-series, or blood group series, begins with the addition of a β1-3-GlcNAc. Next, the Globo series is distinguished by the addition of an α1-4-Gal (also known as the Pk antigen). Third are the gangliosides, which are glycolipids that feature one or more sialic acid residues; membership in this group does not preclude inclusion in the lacto- or globo series. Gangliosides are present in nearly all animal cells, but they are particularly prevalent in the plasma membranes of cells in the central nervous system (37).

Because of the importance of gangliosides in a variety of disease states and the volume of research devoted to them, the biosynthetic processes of this class of GSLs are described briefly. All gangliosides begin as LacCer except for GM4 (α2,3-GalCer). From LacCer, non-GM4 ganglioside biosynthesis continues down one of two branches: the α1-2-Gal (also called the α-pathway) through addition of a GalNAc residue, or into the “ganglioside proper” pathways (a-, b-, and c-pathways) through the addition of one or more sialic acid residues. GalNAc and Gal residues can be added as side chains that may be capped with additional sialic acids. Sialic acids appended to gangliosides can have subtle differences in structure that can result in drastic functional differences; a prime example is 9-O-acetylation of the terminal residue of GD3 that endows this potent inducer of apoptosis with anti-apoptotic properties (38). The N-terminal domains of the promiscuous glycosyltransferases responsible for the construction of the gangliosides specify the distribution of these enzymes within the Golgi stacks, which results in a differential expression pattern (37). A salvage pathway also exists for resynthesizing gangliosides, recycling them from their endosomal breakdown through the Golgi; this recycling pathway dominates in slowly dividing cells, whereas de novo synthesis dominates in highly mitotic cells.

GPI anchors

The discovery that phospholipase C could release alkaline phosphatase from lipid-linked structures on cellular surfaces (39) led to the identification of the glycophasphatidylinositol (GPI) membrane component (40). GPI structures are a synthetic tour de force of nature, combining lipids, carbohydrates, and proteins into a single macromolecule. Certain proteins require GPI anchoring to be functional; for example, L-64E-mediated T-cell activation is critically dependent on its GPI anchor (41), and folate uptake functions efficiently only when its receptors are GPI anchored (42).

The basic structure of the GPI-anchor (Fig. 6) is maintained across all species studied thus far) begins with phosphatidylinositol (PI), which spans the external ER membrane linked to an inositol via a phosphodiester. An oligosaccharide chain, which is attached to the inositol, consists of GlcN (donated by Dol-P-Man donors), three linear Man residues (provided by Dol-P-Man donors). Finally, phosphoethanolamine (P-EtN) is linked to the terminal Man residue (43).

The GPI anchor is synthesized in the ER lumen and trafficked into the Golgi where it can be modified further by sulfation or additional glycosylation before cell-surface presentation. Unlike GalCer-derived GSLs, structures based on GlcCer are ubiquitous. This GSL forms at the cis-Golgi and is translocated to the lumen of the Golgi via the Golgi stack trafficking process (37) to become a substrate for various glycosyltransferase enzymes and complexes. The addition of a Gal residue to GlcCer results in LacCer, which is the foundation for three additional classes of GSLs. First, the (neo-)lacto-series, or blood group series, begins with the addition of a β1-3-GlcNAc. Next, the Globo series is distinguished by the addition of an α1-4-Gal (also known as the Pk antigen). Third are the gangliosides, which are glycolipids that feature one or more sialic acid residues; membership in this group does not preclude inclusion in the lacto- or globo series. Gangliosides are present in nearly all animal cells, but they are particularly prevalent in the plasma membranes of cells in the central nervous system (37).

Because of the importance of gangliosides in a variety of disease states and the volume of research devoted to them, the biosynthetic processes of this class of GSLs are described briefly. All gangliosides begin as LacCer except for GM4 (α2,3-GalCer). From LacCer, non-GM4 ganglioside biosynthesis continues down one of two branches: the α1-2-Gal (also called the α-pathway) through addition of a GalNAc residue, or into the “ganglioside proper” pathways (a-, b-, and c-pathways) through the addition of one or more sialic acid residues. GalNAc and Gal residues can be added as side chains that may be capped with additional sialic acids. Sialic acids appended to gangliosides can have subtle differences in structure that can result in drastic functional differences; a prime example is 9-O-acetylation of the terminal residue of GD3 that endows this potent inducer of apoptosis with anti-apoptotic properties (38). The N-terminal domains of the promiscuous glycosyltransferases responsible for the construction of the gangliosides specify the distribution of these enzymes within the Golgi stacks, which results in a differential expression pattern (37). A salvage pathway also exists for resynthesizing gangliosides, recycling them from their endosomal breakdown through the Golgi; this recycling pathway dominates in slowly dividing cells, whereas de novo synthesis dominates in highly mitotic cells.
Glycan Biosynthesis in Mammals

Figure 7 Metabolic oligosaccharide engineering of sialic acid. Exogenously supplied analogs of ManNAc or sialic acid can intercept the sialic acid biosynthetic pathway (at different points) and be displayed on the cell surface in place of the natural sialic acids Neu5Ac and Neu5Gc. The sampling of non-natural "R" groups shown here was selected from References 84, 92, 93 and 97–101 (note that not all "mix-and-match" permutations of the indicated groups have been reported). Additional information on the enzymes shown can be found in the KEGG databases (see the Legend for Fig. 3).
Hyaluronan
Hyaluronan, or hyaluronic acid, is synthesized at the plasma membrane (rather than in the ER or Golgi apparatus) by one of three distinct hyaluronan synthases, which allows it to be easily secreted directly to the ECM (47). Hyaluronan is the simplest GAG, consisting of the repeating unit \(-\text{GlcA-\beta-1,3-GlcNAc-\beta-1,4-}\); this GAG forgoes postsynthetic modification and remains unbound to surface proteins.

Heparin/heparan sulfate and chondroitin/dermatan sulfate
Heparin/heparan sulfate GAGs (HSGAGs) and chondroitin/dermatan sulfate GAGs (CSGAGs) share a common synthetic origin, with both being linked to a core protein through a specific uronic acid. HSGAGs differ from other GAGs in two major respects. Keratan sulfate differs from other uronic acids in its disaccharide repeat. The first GlcNAc or GlcNAc residue determines whether the GAG will belong to the heparan sulfate or condroitin sulfate family, respectively. HSGAGs consist of the repeating unit \(-\text{GlcNAc-\beta-1,4-Glc} \alpha-N\text{-Acetylglucosamine transferase I and II) and a GlcNAc (by glucosamine acid transferase I) are attached sequentially to complete the tetramer.} \]

Synthesis of the polysaccharide portion of these GAGs begins with the addition of GalNAc (or GlcNAc) and GlcA residues to the O-linked tetramer in an alternating fashion by multidomain glycosyltransferases (48). The addition of the first GlcNAc or GalNAc residue determines whether the GAG will belong to the heparan sulfate or condroitin sulfate family, respectively. HSGAGs consist of the repeating unit \(-\text{GlcNAc-\beta-1,4-GlcA-\alpha-1-3-}, \)-which contains GlcA rather than GlcNAc and employs 1-3- rather than 1-4-glycosidic linkages between the repeating disaccharides; despite the differences in the monosaccharide building blocks used, CSGAGs are also constructed from genes in the EXT family (48). When the HSGAG or CSGAG chain has grown to an appropriate length, it is acted upon by additional enzymes that impart structural uniqueness. 2-0, 3-0, and 6-O-sulfotransferases add sulfate groups at appropriate locations (50). N-deacetylase N-sulfotransferase can exchange the amine groups of GlcNAc, and C5 epoxytransferase converts a portion of GaNAc residues to IdoA. This epimerization results in the distinction between condroitin sulfate and dermatan sulfate (51).

Keratan Sulfate
Keratan sulfate differs from other GAGs in two major respects. First, it can be either N- or O-linked to the core protein (52). Second, its repeating disaccharide unit contains a Gal rather than one of the uronic acids in its disaccharide repeat. The basic repeating unit is \(-\text{Gal-\beta-1,3-GalNAc-\alpha-1,4-}\), assembled by \(\beta\)-1,4-galactosyltransferase (B4GALT1) and a \(\beta\)-1,3-GalNAc transferase (B3GALT1 or B3GALT2). Three classes of keratan sulfate are distinct in their protein linkage. KSI members are N-linked to an Asn of the protein; they are found primarily in the cornea and can be terminated with sialic acids, Gal, or GlcNAc. KSI members are O-linked to a Ser/Thr residue of the core protein; they are primarily found in cartilage, are highly sulfated, and are terminated by sialic acids. KSIIs are found in brain tissue and have a unique linker between the keratan sulfate chain and the protein: a Man O-linked to a Ser of the protein.

Manipulating Mammalian Glycans—Early Steps Toward Sugar-Based Medicines
Although many aspects of glycans remain mysterious, mammalian glycosylation has now been elucidated well enough to provide a basic understanding of glycan biosynthesis, structure, and function that, combined with the realization that many disease states result from glycan abnormalities, have spurred efforts to develop sugar-based medicines. As modern medical research looks for ways to exploit the knowledge that has been gained regarding glycan biosynthesis to create new therapeutics, several challenges inherent in modifying glycosylation in living cells and tissues must be overcome (53). A particular obstacle to biologic intervention is the lack of template for carbohydrate structures akin to the DNA sequence that specifies primary amino acid sequences of proteins that motivates gene therapies. Similarly, the expense of de novo synthesis of complex oligosaccharides coupled with their notoriously poor pharmacologic properties make them nonideal drug candidates and hinders conventional synthetic approaches to drug development. Nonetheless, these limitations, both biology- and chemistry-based approaches remain enticing; moreover, merging these two disciplines into innovative “chemical biology” strategies seems particularly promising. This article concludes by discussing each of these topics briefly, after providing a short synopsis of the contributions of rapidly coalescing glycomics efforts to the study of mammalian glycosylation and development of human therapeutics.

Advances in technology and bioinformatics ease characterization
Efforts to manipulate glycans for medical purposes rely on knowing what oligosaccharide structures exist in health and disease, the biosynthetic machinery that builds these sugars, and specific molecular changes rendered by therapeutic intervention. In the past, the difficulty of characterizing glycans has made tackling even one of these tasks formidable. Now, however, bioinformatics has combined technological advances in carbohydrate analysis with the concept of global characterization found in genomics and proteomics to create the field of “glycomics”—the study of all the glycan structures produced by the cell. Modern methods using mass spectrometry, chromatography, nuclear magnetic resonance, and capillary electrophoresis have identified many carbohydrate structures (3, 54, 55). These techniques, along with high throughput arrays using lectins and other glycan-binding proteins (56), have resulted in a significant amount of information available in databases such as the Online glycan database (http://www.glycanbank.org).
as those available online from the German Cancer Research Center (http://www.glycosciences.de), the Consortium for Functional Glycomics (http://www.functionalglycomics.org), and the AFB-CNR of the University of Provence and the University of the Mediterranea (http://www.cazy.org). These resources complement automated methods for predicting function, structure, and localization of newly discovered glycans and glycan-related enzymes alongside their arrays of references, composition and spatial structures, and gathered NMR shift data (4). These developments are critical both for the glycobiology specialist undertaking further study of the intricacies of glycosylation and well as for the non-specialist, such as a physician, who seeks to apply glyobiology-based technologies in the clinic.

Biological approaches

Genetic approaches

Fascinating findings that viruses alter the glycome by regulating expression of host glycosyltransferases or by expressing their own glycosyltransferases (37) and forward genetics approaches that have identified glycan defects associated with specific genetic abnormalities (58) have motivated efforts to exploit modern genetic tools to manipulate glycosylation. Unfortunately, tools such as knockout mice to eliminate a specific biosynthetic enzyme often have manifold and severe effects such as early lethality (59). By contrast, in other situations, including the “goat” knockout pig created to supply organs for xenotransplantation (60), the removal of a glycosyltransferase did not abolish production of the targeted oligosaccharide epitope. Notwithstanding these difficulties, the compelling links between defects in specific glycosylation enzymes and disease currently refractory to treatment, such as cancer (20), have ensured the continuation of research efforts to develop genetic methods to manipulate glycans.

Molecular biology approaches

In addition to upregulation or knock-down of biosynthetic elements, modern molecular biology offers techniques for more subtle manipulation of the glycosylation process. The biosynthesis of glycan structures relies on precisely localized enzymes for proper construction, for example, the localization of glycosylation enzymes within the ER and Golgi cisternae in the same sequence in which they act to modify oligosaccharide substrates (61). One way that this localization is achieved is based on the thickness of the membranes, which increases from the ER to the cis-, medial-, and trans-Golgi compartments; glycosyltransferase enzymes possess transmembrane domains of a length optimal to anchor them to a specific location in a cell’s secretory organelles (62). It is therefore possible to relocate an enzyme involved in glycosylation by swapping that enzyme’s native transmembrane region with a transmembrane domain of a different length and thereby changing substrate preference (63). Additionally, the stem region, located between the transmembrane and catalytic domains, can also be swapped to tune the activity of a glycosyltransferase (64). Although currently largely laboratory research tools, these studies point the way to a future where fine control over glycosylation may be possible by mix-and-matching the membrane, stem, and catalytic domains of glycan processing enzymes.

Synthetic approaches

Modern synthetic chemistry has been able to reproduce several glycan structures of considerable complexity and biomedical relevance. The pioneering example is the use of synthetic sialyl Lewis X for the treatment of reperfusion injury (65). For the past decade, much effort, which has been facilitated by automated synthesis (66), has focused on the creation of carbohydrate-based vaccines. It is possible to use synthetic carbohydrate analogs of viral and microbial surface polysaccharides as vaccines to elicit an immune response against the microorganism. In fact, because an “artificial” polysaccharide can be carefully designed through precise synthesis, this type of vaccine may be both safer and more effective at lower dosage [i.e., through multivalency (67, 68)] than a naturally derived vaccine such as that of a live or killed microbe that contains a mixture of glycoforms, some of which may be immunogenic (69). Synthetic polysaccharide vaccines have been recently developed for several targets, including Haemophilus influenzae type b (70), human immunodeficiency virus (71), and various cancers (72, 73). In the future, as synthetic strategies are streamlined to become both technically and cost-effective, the possibilities of using carbohydrates to positively impact human health are numerous; for example, human breast milk contains a multitude of oligosaccharides that are distinct from other species such as the cow. Human-specific milk sugars are both developmentally important and have activity against pathogens (74), and the ability to supplement infant formula with these sugars would be valuable especially in the third world nations where malnutrition is endemic and infectious diseases are prevalent.

Chemical biology approaches

Combined chemical–biological synthesis

Combining tried-and-true synthetic methods with emerging chemoselective ligand methodology (75) and chemoenzymatic transformations that use the suite of enzymes cells employ for glycosylation (76) has led to the production of a multitude of carbohydrate structures. These hybrid approaches, exploiting biologic tools for programmable one-pot strategies (77), have several attractive features, including the ability to not only make a carbohydrate moiety but rather an entire glycoprotein (78) or glycolipid (79). Another important aspect of a synthetic strategy is that chemically distinct glycan structures can be produced allowing evaluation of the biological response of an individual glycoform, rather than an averaged response obtained when testing a mixture of the profusion of glycoforms found in nature. To illustrate, prion proteins from diseased and healthy cells have different glycan profiles that are proposed to be critical for disease progression (80), but this hypothesis is difficult to verify without the synthetic ability, which is afforded by the methods described herein, of producing testable quantities of individual prion glycoforms.
Small-molecule switches for controlling glycosylation

The emerging use of small molecules to control the activity of glycans processing enzymes (81) and direct biosynthetic traffic in the Golgi (82) constitutes an intriguing approach for modulating glycan synthesis. These efforts build on the modular nature of glycosyltransferases, where the membrane, stem, and catalytic domains can be swapped without loss of function. In particular, fusion proteins were created that combined the catalytic or localization domain of fusosyltransferase 1 (FUT1) with the rapamycin-binding proteins FRP or FRB. Then, by exploiting rapamycin-mediated heterodimerization of these elements to control Golgi localization of both domains and thereby reconstitute FUT1 activity, Koliher and Bertozzi showed that cell-surface glycosylation could be altered (83). In the future, bringing a rich complement of biologic and chemical tools into play, as exemplified by this approach, will continue to drive progress toward glycanc-based therapies.

Metabolic oligosaccharide engineering

“Metabolic oligosaccharide engineering” is an attractive alternative to the synthetic strategies outlined above not only because of its basic simplicity but also because it provides one of the few ways to alter glycosylation in living systems. This methodology, pioneered by the Reutter laboratory for sialic acid (84) (Fig. 7) and now extended to GaINac (85, 86) and GlcNAc (87), is based on the remarkable ability of certain non-natural monosaccharide analogs to be metabolically incorporated into glycosylation pathways and to replace the corresponding sugar residue in the oligosaccharide complement of a cell.

The opportunity provided by metabolic oligosaccharide engineering is divided into two essential classes. First, the intrinsic “on/off” switch that nature has to fine-tune function and structure (88), and thereby endow glycans with antiviral properties (89), enhance immunogenicity (90), modulate cell adhesion (91), or confer stem cell fate (92). Alternatively, when the analog bears a chemical functional group unique to the cell surface, such as a ketone (93) or an azide (94), such sugars can act as “tags” for the delivery of genes (95), toxins (96), or imaging agents (97) by exploiting chemoselective ligation chemistry that has been developed to be compatible with physiologic conditions (75).

Acknowledgments

The authors thank S.-G. Sampathkumar for critical comments on the manuscript and the National Institutes of Health (R01CA121134-01A1), the National Science Foundation (OSB-0425688), and the Arnold and Mabel Beckman Foundation for financial support.

References

Glycan Biosynthesis in Mammals


Further Reading


See Also

Glycans in Information Storage and Transfer
Glycans, Mammalian
Glycosylation of Proteins in the Golgi Compartment
Glycomics, Major Techniques
Glycan Therapeutics, Engineering of
Eukaryotic cell glycosphingolipids (GSLs) are central components of membrane lipid microdomains that function as trans plasmamembrane signaling foci. GSLs serve as important receptors and coreceptors, primarily to mediate host/microbial pathogen interactions, and undergo unique intracellular trafficking pathways. As such, their chemical modification can generate interventive therapeutic strategies and biologic probes. The process of achieving such goals via chemistry of the carbohydrate is in its infancy, but substitution within the lipid moiety has been used extensively. A consideration of the importance of the lipid moiety in GSL function has led to novel chemical approaches that attempt to retain this property.

Glycosphingolipids (GSLs) are sugar–lipid conjugates expressed on the outer bilayer leaflet of the plasma membrane of all eukaryotic cells. The linear and branched chain carbohydrate structure is determined by a series of Golgi-lumen-located glycosyl transferases that add single sugars from nucleotide sugar donors transported into the Golgi lumen. The acceptor specificity and intraGolgi location of these glycosyl transferases determine the carbohydrate sequence of the GSLs eventually expressed on the plasma membrane. However, there still remains considerable uncertainty as to the mechanism of their sorting via the vesicular secretory pathway and their recycling from the cell surface. Differential sorting of GSLs to the apical and basolateral plasma membrane of polarized cells has been attributed to a combination of carbohydrate and lipid moieties in relation to the domain organization within the membrane (1, 2).

Glycosphingolipid Structure and Metabolism

A typical GSL, globotriaosyl ceramide galactose 1-4 galactose 1-4 glucosyl ceramide (Gb3), is shown in Fig. 1. The sugar sequence of membrane-embedded GSLs can adopt several conformational energy minima because of a restriction around the glucosyl anomeric link as a function of the relative plane of the membrane in which the GSL is contained (3). Molecular modeling of glucosyl ceramide defined nine thermodynamic minima of which the three most favorable are shown for Gb3. Modeling of blood group A GSLs showed that the internal sugar sequence greatly can affect the presentation of the A-epitope from perpendicular to parallel to the cell surface (4). Membrane parallel orientation of GSLs may be limited for longer sugar chains because a calculation of the phi/psi angles of tetra- and pentasaccharide globoseries GSLs showed that these termini can clash with the membrane for some conformers (5). These restrictions are dependent on the relative plane of the plasma membrane in relation to the glucose anomeric link. This relative plane, in turn, is defined by the composition of the ceramide moiety and the phospholipid bilayer. These are distinct in the liquid ordered vs the liquid crystalline phase so it is very likely that the partitioning of GSLs in and out of lipid microdomains significantly affects the conformation of the oligosaccharide.

Ceramide

The synthesis of the ceramide lipid moiety of GSLs is initiated in the endoplasmic reticulum (ER) (6) via the condensation of serine and palmitoylCoA. This enzyme has been cloned (7). The product is 3-keto-sphinganine that then is reduced to sphinganine and acylated by ceramide synthase with a long-chain fatty acid to form dihydroceramide, which is converted to ceramide by ceramide desaturase, which inserts the 4,5-trans double bond. GSLs are highly heterogeneous in terms of fatty acid composition. Fatty acid chain lengths vary from C16 to C24 and can be monounsaturated or 2′hydroxylated in mammalian systems. The functional importance of the heterogeneity has not been defined but may relate to the receptor function of GSLs and their intracellular trafficking (8). Such trafficking, in turn, is related to the membrane organization of GSLs in terms of their ordered or disordered domain structure (9). This membrane organization is itself a property of the GSL lipid moiety. The fatty acid heterogeneity was thought to be a function of the availability and the lack of specificity of the anabolic enzyme, but recently a family of ceramide synthases has been identified (Lass1-6) (10, 11) with a restricted and, in some cases, unique fatty acid
glycosphinoglycans.

Figure 1: Structure of the glycosphingolipid, globotriaosylceramide, Gal β1-3Galβ1-4Glc cermamide. The three lowest energy conformations as calculated by molecular modeling (8) are shown.

Neutral GSLs

Greater than 90% of the GSLs are based on the transfer of glucose to ceramide to give glucosyl ceramide. Dihydroceramide also can serve as a substrate (23). Glucosyl ceramide then is extended by lactosyltransferase to give lactosyl ceramide. Lactosyl ceramide is the precursor of all the major GSL series. Unlike all other glycosyltransferases involved in GSL biosynthesis, glucosyl ceramide synthesis is a cytosolic enzyme (26-28). Except for this initial enzyme reaction, glycolipid biosynthesis occurs within lumen of the Golgi (29). Glucosyl ceramide synthase transfers glucose from UDP-glucose to ceramides to form glucosyl ceramide on the outer cytosolic surface of the Golgi vesicles. Glucosyl ceramide needs to flip into the Golgi lumen to provide the precursor for the majority of GSL biosynthesis. This translocation is achieved, at least in the majority of cells, by the ABC transporter, MDR1. MDR1 (or P-glycoprotein) was discovered first because of its ability to mediate cytotoxic drug efflux in drug-resistant cells. The up-regulation of Pgp has been shown subsequently to be a major mechanism for drug resistance in cancer cells (30), and MDR1 inhibitors commonly are used therapeutically (31).

MDR1 is a member of a family of ATP-dependent drug efflux pumps (32). Several of these pumps, including MDR1, have been shown to mediate lipid translocation and phospholipid bilayers (33). This translocation activity has been implicated in part as the mechanism for drug efflux by which the hydrophobic cytotoxic drug is translocated from the cytosolic to the external leaflet of the plasma membrane bilayer. MDR1 first was shown to translocate glucosyl ceramide analogs from the cytosolic to the external leaflet of the plasma membrane (33). However, the major glucosyl ceramide translocation activity of MDR1 later
was shown to occur in Golgi vesicles (34). The transfection of cells with MDR1 resulted in a major increase in glycolipid biosynthesis that was prevented by MDR1 inhibition (35). In a cell-free microsomal system, exogenous glucosyl ceramide was incorporated exclusively into more complex GSLs in an MDR1-dependent manner (34). In most cells, MDR1 inhibition prevents neutral, but not acidic, GSL synthesis (34), which suggests that the lactosyl ceramide first made from the pool of glucosyl ceramide translocated into the Golgi lumen by MDR1 is unavailable for immediate ganglioside synthesis. The inhibition of MDR1 therefore represents a means to regulate selectively the neutral GSL biosynthesis (36). This regulation, however, raises a problem for the synthesis of gangliosides. Some GlcCer has been shown to be translocated directly to the cell surface without the involvement of transGolgi traffic (37,38). Recently, LacCer and GlcCer generated by cell-surface sialidase-mediated plasma membrane GM3 turnover was found to be incorporated exclusively into gangliosides (39). Thus, potentially, gangliosides could be made from cell-surface GlcCer, derived either by direct transit from the Golgi lumen by MDR1 or by degradation of other plasma membrane GSLs at the cell surface (or lysosomes).

Lactosyl ceramide synthase is a β-1,4 galactosyl transferase (40). Lactosyl ceramide is the only glycosylconjugate that contains lactose (as opposed to lactosamine). Lactosyl ceramide is the substrate for several glycosyl transferases that then define a core sequence of glycosphingolipids. The Globo series is defined by an α-1,4 galactosyl transferase activity (41,42) to make initially globo triosyl ceramide (also defined as the α series) and GD77 (43) and is the substrate for Gb4 synthase (a β-1,3 GalNAc transferase) (44). Gb4 is the major neutral glycolipid of most cells, including human red blood cells. Gb4 (also termed the P antigen (45)) is the precursor of the Forssman antigen formed by the action of an α-1,3 GalNAc transferase (46,47). Ganglio series glycolipids are made by the action of a β-1,3 GalNAc transferase (48) on lactosyl ceramide to give gangliotetraosyl ceramide, which then is the substrate of α-1,4 Gal transferase (type 2 chain); Lacto series glycolipids are made by the action of β-1,3 GalNAc transferase (49) and β-1,4 Gal transferase (type 2 chain). Lacto series GSLs are defined by the action of β-1,3 GalNAc transferase (49) and β-1,4 Gal transferase (type 2 chain) (51).

Gangliosides

Gangliosides are the major acidic GSLs and are derived primarily from the α-2,3 sialylation of lactosyl ceramide to give the simplest monosialylated GSLs, GlcCer (52). The sialic acid of GlcCer can be α-2,8 sialylated further to give GD3 (53), which can be sialylated further to form GT3. GlcCer, GD3, and GT3 all can serve as substrates (in addition to lactosyl ceramide) for the gangliosides of glycosyl transferases to form the “α”, “β”, and “γ” series of gangliosides, respectively (Fig. 3).

The sialic acid (N-acetylneuraminic acid) of gangliosides and sialylglycoproteins can be N-glycolyl or N-acetyl, and particularly in the context of the lectins that bind such glycoconjugates, this reality is of evolutionary significance (54). Many modifications of sialic acid (primarily 0-acetylation at the 7,8,9
Figure 3. Structure of the major acidic GSLs (gangliosides). Lactosylceramide, GM3, GD3, and GT3 are substrates for the same series of glycosyl transferases reactions, namely, GalNac transferase, galactosyl transferase II, and sialyl transferase V. Positions (α2-3 or α2-6) have been described (55) that are primarily α2-3 linked to galactose within sialoconjugates. The total synthesis of the most common gangliosides has been described (56-59). Gangliosides are abundant in neuronal tissue, and mouse knockout studies indicate a stabilizing function in axons in white matter (60). Gangliosides provide the receptors for myelin-associated glycoprotein (MAG) inhibition of neurite outgrowth (61-63), and their mimics may be of therapeutic value in neuronal injury (54).

Gangliosides within lipid rafts also can modulate growth factor receptor- and src kinase-mediated tyrosine phosphorylation signaling cascades (66). This modulation can be a carbohydrate—carbohydrate-based interaction (67). Although the molecular basis of such flexible sugar-sugar binding is not known, a series of selective interactions have been defined (68) within the context of a “glycosignaling domain” (69, 70) or a “glycosynapse” (71). GM3 inhibits phosphorylation in such domains, but deacetylation of the sialic acid results in the promotion of signaling (72), which indicates the central importance of the sialic acid. Indeed, sialyl-sphingosine was found to be the minimum structure to block signaling (73). Ganglioside signal modification plays an important role in distinguishing motile metastatic tumor cells from sessile “normal” cells (74, 75).

Galactosyl ceramide-based GSLs

A few GSLs are based on galactosyl ceramide rather than glucosyl ceramide. Galactosyl ceramide synthase, unlike glucosyl...
Inhibitors of glycosphingolipid biosynthesis

L-cycloheximide is an inhibitor of serine palmityltransferase (85), the first committed step in GSL biosynthesis. However, because this inhibitor prevents the synthesis of the highly bioactive ceramide, an interpretation of GSL depletion is equivocal.

Fumonisin B1 (Fig. 4) is a fungal inhibitor of ceramide synthase because of structural similarity to sphingamine (86). This inhibition also depletes ceramide levels, however. Moreover, the identified fatty acid selective ceramide synthases have proven to be fumonisin resistant (10, 11). From the structure, it is evident that fatty acid chain length-specific ceramide synthases might have a binding site that is too restricted to accommodate the branched terminus of this inhibitor. Sensitive to fumonisin also is present in the sphingolipid base because acylation of sphingosine rather than sphinganine in vitro is nonexistent to fumonisin (87), which is consistent with the saturated alkyl chain of this amino alcohol.

The natural form of ceramide is the D-erythro isomer, and this form is the substrate for glucosyl ceramide synthase, which forms the precursor to the majority of GSLs. The imino glucopyranosyl analog, butyl deoxynojirimycin (DNJ), is a potent inhibitor of glucosyl ceramide synthase and thereby prevents the synthesis of the majority of GSLs. This inhibitor has been used as a substrate reduction approach to several GSL storages. However, DNJ also inhibits glucose oxidase and thereby affects intracellular glucopyranose carbohydrate processing (90) and glucose storage (91), therefore, more selective inhibitors have been developed. An inhibitor of glucosyl ceramide synthase was designed to mimic the product of this reaction structurally. S-Phenyl-2-decanoylamino-3-morpholino-1-propanol (PDMP) (Fig. 5) is a membrane permeable and has been used extensively to examine the role of glucosyl ceramide-based GSLs in a variety of physiologic processes (92). However, PDMP was found to have additional inhibitory activities (93), for example, on other enzymes for which ceramide is a substrate such as ceramide acyltransferase. Several generations of derivatives of PDMP have been designed subsequently with improved specific activity for glucosyl ceramide synthase. The first was to lengthen the hydrocarbon chain from 10 to 16 carbons (84) (PPMP) and then to replace the morpholino ring with a pyrrolidino ring (P4) (95). Later substitutions within the phenyl ring were examined (96) and the 4-hydroxy derivative was found to be the most potent inhibitor (IC50 for the GC synthase is 90 nM).

Although the use of these inhibitors has been highly informative, the functional aspects of GSLs in plasma membranes have been difficult to define. Changes in growth status, differentiation, and malignancy are associated with altered GSL profiles (97). Ganglioside modification of growth factor receptor kinase activity (a coreceptor function) may relate to these changes (98). This difficulty in defining the functional aspects of GSLs in plasma membranes is largely because of the lack of methodologies for varying the expression of any specific GSL.

The overall importance of GSLs to cell physiology has been demonstrated clearly by using molecular biology techniques. Many glycosyltransferases involved in their synthesis have been cloned. As indicated above, the glucosyl ceramide synthase is the first committed step in the synthesis of most GSLs. This enzyme has been knocked out in mice and shown to be embryonic lethal (99). Interestingly, cultured cells survive well without this enzyme (100), which indicates that GSLs are required for the more differentiated functions in embryogenesis. Other enzymes involved in ganglioside biosynthesis, for example, GM3 synthase and GM2 synthase (60), have provided a more selective approach to reduce ganglioside biosynthesis, and knockout mice that lack these enzymes have been shown to have a less severe phenotype, primarily confined to CNS defects in long-term neuronal cell function. Knocking out the galactosyl ceramide synthase gene in mice results in the formation of dysfunctional and unstable myelin (78). A similar phenotype was observed for the 3′sulfotransferase knockout mouse, which indicates that SGC is the major species responsible for these defects (84). Several previous studies had shown the importance of SGC and its glycerol-based analog, sulfoglucosylceramide in spermatogenesis (83) and fertilization (101). Similarly, the sulfation of galactosyl ceramide in the brain is a function of myelination (102). The calcium-mediated interaction between galactosyl ceramide and SGC demonstrated to occur in vitro (103) is implicated strongly in myelin compaction. Ganglioside binding to myelin-associated glycoprotein (MAG) has been shown to be responsible for some negative signals that prevent neuronal cell regeneration (61).

GSL carbohydrate binding complementarity has been proposed as a mechanism for signal transduction, primarily by Hakomori (71), who has shown the binding interaction between GM3 ganglioside and gangliotriosyl ceramide lactosylceramide, SSEA3 and Gb4. These in vitro binding selectivities correlate with cellular adhesion between cells that express high levels of these GSLs (68). Nevertheless, the mechanism by which the flexible GSL carbohydrate chain can interact specifically with another flexible carbohydrate chain in these GSLs is unknown. The Hakomori group also has shown that gangliosides can modulate the growth factor receptor function and has proposed a cis-interaction between the receptor and the...
Glycosphingolipids, Chemistry of

Figure 4 Structure of Fumonisin B1.

1-phenyl-2-decanoylamino-3-morpholino-1-propanol

4′-hydroxy-1-phenyl-2-palmitoylamino-3-pyrrolidino-1-propanol

Figure 5 Structure of chemical inhibitors of glucosyl ceramide synthase: PDMP and hydroxy P4.

ganglioside within the plane of the plasma membrane (75). Sim- lar lateral interactions between receptors and GSLs have been proposed for the α2-interferon receptor (104, 105) and CO19 (106) with globotriaosyl ceramide. Both these transmembrane protein receptors contain an extracellular N-terminal sequence with similarity to the sequence within the receptor binding domain of the verotoxin (Shiga toxin) B subunit (107). Interest ingly, α2-interferon receptor signaling has been compared with that of γ-interferon. Both these receptors are internalized by clathrin-coated pits, but only α2-interferon signaling is pre- vented by the inhibition of internalization (108).

Lipid Rafts

The concept of cellular lipid rafts was introduced first by Simons (1, 109) to explain differential GSL trafficking in polarized epithe- lial cells. These domains are cholesterol and GSL enriched. The H-bond donor and acceptor capability of sphingolipids result in a close, rigid cholesterol complex—a lipid-ordered domain—as compared with the liquid crystalline, predominantly glycerolipid bilayer. In addition, membrane proteins anchored through glycoprophosphoinositol linkages preferentially accumu- late in these domains. Palmitoylated or myristylated cytosolic enzymes involved in downstream signal transduction pathways [e.g., src family kinases (110, 111)] accumulate in the cor- responding cytosolic domains of such cell surface lipid rafts. Lipid rafts serve as foci for a variety of signal transduction pathways (112) and also as access points for many pathogenic microorganisms (113). Direct GSL-mediated signaling within such rafts has been documented by using bacterial toxins that specifically recognize GSLs within these domains (114, 115). Indeed, cholera toxin binding to its receptor GSL, GM1 ganglio- side, is the gold standard marker for lipid rafts (116) in virtually all studies. However, lipid rafts as monitored by CT binding can be excluded from clathrin-coated pits (117). Moreover, recent studies have shown that the binding of cholera toxin does not correlate always with GM1 content (118) but rather fucosylated GM1 to which cholera toxin also binds (119). This finding cer- tainly will require a reevaluation of the basis of the association of cholera toxin with lipid rafts.

Phase separation in model lipid membranes is established clearly (120), but the significance of lipid rafts in cell mem- branes has been controversial (121) largely because of the harsh conditions of detergent resistance. More re- cently, methods have been developed to visualize lipid rafts in living cells (122, 123) and to identify proteins within them by less invasive methods (124).

The functional role of GSLs with in rafts is speculative but may involve linkage to cytosolic, raft-associated signal trans- duction cascades such as src family kinases (125). In this regard, we have demonstrated recently the structural similar- ity of 3′-sulfogalactose, found in GSLs, and tyrosine phosphate (126). This structural similarity is represented in Fig. 6.
that bound tyrosine phosphate and/or tyrosine sulfate (including SH2 domains from src) were found also to bind 3′sulfogalactose GSLs. Furthermore, the binding of such ligands to tyrosine sulfate lipid conjugates was found to depend on the lipid structure, a property typical of GSL recognition. We interpret this finding to indicate that the recognition of GSL carbohydrate within the context of the cell membrane and the recognition of the selective ligand binding of specific tyrosine phosphate groups in the context of a specific polypeptide background may share common principles. In modeling studies, we showed that a galactose hydroxyl group could replace a bound water molecule in tyrosine phosphate-ligand complexes, as determined by crystallography. A similar substitution binding of the carbohydrate could provide an entropic advantage such that carbohydrate-based inhibitors of specific tyrosine phosphatases/kinases might be feasible.

Terminal N-acetyl hexosamine

We described a procedure for the selective removal of terminal N-acetyl hexosamines from GSLs (132). In this procedure, the N-acetyl group first is deacetylated selectively with an aqueous base. Then, in the presence of H₂O₂, the amino function at the 2 position of the terminal amino sugar undergoes a free radical rearrangement reaction that results in the loss of the terminal sugar and some additional peeling reaction, with the likely release of NO. Internal N-acetyl hexosamine residues are unaffected. We propose a mechanism in Fig. 7. Peroxide oxidation of the amine to a nitroso group (133) results in the electronic reconfiguration to cleave the glycosidic bond and liberate NO. Although free amino sugars are rare, they might occur in lysosomes in the context of peroxide radicals, which might make this reaction a biologic source of NO.

Sphingosine double bond

Oxidative cleavage of the sphingosine double bond by ozonolysis was reported some 30 years ago (134). However, this reaction could be carried out only in large scale, and yields of the truncated GSL were very poor and required an ozone generator. This procedure was carried out with globoside and hematoside, and the equivalent glycoceramidic acid was characterized as the truncated product. We developed a new, more versatile procedure for the oxidative cleavage of the double bond of the sphingosine of glycolipids by using a more controllable potassium permanganate/potassium iodate oxidation system. This procedure allowed the oxidation of GSLs at the microgram level and under neutral conditions and generated the ceramidic acid in quantitative yield. This occurrence allowed a subsequent coupling to various inert supports (135). This coupling route has been used to conjugate glycolipids to protein carriers for generating GSL immunogens as potential anticancer therapies. However,
under basic conditions this oxidation resulted in greater truncation of the sphingoid base to give the glycosyl serine acid rather than the ceramide acid produced under neutral conditions (136). A hydroxy-acyl intermediate was identified; the oxidation of which was pH dependent. Unsaturated fatty acids also were cleaved such that the oxidation procedure could be used to define the aglycone composition of natural GSLs. The mass spectrometry analysis of the oxidative cleavage of glucosyl ceramide is shown in Table 1.

**Ganglioside coupling reactions**

Procedures for the selective deacylation of the ceramide moiety of gangliosides (137) have been used as routes to form ganglioside conjugates, primarily for immunogen development.

---

**Table 1** Mass spectroscopy of oxidized glucosyl ceramide that shows that the oxidation procedure allows the quantitation of a degree of unsaturation within the fatty acid moiety (taken from Reference 136)

<table>
<thead>
<tr>
<th>Gb4/Cer</th>
<th>before oxidation</th>
<th>after oxidation</th>
<th>oxidized intermediate</th>
<th>ceramide monosaccharide</th>
<th>ceramide disaccharide</th>
<th>carboxylic acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gb4/Cer</td>
<td>910.0 (100)</td>
<td>991.0 (45)</td>
<td>991.0 (51)</td>
<td>796.6 (50)</td>
<td>614.8 (36)</td>
<td>806.6 (39)</td>
</tr>
<tr>
<td>Gb4/Cer</td>
<td>910.0 (39)</td>
<td>991.0 (21)</td>
<td>991.0 (23)</td>
<td>796.6 (23)</td>
<td>614.8 (16)</td>
<td>806.6 (17)</td>
</tr>
<tr>
<td>Gb4/Cer</td>
<td>1017.2 (69)</td>
<td>1040.9 (53)*</td>
<td>1040.9 (53)*</td>
<td>832.0 (43)</td>
<td>874.6 (58)</td>
<td>806.6 (58)</td>
</tr>
<tr>
<td>Gb4/Cer</td>
<td>1045.2 (50)*</td>
<td>1070.9 (46)*</td>
<td>1070.9 (46)*</td>
<td>880.8 (31)</td>
<td>902.8 (31)</td>
<td>900.8 (24)</td>
</tr>
<tr>
<td>Gb4/Cer</td>
<td>1043.2 (78)</td>
<td>1071.2 (54)*</td>
<td>1071.2 (54)*</td>
<td>C12: 806.6 (39)</td>
<td>806.6 (39)</td>
<td></td>
</tr>
</tbody>
</table>

---

*Relative intensities are given in parentheses, and isotopic masses are indicated by superscript numbers.*

---

Figure 7: H2O2-mediated, nonreducing terminal hexosamine cleavage. Terminal amino sugar loss is the major product (but some peeling of the n-1, n-2 sugar is seen), possibly via the rearrangement of a nitroso oxidation product to liberate NO and generate a free radical that rearranges to induce the partial loss of the next sugar, and so forth.
Gb3 was internalized by a clathrin-independent mechanism in the cells with BSA. In addition, whereas BODIPY-labeled these BODIPY- or NBD-labeled GSLs, when incorporated into the membrane (9, 148). This effect is illustrated by the fact that ficking can be affected markedly by the lipid moiety within the ficking and partition into membranes does not mimic necessarily in terms of cell uptake. C6 NDB ceramide was used as a fluorescent probe of retrograde transport from the cell surface in cultured cells, Shiga toxin internalization via endogenous Gb3 was clathrin dependent in the same cells (149).

Nevertheless, the use of fluorescent GSL analog trafficking within cells to define aberrant intracellular trafficking pathways in cells from lysosomal storage disease patients has proven valuable in establishing the link between cholesterol and GSL trafficking. Indeed, the aberrant trafficking of BODIPY-GSLs can be used as a diagnostic tool in these genetic diseases (150).

A recent alternative procedure for the fluorescent labeling of GSLs has been described; it attempts to reduce the impact of the fluor on the structure of the GSL lipid moiety (147). A strong alkali hydrolysis of gangliosides was reported to cleave the ceramide fatty acid and deacylate the sialic acid group without affecting the N-acetyl galactosamine residue. (However, in our hands we find that the strong base simultaneously deacylates N-acetyl amino sugars while decaying the ceramide moiety of neutral GSLs; see Fig. 9). The protection of the sphingosine amine with fmoc then can allow the selective reacylation of the sialic acid with acetic anhydride. The removal of the fmoc group with piperidine in DMF allows the selective coupling of the lyso-ganglioside to azido octodecanoate. The selective re-duction of the azido group without reducing the sphingosine double bond was achieved by using hydrogen sulfide. NBD fluoride then was coupled to this 2-amino function within the lipid moiety of the ganglioside (Fig. 8). In previous coupling procedures for the generation of fluorescent glycosphingolipids, the NBD has been coupled directly to the sphingosine amine. In this procedure, the advantage is that the lipid moiety of the fluorescent ganglioside is more similar to that of the native species. Because many receptor functions of GSLs are modulated by the lipid moieties, this approach provides a mechanism by which such effects will be retained, at least in part, within the derivatized GSL. This conversion is supported by the effect that the GM2 activator was shown to bind the fluorescent GM2 analog as effectively as the natural GM2 ganglioside. The disturbing effect of such chromophore conjugations within the lipid moiety has been reduced more by the use of polynucleotide substitution with fatty acids and coupling to sphingosine (151).

GSL fatty acid substitutions

Many studies have substituted the fatty acid moieties of GSLs with various fluorophores to study GSL cellular trafficking and ligand binding properties. Nitrobenzoxadiazol (NBD) has been a popular choice (143). Early studies use the direct coupling of NBD via N-hydroxysuccinimidyl ester condensation with the sphingosine amine of lyso-sphingolipids. NBD with various fatty acid chain lengths were coupled, but C6 was most useful in terms of cell uptake. C6 NBD ceramide was used as a fluorescent probe of retrograde transport from the cell surface to the Golgi apparatus and metabolized to NBD-GSLs (144). Procedures were developed later to couple the fluor within the GSL lipid moiety selectively. With neutral GSL, this coupling is achieved easily by base hydrolysis (145). For gangliosides, selective cleavage of the base labile ceramide was required first (137).

These studies were superceded largely by the use of boron dipyrromethene difluoride (BODIPY), which has the useful property of showing a differential fluorescence emission spectrum according to concentration (146). The BODIPY moiety could be coupled with a variety of hydrocarbon spacer groups, and as with NBD the 5-carbon fatty acid analog has been used frequently. BODIPY analogs of galactose, lactose sulfates, GM1 ceramide, and sphingomyelin have been synthesized. The structure of examples of NBD- and BODIPY-labeled GSLs is shown in Fig. 8. However, these fluorophore substitutions considerably alter the lipid characteristics of these analogs such that their trafficking and partition into membranes does not mimic necessarily the natural GSL, because it is an intracellular trafficking can be affected markedly by the lipid moiety within the membrane (9, 148). This effect is illustrated by the fact that these BODIPY- or NBD-labeled GSLs, when incorporated into cell membranes, can be back-extracted readily by the treatment of the cells with BSA. In addition, whereas BODIPY-labeled Gb3 was internalized by a clathrin-independent mechanism in the culture of examples of NBD- and BODIPY-labeled GSLs is shown in Fig. 8. However, these fluorophore substitutions considerably alter the lipid characteristics of these analogs such that their trafficking and partition into membranes does not mimic necessarily the natural GSL, because it is an intracellular trafficking can be affected markedly by the lipid moiety within the membrane (9, 148). This effect is illustrated by the fact that these BODIPY- or NBD-labeled GSLs, when incorporated into cell membranes, can be back-extracted readily by the treatment of the cells with BSA. In addition, whereas BODIPY-labeled Gb3 was internalized by a clathrin-independent mechanism in cultured cells, Shiga toxin internalization via endogenous Gb3 was clathrin dependent in the same cells (149).

Nevertheless, the use of fluorescent GSL analog trafficking within cells to define aberrant intracellular trafficking pathways in cells from lysosomal storage disease patients has proven valuable in establishing the link between cholesterol and GSL trafficking. Indeed, the aberrant trafficking of BODIPY-GSLs can be used as a diagnostic tool in these genetic diseases (150).

A recent alternative procedure for the fluorescent labeling of GSLs has been described; it attempts to reduce the impact of the fluor on the structure of the GSL lipid moiety (147). A strong alkali hydrolysis of gangliosides was reported to cleave the ceramide fatty acid and deacylate the sialic acid group without affecting the N-acetyl galactosamine residue. (However, in our hands we find that the strong base simultaneously deacylates N-acetyl amino sugars while decaying the ceramide moiety of neutral GSLs; see Fig. 9). The protection of the sphingosine amine with fmoc then can allow the selective reacylation of the sialic acid with acetic anhydride. The removal of the fmoc group with piperidine in DMF allows the selective coupling of the lyso-ganglioside to azido octodecanoate. The selective re-duction of the azido group without reducing the sphingosine double bond was achieved by using hydrogen sulfide. NBD fluoride then was coupled to this 2-amino function within the lipid moiety of the ganglioside (Fig. 8). In previous coupling procedures for the generation of fluorescent glycosphingolipids, the NBD has been coupled directly to the sphingosine amine. In this procedure, the advantage is that the lipid moiety of the fluorescent ganglioside is more similar to that of the native species. Because many receptor functions of GSLs are modulated by the lipid moieties, this approach provides a mechanism by which such effects will be retained, at least in part, within the derivatized GSL. This conversion is supported by the effect that the GM2 activator was shown to bind the fluorescent GM2 analog as effectively as the natural GM2 ganglioside. The disturbing effect of such chromophore conjugations within the lipid moiety has been reduced more by the use of polynucleotide substitution with fatty acids and coupling to sphingosine (151).

Radiolabeling of GSL

GSL radiolabeling is achieved commonly by the catalytic re-
duction of the sphingosine double bond using palladium and sodium borotrifluoride (152), although this, of course, generates the dihydro species that may behave differently from the native ceramide-containing glycosphingolipid. Radiolabeling with galactose oxidase is a feasible alternative that produces an aldehyde at the C6 position of terminal galactose residues, which can be reduced again with sodium borotritide (153, 154). This procedure can be adapted to label only cell surface GSLs (155). Interestingly, this reaction was found to fail when the glycolipid had a glycerol, as opposed to a sphingosine, backbone (156), which indicates the that availability of the C6 position of galac-
tose to the enzyme is restricted in the case of the glycolipid species.

Surprisingly, a good chemical method for radiolabeling all glycosphingolipids has yet to be devised. The decylation of the sphingosine under basic conditions and the recylation of...
generated amino function with radiolabeled fatty acid is a feasible method, although this method does eliminate the fatty acid heterogeneity of the native species. However, this approach is problematic when it comes to gangliosides because the sialic acid also is base labile. Similarly, in amino sugar-containing GSLs, basic hydrolysis will result in the deacetylation of the N-acetyl amino sugar, which thereby precludes the subsequent selective acylation of the sphingosine amine. We have overcome this problem in terms of amino sugar-containing GSLs by developing a procedure for the selective reacylation of either the sphingosine amine or the galactose (glucose) amino function (Fig. 9).

Chemical Glycosphingolipid Synthesis

The carbohydrate moieties of GSLs have been synthesized, but these procedures are essentially no different from the chemical synthesis of oligosaccharides found on glycoproteins (157). The major difference is the coupling of the oligosaccharide to the aglycone. Complex GSLs have been synthesized by the assembly of appropriately protected oligosaccharide donors and the coupling to a ceramide acceptor and deprotection under mild conditions, for examples, see References 158-160. These approaches are restricted by the complexity of the protection chemistry, intermediate purification, and low product yields. Combined enzymatic and chemical syntheses can address these problems. A recent innovative approach has been to reverse the hydrolytic activity of the endoglycoceramidase that normally removes the carbohydrate from GSLs, to generate a novel glycosynthase (161). In this approach, the catalytic nucleophile residue of the hydrolase is mutated to an inactive amino acid, and the reverse reaction with the appropriate fluoride sugar donor then becomes readily apparent. This process removes problems with protecting group chemistry and ensures the correct stereochemistry of the adduct.

Thioglucosyl ceramide and thiolactosyl ceramide also have been synthesized to study cellular glycolipid traffic and metabolism (162). Both \( \beta \)-thiolactosyl ceramide and lactosyl \( \beta \)-thioceramide were made. The thioglycosidic link is resistant to glycosidase degradation to ceramide, and these analogs therefore can be used as precursors to monitor cellular GSL anabolism alone. The \( \beta \)-thiolactosyl ceramide was made by coupling a protected thio-Glc-β-Gal to C-4 of a protected galactosyl sphingosine. The inversion of configuration during the \( S_N2 \) reaction generated the correct \( \beta \)-thiolactosyl ceramide. Lactosyl \( \beta \)-thioceramide and glucosyl \( \beta \)-thioceramide were made by condensation of the protected thiourea glucoside or the thioacetyl lactoside donor with a protected iodosphingosine.

The synthesis of glycolipid analogs suitable for microarrays is an important component for defining carbohydrate-binding phenotypes in “glycomics.” These conjugation procedures that allow the solid-phase presentation of the glycolipid carbohydrate moiety are most physiologic when as much of the intact ceramide moiety of the parent GSL as possible is maintained. Serine-based GSL analogs, at least in part, accomplish this. This has been achieved by the use of fluoro hydroxy methophenoxyl acid as the linker to conjugate to microtitre plates (163).
Glycosphingolipids, Chemistry of

A new procedure for GSL synthesis via olefin cross metathesis (164) is highly versatile in terms of the hydrophobic aglycone. A protected 5 carbon amino alkene diol is the central building block to which the protected carbohydrate donor, long chain fatty acid, or, by olefin cross metathesis, the long alkenyl chain of the base can be coupled, in a variety of sequences. This atypical synthetic flexibility should allow a structural approach to dissecting the role of the lipid moiety in GSL receptor function and intracellular trafficking.

Glycosphingolipid mimics

GSLs often serve a receptor function in microbial pathogenesis. Therefore, one therapeutic approach to such infections is the development of soluble inhibitors of this interaction, and such inhibitors are for the most part based on the generation of high-affinity receptor mimics. Ligands that bind GSL sugars that are distal to the glycone are less likely to be affected by the modification or removal of the lipid moiety. Cholera toxin binding to GM1 ganglioside is affected marginally by the nature of the lipid moiety (165). This occurrence has allowed the development of GM1 mimics in which the ceramide is absent. CT binds to a conformer that is similar to the global energy minimum of the free GM1 oligosaccharide (166), and the sugar has been paired away such that the sialic acid is reduced to lactic acid (167) linked to an appropriate conformationally restricted 1,2 dicarboxy-4,5-cyclohexanediol (168) that serves as a simplified mimic of the 3,4 substituted core galactose. This has been multimerized by using a tetraevalent scaffold to improve the affinity for the multivalent toxin B subunit pentamer (169).

Although the dendritic presentation of carbohydrates for glycoprotein binding lectins has been well established (170, 171), the potential multivalency of GSLs within the plasma membrane bilayer is far greater than for glycoproteins, and thus the chemical solution to the generation of synthetic oligosaccharide multimers that retain a binding affinity of the order of the natural membrane GSL has been particularly difficult. In terms of the Escherichia coli-derived verotoxin (also termed Shiga toxin), which is responsible for the development of hemolytic uremic syndrome primarily in children (172), the importance of the lipid moiety of its GSL receptor, globotriaosyl ceramide (Gb3), is clearly apparent (173). Moreover, the pentameric receptor

Figure 9 Selective ceramide deacylation of N-acetylhexosamine that contains GSLs. Gangliotriaosyl ceramide is shown as example. In organic solvent, 1 M base hydrolysis results in the cleavage of both amides, but acetylation in aqueous buffer only affects the amino sugar.

Dry 3M NaOH / MeOH
70 °C / 4 days
AC2O / PBS

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

11 - 346 -

- 346 -
binding B subunit indicates at least five binding sites, which are suggested in the original crystal structure of the VT1 B subunit to be within the intersubunit cleft (174). Modeling studies supported this conclusion (175). However, subsequent co-crystallographic studies with the Gb3 oligosaccharide indicated as many as three binding thermodynamic minimalmonomer (35 in total) (176), which makes the design of inhibitors difficult. The interplay of these potential receptor sites still remains uncertain (177). Although the binding affinity of the lipid-free Gb3 oligosaccharide was of the order of 5 log weaker than that of the intact GSL (178), this affinity was considered an avidity problem, which could be solved by the design of an appropriate dendrimer presentation. The first, and probably the most elegant, approach was the synthesis of a pentamer of globotriaose dimers. This de-cavalent pentamer was based on coupling a globotriaose dimer to a central glucose moiety (179). The intersubunit cleft site was not the major binding site but rather a shallow trough on the surface opposite the bound membrane of each subunit monomer. This surface binding site, site 2, was targeted for occupancy by the pentameric globotriaose array. This “starfish” array proved an effective neutralizing ligand for VT1 but less so for VT2 in terms of both binding inhibition and prevention of cell cytotoxicity in vitro. Subsequent modifications of this pentameric array such that the separation between the globotriaose dimers was increased to accommodate potentially the distance between site 2 and the inter subunit cleft site (site 1) improved efficacy, and this was attributed to dual binding site occupancy (180); however, this was not seen in cocyrstals. Rather, each globotriaose within each dimer occupied equivalent sites in adjacent B subunit pentamers—a pentameric sugar sandwich—and crosslinked adjacent B subunit pentamers to inhibit cytotoxicity. Subsequent studies found that a trimer, rather than the pentamer, of dimers was a more effective ligand substitute for VT2 (181). The crystal structure of VT2 (182) has shown that the C-terminus of the A subunit that penetrates the B subunit pentamer would obstruct receptor binding in the site 3 equivalent of this toxin. The loop that contains the equivalent of site 2 is in a different conformation that also would not allow Gb3 carbohydrate binding. Thus, VT2 may show receptor binding in site 1 only. The other procedure described for the generation of multivalent globotriaose dendrimers for protection against VT1 or VT2 cytotoxicity in vivo is based on carbosilane multimers (183). These multimers are more systematic expansions of trimeric structures that are based on multiples of the tetrameric carbosilane (184). In these studies, coupling the two terminal silicon residues of a linear carboxilane trimer (“super twig”) to either 3 or 9 globotriaose oligosaccharide and the terminal to be substituted with three additional carbosilane units proved the most effective in neutralizing verotoxin in mice. Site-specific mutational studies on the B subunits of VT1 and VT2 showed that sites 1 and 2 were required for “super twig” VT1 binding but that site 3 was central for VT2 binding (184). However, this finding does not reconcile with the VT2 crystal structure that shows that site 3 is inaccessible (182). It is significant that whereas these more densely displayed globotriaose dendrimers proved more effective than the “rational designed” pentameric displays discussed above, both of these synthetic dendrimeric presentations proved less effective than the random coupling of globotriaose to a polyacrylamide support (185), which indicates that a considerable gap still exists in relating the structure of the Gb3 oligosaccharide-verotoxin B subunit complex to the Gb3 GSL binding on target cells (177). The other major GSL receptor function is that of GM1 ganglioside for cholera toxin. In contrast to the verotoxin-Gb3 system, no attempts to generate GM1 receptor mimics to prevent cholera toxin binding have been reported. Although the lipid moiety of GM1 ganglioside plays a less significant role in cholera toxin binding than that which Gb3 plays in VT1 recognition, a chemical substitution of the lipid moiety of GM1 ganglioside has been shown to modify cholera toxin cytotoxicity (186). Substitution of the ceramide moiety for cholesterol, aliphatic amines, or aminophospholipids has been shown to modify the cholera toxin response. The increased aliphatic chain length increased efficacy. The lower functional limit (C12) is similar to the fatty acid chain length dependency for the verotoxin receptor GSL function (186). The GM1-cholesterol analog proved more effective than the native GM1, which may relate to the GM1 raft requirement for cholera toxin cytopathicity (116,116), and implies differential intracellular trafficking of the toxin receptor mimic complexes. GM1 ganglioside is the gold standard marker of lipid raft assembly in plasma membranes (187). A mutation of the cholera toxin B subunit such that a nonraft ganglioside was bound preferentially resulted in the abolition of cholera toxin cytopathicity (116,116). Thus, although the lipid moiety may not be crucial for cholera toxin binding, it is crucial for plasma membrane receptor organization, which in turn defines toxin efficacy. HIV envelope adhesion gp120 binding to several GSLs has been reported (189, 190), and GSL depletion prevents HIV infection (191). HIV binding to galactosyl ceramide is implicated in the infection of CD4 negative cells (192), which has spurred the synthesis of several analogs as a means to control this infection (193–195). The glucose ring of lactose was opened and coupled to aminoundecanoic sodium carboxylate. The amino group then was C16 acylated more. This soluble analog inhibited HIV infection and syncytium formation in vitro (193). Gal Cer analogs were made in which the sphingosine base was truncated via the construction of galactosyl “seritol” (196). This was not bound by gp120, showing the importance of the hydrocarbon chains. Although conjugates that contain a C9 sphinganine did not bind gp120 (196), these could inhibit gp120/GalCer binding according to aglycone structure. Hydroxylation at the C2 position of the fatty acid (even for short chain fatty acids) promoted this inhibition (196). OH and increased hydrocarbon chain length correlated with the calculated head group conformation—at a right angle to the hydrocarbon axis, which indicates that this is the major aglycone effect. Water-soluble C-glycoside analogs of o-lyso-galactosyl ceramide were made (197) via C-glycosyl aldehyde condensation with Wittig reagent to form the trans double bond that was then reduced; the ring was opened and oxidized to form a protected C-glycosyl amino acid. This C-glycosyl amino acid then was coupled to aliphatic amines of increasing chain length. Gp120 binding of the C16 species was equivalent to o-lyso-galactosyl ceramide, and binding was hydrocarbon chain length-dependent.
The GSL carbohydrate binding specificity of gp120 in vitro is lax. The binding to N-stearyl-deoxylactosaminimidocyclon (198) prompted the synthesis of stearyl C-glycoside and azo-C-glycoside analogs of galactosyl ceramide (199). A common precursor, a C1 stearyl galactal, was constructed; it was hydroborated to give the C-glycoside, or it was transformed to a diketone from which the azo-C-glycoside was derived by reduction and amination. By Langmuir trough measurements, these stearyl glycosides showed equal or greater binding to gp120 than galactosyl ceramide. An assay of their efficacy against HIV infection will be of interest. In contrast to the O-glycosides, fatty acid–OH and head group conformation were not found important. This finding could relate to the -OH/anomeric oxygen H-bond in the O-glycosides that restricts the headgroup conformation (196). In the C-glycosides, rotation will be reduced, and this may not be necessary. In addition, a substitution of the anomeric oxygen will make the interface between the “sugar” and the hydrocarbon chain more hydrophobic, which will alter packing and thereby might reduce the effect of the hydrocarbon chain length seen for the O-glycoside. Galactosyl ceramide analogs with hydrophilic (ethylene glycol) spacer groups of increasing length might be restricted to facilitate ligand binding. We have re-considered those sugar sequences adjacent to the ceramide moiety, might be restricted to facilitate ligand binding. We have replaced the fatty acid of several GSLs with rigid carbon frames such as adamantane (202). These structures are surprisingly water-soluble and yet, unlike the lipid-free carbohydrate, retain the receptor function of the native membrane-bound GSL. Our premise is that selection of appropriate chemical substitutions within these frames can promote ligand binding and indeed modulate intracellular trafficking pathways. Adamantyl Gb3 (Pip 3D) is an effective inhibitor of verotoxin GB3 binding (203) and can protect cells against VT cytopathology in culture. However, adamantyl Gb3 is not effective in protecting against VT2 cytopathology in the mouse model (204). Adamantyl Gb3 is, nevertheless, a highly effective ligand for the HIV envelope-protein gp120 (205) and prevents HIV infection in vitro, irrespective of strain or drug-resistant status (206), via inhibition of viral-host gp120-dependent cell fusion. Using a Langmuir trough, we established that the molecular area of adamantyl Gb3 is greater than that of native Gb3. Moreover, an adamantyl Gb3 monolayer was more rigid than that of Gb3. HIV gp120 interaction with Gb3 monolayers was slow and with sigmoidal kinetics, whereas that with adamantyl Gb3 monolayers was rapid and exponential (205). Gp120 interaction with Gb3 became exponential in the presence of cholesterol, which suggests that adamantyl Gb3 may mimic a Gb3:cholesterol complex.

Members of the hsp70 family of stress proteins bind the GSL 3′sulfogalactosyl ceramide (SGC) (207). The binding site is in the N-terminal ATPase domain (208). Adamantyl SGC, similarly generated via fatty acid replacement, similarly has proven water soluble and is an effective inhibitor of hsp70-ATPase activity (209) in vitro and therefore may modulate its chaperone function in cells. Such an effect also has therapeutic potential (210). The high water solubility of these adamantyl GSL derivatives was surprising considering that the adamantane frame is at least as hydrophobic as the fatty acid it replaced. It is possible, however, that the adamantane frame together with the sphingosine acyl chain may pack sufficiently to be below the hydrophobic size that is necessary to disturb the molecular organization of water (211). This situation may provide the basis of their aqueous solubility characteristics.

Acknowledgments

Studies from our laboratory have been supported by grants from CIHR and the Ontario HIV Treatment Network.

Figure 10 Adamantyl Gb3 is a soluble mimic of Gb3 GSL.

The GSL carbohydrate binding specificity of gp120 in vitro is lax. The binding to N-stearyl-deoxylactosaminimidocyclon (198) prompted the synthesis of stearyl C-glycoside and azo-C-glycoside analogs of galactosyl ceramide (199). A common precursor, a C1 stearyl galactal, was constructed; it was hydroborated to give the C-glycoside, or it was transformed to a diketone from which the azo-C-glycoside was derived by reduction and amination. By Langmuir trough measurements, these stearyl glycosides showed equal or greater binding to gp120 than galactosyl ceramide. An assay of their efficacy against HIV infection will be of interest. In contrast to the O-glycosides, fatty acid–OH and head group conformation were not found important. This finding could relate to the -OH/anomeric oxygen H-bond in the O-glycosides that restricts the headgroup conformation (196). In the C-glycosides, rotation will be reduced, and this may not be necessary. In addition, a substitution of the anomeric oxygen will make the interface between the “sugar” and the hydrocarbon chain more hydrophobic, which will alter packing and thereby might reduce the effect of the hydrocarbon chain length seen for the O-glycoside. Galactosyl ceramide analogs with hydrophilic (ethylene glycol) spacer groups of increasing length might be restricted to facilitate ligand binding. We have re-considered those sugar sequences adjacent to the ceramide moiety, might be restricted to facilitate ligand binding. We have replaced the fatty acid of several GSLs with rigid carbon frames such as adamantane (202). These structures are surprisingly water-soluble and yet, unlike the lipid-free carbohydrate, retain the receptor function of the native membrane-bound GSL. Our premise is that selection of appropriate chemical substitutions within these frames can promote ligand binding and indeed modulate intracellular trafficking pathways. Adamantyl Gb3 (Pip 3D) is an effective inhibitor of verotoxin GB3 binding (203) and can protect cells against VT cytopathology in culture. However, adamantyl Gb3 is not effective in protecting against VT2 cytopathology in the mouse model (204). Adamantyl Gb3 is, nevertheless, a highly effective ligand for the HIV envelope-protein gp120 (205) and prevents HIV infection in vitro, irrespective of strain or drug-resistant status (206), via inhibition of viral-host gp120-dependent cell fusion. Using a Langmuir trough, we established that the molecular area of adamantyl Gb3 is greater than that of native Gb3. Moreover, an adamantyl Gb3 monolayer was more rigid than that of Gb3. HIV gp120 interaction with Gb3 monolayers was slow and with sigmoidal kinetics, whereas that with adamantyl Gb3 monolayers was rapid and exponential (205). Gp120 interaction with Gb3 became exponential in the presence of cholesterol, which suggests that adamantyl Gb3 may mimic a Gb3:cholesterol complex.

Members of the hsp70 family of stress proteins bind the GSL 3′sulfogalactosyl ceramide (SGC) (207). The binding site is in the N-terminal ATPase domain (208). Adamantyl SGC, similarly generated via fatty acid replacement, similarly has proven water soluble and is an effective inhibitor of hsp70-ATPase activity (209) in vitro and therefore may modulate its chaperone function in cells. Such an effect also has therapeutic potential (210). The high water solubility of these adamantyl GSL derivatives was surprising considering that the adamantane frame is at least as hydrophobic as the fatty acid it replaced. It is possible, however, that the adamantane frame together with the sphingosine acyl chain may pack sufficiently to be below the hydrophobic size that is necessary to disturb the molecular organization of water (211). This situation may provide the basis of their aqueous solubility characteristics.

Acknowledgments

Studies from our laboratory have been supported by grants from CIHR and the Ontario HIV Treatment Network.
45. Spitalnik P, Spitalnik S. The P blood group system: Bio-
44. Okajima T, Nakamura Y, Uchikawa M, Haslam DB, Nu-
43. Mangeney M, Richard Y, Coulaud D, Tursz T, Wiels J. CD77:
42. Steffensen R, Carlier K, Wiels J, Levery SB, Stroud M, Cederen
49. Togayachi A, Akashima T, Ookubo R, Kudo T, Nishihara S,
48. Nagata Y, Yamashiro S, Yodoi J, Lloyd KO, Shiku H, Fu-
47. Fujii Y, Numata SI, Nakamura Y, Honda T, Furukawa K, Urano
46. Haslam DB, Baenziger JU. Expression cloning of Forssman gly-
39. Valaperta R, Chigorno V, Basso L, Prinetti A, Bresciani R,
37. van IJzendoorn SC, Zegers MM, Kok JW, Hoekstra D. Segrega-
20. Henning TR, Zhou D, Wolfer DP, Jungalwala FB, Herent T,
Glycosphingolipids, Chemistry of


--- 350 ---
**Glycosphingolipids, Chemistry of**


147. Aref S, Lingwood C. Intracellular targeting of the endoplasmic reticulum membrane envelope by retrograde transport may determine cell hypersensitivity to Verotoxin: sodium butyrate or selective deletion of resistance may induce nuclear toxin targeting via glucosylceramide fatty acid isoform traffic. J. Cell Physiol. 1999;177:646–660.


Starch is a major storage compound in plants that is present both in leaves and in storage tissues. Biochemical and molecular biological data show that ADP-glucose is the glucosyl donor for plant starch synthesis, and its synthesis is catalyzed by ADP-glucose pyrophosphorylase. Subsequently, starch synthases catalyze the transfer of the glucosyl residue from ADP-glucose to the oligosaccharide chains of the starch components amylose and amylpectin to form new α-1,4-glucosidic residues. After elongation of these α-1,4-glucosidic chains, the branching enzyme catalyzes a cleavage of the elongated chain and transfers the cleaved portion of the oligosaccharide chain to either another region in the amylopectin molecule or to a new amylpectin and forms a new α-1,6-glucosidic linkage. Amylose synthesis is catalyzed by the granule-bound starch synthase. Regulation of starch synthesis occurs at the ADP-glucose pyrophosphorylase step. The enzyme from higher plants, green algae, and cyanobacteria is activated allosterically by 3-phosphoglycerate and inhibited by inorganic phosphate. Isolation of mutants and control analyses indicate that the allosteric activation and inhibition are of physiological and functional importance in the regulation of starch synthesis. Furthermore, evidence indicates that ADP-glucose pyrophosphorylases can also be regulated by a redox mechanism. The current knowledge of the enzyme structures and critical amino acids necessary for substrate binding, allosteric effector binding, regulation, and catalysis for the ADP-glucose pyrophosphorylase is reviewed.

The enzymatic reactions involved in starch synthesis in higher plants and algae and in glycogen synthesis in cyanobacteria are reviewed in this article. Regulation of α-1,4 and α-1,6 glucan synthesis at the enzymic level is discussed, and mutants that cause specific enzymatic deficiencies that affect starch structure are also reviewed. Recent reviews on starch synthesis have been published and are noted (1–6).

Localization of Starch in Plants

Starch is present in every type of tissue, such as leaves, fruit, pollen grains, roots, shoots, and stems. Formation of starch granules occurs in the chloroplast during exposure of leaves to bright light because of carbon uptake during photosynthesis. Loss of leaf starch occurs in low light or in long dark periods of 24-48 hours when it is degraded to products that are used for sucrose synthesis. Arabidopsis thaliana mutants that cannot synthesize starch grow as well as wild type when placed in continuous light as they can synthesize sucrose (7). However, if grown in a day-night regime, the growth rate is dramatically decreased because starch is required for sucrose synthesis at night. Thus, leaf starch synthesis and use is a dynamic process with diurnal fluctuations.

Starch synthesis occurs during development and maturation of plant storage organs. During sprouting or germination of the seed or tuber or in fruit ripening, the reserved starch is degraded, and its metabolites are sources for carbon and energy. Thus, reserved starch degradation and synthesis are time-separated events. Cereal grain starch-storage tissue is the endosperm, with starch granules located in the amyloplasts. Reserve starch content varies from 65% to 90% of the total dry weight.

The starch granule contains two polysaccharide types: amylose, which is a linear polymer, and amylpectin, which is a branched polymer. Amylose is composed of linear chains of about 800 to 22,000 α-D-glucopyranosyl units in (α-1,4) linkage.
Figure 1  Cluster model of amylopectin (10). A indicates A chains. A chains have no other oligosaccharide chains attached to them. B indicates B chains that are defined as having one or more oligosaccharide chains (either A or B) attached to them. The C chain, which is not labeled, has the only reducing end group \( \Omega \) in the polysaccharide. B3 chains are longer than B2 chains, which are longer than the B1 chains. The B2, B3, and B4 (not shown) chains extend into 2, 3 and 4 cluster regions, respectively. The average chain lengths are B1, 19; B2, 41; B3, 169; and B4, 109. For the A chains, the shortest chain length is about 13.

(molecular size \( \sim 1.36 \times 10^5 \) to \( 3.5 \times 10^6 \)). Amylose molecules may be slightly branched (branch chain, 170 to 500 glucosyl units). Amylopectin, which comprises about 70-80% of the starch granule, has 4-5% of the glucosidic linkages as branch points.

The accepted models of proposed amylopectin structures, which are known as cluster models, are those of Robin et al. (8), Manners and Matheson (9), and Hizukuri (10). Reviews that detail chemical and physical properties of starch are given by Morrison and Karkalis (11) and Hizukuri (12). Figure 1 shows the proposed cluster structure.

Starch Synthesis Reactions in Plants and Algae and Glycogen Synthesis in Cyanobacteria

ADP-glucose is used for synthesis of \( \alpha-1,4 \) glucosidic linkages in amylose and amylopectin. Its synthesis is catalyzed by ADP-glucose pyrophosphorylase (reaction 1, E.C. 2.7.7.27; ATP: \( \alpha \)-D-glucose-1-phosphate adenylyltransferase).

\[
\text{ATP + } \alpha\text{-glucose 1-P } \rightleftharpoons \text{ADP-glucose + PPI} \tag{1}
\]

Synthesis of the \( \alpha-1,4 \) glucosidic bond in starch is catalyzed by starch synthase (E.C. 2.4.1.21; ADP-glucose: 1,4-\( \alpha \)-D-glucan 4-\( \alpha \)-glucosyltransferase; reaction 2) and was first reported by Leloir (13, 14). Similar reactions are seen for glycogen synthesis in bacteria (15), and the enzyme is termed glycogen synthase (also E.C. 2.4.1.21).

\[
\text{ADP-glucose + 1, 4-\alpha-glucan } \rightarrow \text{1, 4-\alpha-glucosyl-1, 4-\alpha-glucan + ADP} \tag{2}
\]

Synthesis of the \( \alpha-1,6 \) glucosidic bonds in amylopectin is catalyzed by a branching enzyme (E.C. 2.4.1.18; 1,4-\( \alpha \)-D-glucan 6-\( \alpha \)-glucosyltransferase; reaction 3). Amylopectin contains longer chains (about 20-24 glucosyl units) and has less branching (~5% of the glucosidic linkages are \( \alpha-1,6 \)) than animal or bacterial glycogen (10-13 glucosyl units and 10% of linkages, \( \alpha-1,6 \)).

\[
\text{Elongated } \alpha-1,4\text{-malto oligosaccharide chain } \rightarrow \text{1, 4-1, 6 branched- } \alpha\text{-Glucan} \tag{3}
\]

Isozymes of plant starch synthases (2, 16–20) and branching enzymes (2, 16, 21–26) have been characterized. Presumably, they play distinct roles in amylopectin and amylose synthesis. Also, plant and Chlamydomonas reinhardtii granule-bound starch synthases (27–33) are responsible for synthesis of amylose. Mutants defective in this enzyme are known as waxy mutants and contain starch granules that have little or no amylose.

A debranching enzyme called isoamylase is involved in synthesis of the starch granule and its polysaccharide components (34–37). Mutant plants deficient in isoamylase activity accumulate a soluble \( \alpha \)-glucan designated phytoglycogen (3, 37, 38) and little starch.

Properties of the Plant 1,4-\( \alpha \)-Glucan Synthesizing Enzymes

ADP-glucose pyrophosphorylase: kinetic properties

The ADP-glucose pyrophosphorylases (ADP-Glc PPases) of higher plants, green algae, and the cyanobacteria are activated allosterically by 3-phosphoglycerate (3-PGA) and inhibited by inorganic phosphate (Pi). These effects are important in regulation of starch synthesis. Both potato tuber (2, 39-49) and spinach leaf ADP-Glc PPases (50–52) have been studied in detail with respect to kinetic properties. The kinetic and regulatory properties of the ADP-Glc PPases of several leaf extracts are similar to those of the spinach leaf enzyme (50). The results obtained with potato tuber and spinach leaf enzyme are summarized in Table 1.
heterotetrameric-subunit structure. The potato tuber ADP-Glc PPase has been shown to be heterotetramers with two homologous subunits in plants and green algae, however, the ADP-Glc PPase is homotetrameric in structure. Bacterial ADP-Glc PPase is composed exclusively of 85–95% identity (54). The small subunit of many higher plant ADP-Glc PPases is active in the presence of high concentrations of the activator 3-PGA. The PPase small subunit can be expressed as a homotetramer and is the activator 3-PGA than the heterotetramer (A0.5 = 0.63 mM) (42). The large subunit of the potato greatly increases the affinity of the small (catalytic) subunit for 3-PGA and lowers the affinity for the inhibitor Pi (41, 42).

In plants, only one conserved small (catalytic) subunit and several large (regulatory) subunits are distributed in different parts of the plant (56, 57). This finding is of physiological significance as expression of different large subunits in different plant tissues may confer distinct allosteric properties to the ADP-Glc PPase needed for the plant tissue’s distinct need for starch.

It has been shown with Arabidopsis ADP-Glc PPase that expression of its small-subunit APS1 with the different Arabidopsis large subunits, which include APL1, APL2, APL3, and APL4, resulted in heterotetramers with different regulatory and kinetic properties (56) (Table 2 and Table 3). The heterotetramer of the small-subunit APS1 with APL1, which is the predominant leaf large subunit, had the highest sensitivity to the allosteric effectors 3-PGA and Pi as well as the highest apparent affinity for the substrates ATP and Glc-1-P. The heterotetrameric pairs of APS1 with either APL3 or APL4, which are large subunits prevalent in sink or storage tissues (57), had a lower apparent affinity for the substrates ATP and Glc-1-P and a higher affinity for the inhibitor Pi (I0.5 = 0.08 mM in the presence of 3-mM 3-PGA) as compared with the heterotetramer (I0.5 = 0.63 mM) (42).

Table 1: Kinetic constants of ADP-Glc PPases from Spinach leaf and potato tuber. S0.5 is the concentration of substrate ATP or Glc-1-P required to attain 50% of Vmax. A0.5 is the concentration of activator 3-PGA required for 50% maximal activation. I0.5 is the concentration of inhibitor need for 50% inhibition. Table 2: Kinetic parameters for the 3-PGA of A. thaliana recombinant ADP-Glc PPase in the synthesis direction (56)
The similarity between the small and large subunits (large subunits
relationships conferred by the Arabidopsis large subunits were found in vitro. Differences noted for source and sink large-subunit pro-
teins strongly suggest that starch synthesis is modulated in a tissue-specific manner in response to 3-PGA and Pi, as well as to the substrate levels. APS1 and APL1 would be regulated finely in source tissues by both effectors and substrates, whereas in sink tissues, the heterotetramers of APS1 with APL2, APL3, or APL4 with lower sensitivity to effectors, and substrates would be controlled more by the supply of substrates. Based on mRNA expression, APS1 is the main small subunit or catalytic isoform responsible for ADP-Glc PPase activity in all tissues of the plant. APL1 is the main large subunit in source tissues, whereas APL3 and APL4 are the main isoforms present in sink tissues. It was also found that sugar regulation of ADP-Glc PPase genes was restricted to APS1 and APL4 in leaves (57). Sucrose induction of APL3 and APL4 transcription in leaves allowed formation of heterotetramers less sensitive to the allosteric effectors, which resembles the situation in sink tissues.

Relationship between the small and large subunits

The similarity between the small and large subunits (~50 to 60% identity) suggests a common origin (54). In both sink and source tissues, the small subunit has catalytic activity, whereas catalytic activity is only observed for the large subunits that may reside in the leaf and not in the sink large subunits (55). Gene duplication and divergence probably has led to different and functional catalytic and regulatory roles for the subunits. The ancestor of small and large subunits possibly is a bacterial subunit that has both catalytic as well as regulatory function in the same subunit. This suggestion is supported by the similarity between the two plant subunits with many active bacterial ADP-Glc PPases (5, 53).

The large subunit from the potato (Solanum tuberosum L.) tuber ADP-Glc PPase was shown to bind substrates (44). Therefore, the plant heterotetramer, and bacterial homotetramers bind four ADP- [14 C] glucose molecules (44, 58). It can be postulated that the large subunit maintained its structure needed for binding of substrate, but catalytic ability was eliminated by mutations of essential residues. To test this hypothesis, it was attempted to create a large subunit with significant catalytic activity by mutating as few residues as possible (49). Thus, sequence alignments of ADP-Glc PPase large and small subunits with reported activity were compared to identify critical missing residues for catalytic activity in the large subunit (49). The subset of the residues absent in the large subunit was of particular interest. In the large subunit of potato tuber, Lys44 and Thr54 were selected as the best candidates to study because the homologous residues Arg33 and Lys43 in the small (catalytic) subunit were conserved completely in the active bacterial and plant catalytic subunits. Moreover, Lys44 and Thr54 are in a highly conserved region of ADP-Glc PPases (Table 4).

The modulatory large subunits Lys44 and Thr54 were mutated to Arg44 and Lys44, respectively. The mutant, LargeK44R/T54K, was expressed in the absence of the small subunit, and it had no activity. Possibly the large subunit cannot form a stable tetramer in absence of the small subunit as observed earlier with the Arabidopsis enzyme (56). Because a wild-type small subunit has intrinsic activity, the activity of the large-subunit mutants cannot be tested when coexpressed. Thus, the large-subunit mutants were coexpressed with inactive small-subunit D145N, in which the catalytic residue Asp145 was mutated (59), and small subunit activity was reduced by more than three orders of magnitude (Table 5). Coexpression of the large-subunit double mutant K44R/T54K with smallD145N generated an enzyme that had 10% and 18% of the wild-type enzyme in the ADP-glucose synthetic and pyrophosphorylating directions, respectively (Table 5). Single mutations of K44R or T54K generated enzymes with no significant activity. The combination of both mutations in the large subunit (smallD145N/largeK44R/T54K) gave the most dramatic effect (Table 5). Therefore, it was concluded that the two residues Arg44 and Lys54 are needed for restoring catalytic activity to the large subunit. Replacement of the homologous two residues with Lys and Thr in the small subunit (mutations R33K and K43T) decreased the activity by one and two orders of magnitude, respectively (Table 5).

The similarity between the small and large subunits (~50 to 60% identity) suggests a common origin (54). In both sink and source tissues, the small subunit has catalytic activity, whereas catalytic activity is only observed for the large subunits that may reside in the leaf and not in the sink large subunits (55). Gene duplication and divergence probably has led to different and functional catalytic and regulatory roles for the subunits. The ancestor of small and large subunits possibly is a bacterial subunit that has both catalytic as well as regulatory function in the same subunit. This suggestion is supported by the similarity between the two plant subunits with many active bacterial ADP-Glc PPases (5, 53).

Table 5

<table>
<thead>
<tr>
<th>APS1</th>
<th>ATP Glc-1-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>APS1/APL1</td>
<td>0.402 ± 0.04</td>
</tr>
<tr>
<td>APS1/APL2</td>
<td>0.067 ± 0.008</td>
</tr>
<tr>
<td>APS1/APL3</td>
<td>0.575 ± 0.03</td>
</tr>
<tr>
<td>APS1/APL4</td>
<td>0.094 ± 0.008</td>
</tr>
<tr>
<td>APS1</td>
<td>0.116 ± 0.01</td>
</tr>
</tbody>
</table>

20 mM of 3-PGA was used for the APS1 enzyme in determining the ATP and the Glc-1-P kinetic constants. For the APS1/apl1 kinetic study, 0.1 mM of 3-PGA was used. 4 mM of 3-PGA was used for the APS1/apl2. 2 mM of 3-PGA was used for the APS1/apl3, and 1 mM of 3-PGA was used for the APS1/apl4 kinetic studies. These concentrations were five times the 3-PGA A0.5 of each enzyme, respectively.

Intermediate sensitivity to the allosteric effectors and intermediate affinity for the substrates ATP and Glc-1-P (57). APL2 also present mainly in sink tissues had low affinity for either 3-PGA or Pi (56). Thus, differences on the regulatory properties conferred by the Arabidopsis large subunits were found in vitro. Differences noted for source and sink large-subunit proteins strongly suggest that starch synthesis is modulated in a tissue-specific manner in response to 3-PGA and Pi, as well as to the substrate levels. APS1 and APL1 would be regulated finely in source tissues by both effectors and substrates, whereas in sink tissues, the heterotetramers of APS1 with APL2, APL3, or APL4 with lower sensitivity to effectors, and substrates would be controlled more by the supply of substrates. Based on mRNA expression, APS1 is the main small subunit or catalytic isoform responsible for ADP-Glc PPase activity in all tissues of the plant. APL1 is the main large subunit in source tissues, whereas APL3 and APL4 are the main isoforms present in sink tissues. It was also found that sugar regulation of ADP-Glc PPase genes was restricted to APS1 and APL4 in leaves (57). Sucrose induction of APL3 and APL4 transcription in leaves allowed formation of heterotetramers less sensitive to the allosteric effectors, which resembles the situation in sink tissues.
and not smallD145N was the catalytic subunit. In the wild-type small subunit did not indicate the large subunit double mutant, K213R mutation of a large subunit severely decreased the apparent affinities of Lys198 in the small subunit of the wild-type enzyme, decreased the Glc-1-P affinity was observed; disruption of the substrate site domain has been conserved. The apparent affinities for the substrates and the allosteric properties of small subunit were very similar kinetic properties, which indicated that the substrate site domain has been conserved. The apparent affinities for the substrates and the allosteric properties of small subunit were very similar kinetic properties, which indicated that the subunit is essential for catalysis, but homologous Asp160 in the large subunit is not essential for catalysis (59). In addition, mutation of D160 to N or E in the active large subunit, LK44R/T54K abolished activity. This result confirms that catalysis of smallD145Nlarge K44R/T54K does occur in the large subunit.

A comparative model of LK44R/T54K shows the predicted role of Arg44 and Lys54 (Fig. 2). In the model, Asp160, which is homologous to the catalytic Asp145 in the small subunit and catalytic Asp142 in the E. coli ADP-Glc PPase (59, 60), interacts with Lys54. This type of interaction (Lys54-Asp145) has also been observed in crystal structures of enzymes that catalyze similar reactions, such as dTDP-glucose pyrophosphorylase (dTDP-Glc PPase) and UDPN-N-acetyl-glucosamine pyrophosphorylase (UDPGalNAc PPase). It is postulated to be important for catalysis by orienting the aspartate residue correctly (61–63). Lys54 interacts with the oxygen that bridges the γ-phosphates of ATP, which correlates with the transition state and makes PPi a better leaving group. Arg44 interacts in the model with the ζ-phosphates as it has been observed in the crystal structure of E. coli dTDP-Glc PPase (63). The interaction neutralizes a negative charge density to stabilize the transition state and make PPi a better leaving group. A n44 interacts in the model with the ζ- and γ-phosphates of ATP, which correspond to the PPI byproduct (Fig. 2). Likewise, Arg15 in the E. coli dTDP-Glc PPase (63).

Table 5 Activity of potato tuber small (catalytic) and large (regulatory) subunit ADP-Glc PPase mutants (49)

<table>
<thead>
<tr>
<th>Subunits</th>
<th>ADP-glucose synthesis Units/mg</th>
<th>ATP synthesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>WT</td>
<td>32 ± 1</td>
</tr>
<tr>
<td>D145N</td>
<td>WT</td>
<td>0.017 ± 0.001</td>
</tr>
<tr>
<td>D145N</td>
<td>K44R</td>
<td>0.031 ± 0.001</td>
</tr>
<tr>
<td>D145N</td>
<td>T54K</td>
<td>0.92 ± 0.08</td>
</tr>
<tr>
<td>D145N</td>
<td>K44R/T54K</td>
<td>3.2 ± 0.2</td>
</tr>
<tr>
<td>R33K</td>
<td>WT</td>
<td>3.6 ± 0.2</td>
</tr>
<tr>
<td>K43T</td>
<td>WT</td>
<td>0.32 ± 0.1</td>
</tr>
</tbody>
</table>

The enzymes activities of purified coexpressed small and large subunits were measured for ADP-glucose synthetic activity or ADP-glucose pyrophosphorylase (ATP synthesis) activity. For ADP-glucose synthesis, 4 mM of 3-PGA (activator), 2 mM of ATP, and 0.5 mM of Glc-1-P were used. For pyrophosphorylase, 4 mM of 3-PGA, mM of ADP-Glc, and 1.4 mM of PPI were used.
The crystal structure of potato tuber homotetrameric small subunits ADP-Glc PPase was also inhibited by inorganic sulfate with the $K_i$ value of 2.8 mM in the presence of 6 mM 3-PGA (48). Sulfate is considered as an analog of phosphate, which is the allosteric inhibitor of plant ADP-Glc PPases. Thus, the atomic resolution structure of the ADP-Glc PPase probably presents a conformation of the allosteric enzyme in its inhibited state. The crystal structure of the potato tuber ADP-Glc PPase (48) allows one to determine the location of the activator and substrate sites in the three-dimensional structure and their relation to the catalytic residue Asp145. The structure also provides insights into the mechanism of allosteric regulation.

Supporting data for the physiological importance of regulation of ADP-glucose pyrophosphorylase

The reaction catalyzed by plant ADP-Glc PPases is activated by 3-P-glycerate and inhibited by orthophosphatase; it is an important step for regulation of starch synthesis in higher plants as well as in the cyanobacteria (see reviews in Reference 5 and 33). In addition, it has been shown that in higher plants the enzyme activity can also be regulated by its redox state (46, 47).

Considerable experimental evidence is available to support the concept that ADP-Glc PPase is an important regulatory enzyme for the synthesis of plant starch. A class of C. reinhardtii starch-deficient mutants have been isolated and shown to contain an ADP-Glc PPase not activated by 3-P-Glca (64, 65). Evidence for the allosteric regulation by ADP-Glc PPase being pertinent in vivo has also been obtained with A. thaliana (66, 67). Mutant TL25 lacked both subunits and accumulated less than 2% of the limits of detection of starch observed in the normal plant (66), which would indicate that starch synthesis is almost completely dependent on the synthesis of ADP-glucose. The mutant TL 46 is also starch deficient, and it lacked the regulatory 54-kDa subunit (67). The mutant had only 7% of the wild-type ADP-Glc PPase activity in potato tuber and wheat endosperm (71) leads to increase of starch. This finding also has been shown for rice (72) as well as for cassava root (73). These results confirm that ADP-Glc synthesis is rate limiting for starch synthesis. Thus, data observed with the allosteric mutant ADP-Glc PPases of C. reinhardtii (64), maize endosperm (68), and Arabidopsis (66, 67) present strong evidence that in vitro allosteric effects are functional in vivo.

**Phylogenetic analysis of the large and small subunits**

A phylogenetic tree of the ADP-Glc PPases present in photosynthetic eukaryotes also sheds information about the origin of the subunits (Fig. 3). The tree shows that plant small and large subunits can be divided into two and four distinct groups, respectively (49). The two main groups of small subunits are from dicot and monocot plants, whereas large-subunit groups correlate better with their documented tissue expression. The first large-subunit group, which is termed group 1, is generally expressed in photosynthetic tissues (49) and comprises large subunits from dicots and monocots. These subunits have been shown recently to have catalytic activity and have in their sequences Arg and Lys in the equivalent residues of 102 and 112 of A. thaliana large subunit, APL1. Group 2 displays a broader expression pattern, whereas groups 3 and 4 are expressed in storage organs (roots, stems, tubers, seeds). Subunits from group 3 are only from dicot plants, whereas group 4 includes seed-specific subunits from monocots. These last two groups stem from the same branch of the phylogenetic tree and split before monocot and dicot separation (49). These subunits are probably inactive in catalytic activity as they are lacking Arg and Lys in the homologous residues observed in A. thaliana APL1 and APL2.

**Crystal structure of potato tuber ADP-Glc PPase**

The crystal structure of potato tuber homotetrameric small subunits ADP-Glc PPase has been determined to 2.1Å resolution (48). The structures of the enzyme in complex with ATP and ADP-Glc were determined to 2.6 and 2.2 Å resolution, respectively. Ammonium sulfate was used in the crystallization process and was found tightly bound to the crystalline enzyme. It was also shown that the small-subunit homotetrameric potato tuber ADP-Glc PPase was also inhibited by inorganic sulfate with the $K_i$ value of 2.8 mM in the presence of 6 mM 3-PGA (48). Sulfate is considered as an analog of phosphate, which is the allosteric inhibitor of plant ADP-Glc PPases. Thus, the atomic resolution structure of the ADP-Glc PPase probably presents a conformation of the allosteric enzyme in its inhibited state. The crystal structure of the potato tuber ADP-Glc PPase (48) allows one to determine the location of the activator and substrate sites in the three-dimensional structure and their relation to the catalytic residue Asp145. The structure also provides insights into the mechanism of allosteric regulation.

**Figure 2** Involvement of large (regulatory) subunit mutant K44R/T54L in ADP-Glc PPase catalysis. The WT and double-mutant large subunits were modeled based on the dTDP-Glc PPase and UDP-UDPGlcNAc PPase crystal structures as indicated in the text. Portions of ADP-Glc PPase residues 31–73 and 111–134 are shown. The dessolvation triphosphate portion common to dTTP and ATP is modeled with Mg$^{2+}$ as a black sphere. The dotted lines depict hydrogen bonds.
Plant ADP-Glc PPases can be activated by thioredoxin

ADP-Glc PPase from potato tuber has an intermolecular disulfide bridge that links the two small subunits by the Cys12 residue; it can be activated by reduction of the Cys12 disulfide linkage (46). At low concentrations (10\(\mu\)M) of 3-PGA, both spinach leaf reduced thioredoxin f and m reduce and activate the enzyme. Fifty percent activation was observed for 4.5- and 8.7- \(\mu\)M reduced thioredoxin f and m (47). The activation was reversed by oxidized thioredoxin. Cys12 is conserved in the ADP-Glc PPases from plant leaves and other tissues except for the monocot endosperm enzymes. In photosynthetic tissues, this reduction may also be physiologically pertinent in the fine regulation of the ADP-Glc PPase.

Both potato tuber and potato leaf ADP-Glc PPases are plastidic; the leaf enzyme is in the chloroplast, and the tuber enzyme is in the amyloplast (74). The ferredoxin/thioredoxin system is located in the chloroplast and thus, with photosynthesis, reduced thioredoxin is formed and activated within the leaf ADP-Glc PPase. At night, oxidized thioredoxin is formed; it oxidizes and inactivates the ADP-Glc PPase. This activation/inactivation process during the light/dark cycle allows a fine tuning and dynamic regulation of starch synthesis in the chloroplasts. Thioredoxin isoforms are present in many different subcellular locations of plant tissues: cytosol, mitochondria, chloroplasts, and even nuclei (75) and are also present in amyloplasts (76).

It has been shown starch synthesis that occurs in potato tubers from growing plants is inhibited within 24 hours after detachment (77) despite having high in vitro ADP-Glc PPase activity and high levels of substrates, as well as ATP and Glc-1-P and an increased 3-PGA/Pi ratio. In the detached tubers, the small subunit in nonreducing SDS-PAGE is solely in dimeric form and relatively inactive in contrast to the enzyme form of growing tubers where it was composed as a mixture of monomers and dimers. The detached tuber enzyme had a great decrease in affinity for the substrates as well as for the activator. Treatment of tuber slices with either DTT or sucrose reduced dimerization of the ADP-Glc PPase small subunit and stimulated starch synthesis in vivo. These results indicate that inactivative activation, which was observed in vitro of the tuber ADP-Glc PPase (46, 47), is important for regulating starch synthesis (77). A strong correlation between sucrose content in the tuber and the reduced/activated ADP-Glc PPase was noted.

Characterization of ADP-Glc PPases from some different sources

As previously indicated, Table 1 summarizes the kinetic and regulatory properties of purified potato tuber and spinach leaf ADP-Glc PPases. As reviewed in References 5 and 53, the properties of many other plant, algae, and cyanobacterial ADP-Glc PPases are similar. However, some ADP-Glc PPases within plant reserve tissue show some differences with respect to allosteric properties and their nonplastidic location. These examples are summarized below.
Barley

The barley leaf ADP-Glc PPase has been purified to homogeneity (93 U/mg), and it shows high sensitivity toward activation by 3-PGA and inhibition by phosphate (78). Substrate kinetics and product inhibition studies in the synthesis direction suggested a sequential (ordered) Bi Bi kinetic mechanism (78). ATP or ADP-Glc bind first to the enzyme in the synthesis or pyrophosphorylation direction, respectively, which is similar to the E. coli enzyme (58).

Partially purified barley endosperm ADP-Glc PPase was shown to have low sensitivity to the regulators 3-PGA and Pi (78). However, 3-PGA decreased up to threefold the $S_{0.5}$ for ATP and the Hill coefficient. At 0.1-mM ATP, the activation by 3-PGA was around fourfold (79), and phosphate 2.5 mM reversed the effect. A recombinant enzyme with a (His)$_6$-tag from barley endosperm was expressed using the baculovirus insect cell system (80). It shows no sensitivity to regulation by 3-PGA and Pi. However, the enzyme was assayed at saturating concentration of substrates and only in the pyrophosphorylation direction. For ADP-Glc PPases, the synthetic direction is more sensitive to activation. When the recombinant enzyme without the (His)$_6$-tag is expressed in insect cells, the heterotetrameric form still was not activated by 3-PGA nor inhibited by Pi at saturating levels of substrates (82). Whether 3-PGA had any effect on the affinity for the substrates as shown in the enzyme purified from the endosperm was not reported. Of interest is that the small (catalytic) subunit, when expressed alone it is very sensitive to activation by 3-PGA, and inhibition by Pi is opposite of what is observed for large subunits of potato tuber (42, 45) and Arabidopsis (56).

Identification of Important Amino Acid Residues Within the ADP-Glc PPases

Amino acid residues that play important roles in the binding of substrates and allosteric regulators have been identified in the ADP-Glc PPases mainly by chemical modification and site-directed mutagenesis studies. Thus, photoaffinity analogs of ATP and ADP-Glc, 8-azido-ATP, and 8-azido-ADP-Glc were used to identify Tyr114 as an important residue in the enzyme from E. coli (95, 96). Site-directed mutagenesis of this residue rendered a mutant enzyme that exhibits a marked increase in $S_{0.5}$ for ATP as well as a lower apparent affinity for Glc-1-P and the activator Fru1,6-bisP (95). The Tyr residue must be close to the adenine ring of ATP or ADP-Glc but probably also near the Glc-1-P and the activator regulatory sites. The homologous Tyr114 in the enzyme from plants is a Phe residue (54), which suggests that the functionality is not given by the specific residue but by its hydrophobicity.

Chemical modification studies on the E. coli ADP-Glc PPase that showed involvement of Lys198 in the binding of Glc-1-P (97, 98) were confirmed by site-directed mutagenesis (99). Site-directed mutagenesis was also used to determine the role of this conserved residue in the small subunit Lys198 and large-subunit Lys213 of the potato tuber ADP-Glc PPase (44). Mutation of Lys198 of the small subunit with Arg, Ala, or Glu decreased the apparent affinity for Glc-1-P 135-fold to 550-fold. Little effect is observed on kinetic parameters for ATP, Mg$^+$, 3-PGA, and Pi. The results show that the Lys198 in the small subunit is involved directly in the binding of Glc-1-P. However, the homologous site residue Lys213 in the large subunit does not seem to be involved because similar mutations on Lys213 had little effect on the affinity for Glc-1-P (44). This finding is consistent with the view that the potato tuber large subunit is a modulatory subunit and does not have a catalytic role (42).

Maize endosperm

Partially purified maize endosperm ADP-Glc PPase (34 U/mg) was found to be activated by 3-PGA and Fru 6-P (23- and 17-fold, respectively) and inhibited by Pi (94). The heterotrimeric endosperm enzyme has been cloned and expressed in E. coli, and its regulatory properties were compared with an isolated allosteric mutant less sensitive to Pi inhibition (69). As indicated above, the increase of starch noted in the mutant maize endosperm ADP-Glc PPase insensitive to Pi inhibition supports the importance of the allosteric effects of 3-PGA and Pi in vivo. Also as indicated above, it is believed that the major endosperm ADP-Glc PPase isoform is located in the cytosol (83).
potato tuber ADP-Glc PPase. The homologous residues of the small subunit Asp145 and large subunit Asp160 were replaced separately by either Asn or Glu residues (59). Mutation of the Asp145 of the small subunit rendered the enzymes almost completely inactive. D145N mutant activity decreased by four orders of magnitude, whereas D145E, which is a more conservative mutation, decreased in specific activity by just two orders of magnitude. The homologous mutations in the large subunit alone (D160) did not alter the specific activity, but they did affect the apparent affinity for 3-PGA (59). Thus, these results agree with the view that each subunit in potato tuber ADP-Glc PPase plays a particular role: the small subunit controls catalysis and the large subunit plays a regulatory role.

Pyridoxal-5-phosphate (PLP) could be considered to have some structural analogy to 3-PGA, and it was found to activate both the enzymes from spinach leaf and *Anabaena*. In spinach ADP-Glc PPase, PLP bound at Lys440, which is very close to the C-terminal of the small subunit, and bound to three Lys residues in the large subunit. Binding to these sites was prevented by the allosteric effector 3-PGA, which indicated that they are close to or are involved directly in the binding of this activator (100, 101).

With *Anabaena* ADP-Glc PPase, PLP modified Lys439 and Lys382. That these residues were regulatory binding sites was confirmed by site-directed mutagenesis of the *Anabaena* ADP-Glc PPase (102, 103). Mutation of the homologous Lys residues in the potato tuber enzyme small-subunit Lys441 and Lys404 indicated that they were also part of the 3-PGA site in heterotetrameric ADP-Glc PPases and that they contribute additively to the binding of the activator (45). Moreover, mutation on the small subunit yielded enzymes with lesser affinity to 3-PGA, measured by 3090-fold and 54-fold, respectively, than the homologous mutants on the large subunit. Results indicate that Lys404 and Lys441 on the potato tuber small subunit are more important than their homologous counterparts on the large-subunit Lys417 and 455. It seems that the large subunit seems to contribute to the enzyme activation by making the activator sites already present in the small subunit more efficient rather than by providing more effective allosteric sites (45).

Arginine residues were found in ADP-Glc PPases from cyanobacteria to be functionally important as shown by chemical modification with phenylglyoxal (104, 105). A loss, the role played by Arg294 in the inhibition by Pi of the enzyme from *Anabaena* PCC 7120 was shown by Ala-scanning mutagenesis studies (105). More recently, it was found that replacement of this residue with Ala or Gin produces mutant enzymes with a changed pattern of inhibitor specificity; it was found that they have NADPH rather than Pi as the main inhibitor (126). All these results suggest that the positive charge of Arg294 may play a key role in determining inhibitor selectivity rather than being specifically involved in Pi binding. However, studies on the role of Arg residues located in the N-terminal of the enzyme from *A. tumefaciens* tumefaciens demonstrated the presence of separate subsites for the activators Fructose-6-P and pyruvate as well as a deamidization of R33A and R45A mutants to Pi inhibition (107).

Random mutagenesis experiments performed on the potato tuber ADP-Glc PPase have been useful to identify residues that are important for the enzyme. Mutation of Asp253 on the small subunit showed a specific effect on the Km for Glc-1-P, which increased 10-fold with respect to the wild-type enzyme (108). The small magnitude in the increase (only one, rather than three to four orders of magnitude) suggests that the Asp253 residue is not involved in directly Glc-1-P binding. This residue, however, is conserved in the NDP-sugar PPases that have been crystallized and the structure solved (61-63). Alignment of Asp253 in the structure according to the secondary structure prediction (109) places the residue close to the substrate site without direct interaction with Glc-1-P. This suggests that substitution of the Asp 253 causes an indirect effect on the Glc-1-P by alteration of the Glc-1-P-binding domain. Another random mutagenesis study (110) concerns A. sp416 (described in the article as Asp413) of the potato tuber ADP-Glc PPase small subunit and its effect on 3-PGA activation. This residue is adjacent to Lys417, which has been shown to be a site for PLP binding and 3-PGA activation (100). Also, several modifications on the C-terminus caused modifications on the regulation of different plant ADP-Glc PPases (69, 111).

The finding that Lys and Arg residues are important in allosteric effector binding and are situated at the C-terminus in ADP-Glc PPases of plants and cyanobacteria is different with what is observed for the bacterial ADP-Glc PPases. Lys338 (E. coli) and Arg residues in the N-terminal of the *A. tumefaciens* enzyme were shown to be important for the interaction of the activators and inhibitors (97, 98, 107, 112). These results suggest that the regulatory domains may be at different sites in the bacterial and plant enzymes. Other studies, however, with chimeric ADP-Glc PPases constructed from *E. coli* and *A. tumefaciens* have shown that the C-terminus in the bacterial and plant enzymes is functional in determining effector specificity and affinity (113). Regulation of ADP-Glc PPases most likely is determined by interactions that occur between the N- and C-termini in the enzyme.

**Characterization of the regulatory domain**

Truncation of 10 amino acids in the small subunit of potato tuber ADP-Glc PPase affected its regulatory properties by increasing the apparent affinity for the activator 3-PGA and decreasing the inhibitor Pi affinity (42). When the large (regulatory) subunit was truncated by 17 amino acids at the N-terminal, similar results were observed (114). The regulatory properties of the E. coli enzyme were also affected when the N-terminal is shortened by 11 amino acids (115, 116). The N-terminal region of the ADP-Glc PPase is predicted to be a loop, and the data suggests it regulates enzyme activity by acting as the “allosteric switch”; it is involved in the transition between the activated and nonactivated conformations of the enzyme. A shorter N-terminus allows the enzyme to be in an “activated” conformation.
Starch Biosynthesis In Plants

Table 6 Properties of maize endosperm-branching isozymes

<table>
<thead>
<tr>
<th>Branching enzymes</th>
<th>BEI</th>
<th>BEIIa</th>
<th>BEIIb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphorylase stimulation (a)</td>
<td>1196</td>
<td>795</td>
<td>994</td>
</tr>
<tr>
<td>Branching linkage assay (b)</td>
<td>2.6</td>
<td>0.32</td>
<td>0.14</td>
</tr>
<tr>
<td>Iodine stain assay (c)</td>
<td>800</td>
<td>29.5</td>
<td>39</td>
</tr>
<tr>
<td>Amylose (c1)</td>
<td>24</td>
<td>58</td>
<td>63</td>
</tr>
<tr>
<td>Amylopectin (c2)</td>
<td>460</td>
<td>2484</td>
<td>7100</td>
</tr>
<tr>
<td>Ratio of activity a/b</td>
<td>1.5</td>
<td>27</td>
<td>25</td>
</tr>
<tr>
<td>a/c1</td>
<td>49.8</td>
<td>13.5</td>
<td>15.8</td>
</tr>
<tr>
<td>a/c2</td>
<td>0.03</td>
<td>2</td>
<td>1.6</td>
</tr>
<tr>
<td>c2/c1</td>
<td>0.03</td>
<td>2</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Phosphorylase stimulation and branching linkage assays units are µmol/min (126). The iodine stain assay unit is (127), a decrease of one Absorbance unit per min.

Starch Synthases and Branching Enzymes

Starch synthase catalyzes transfer of the glucosyl moiety of ADP-glucose either to a maltosaccharide or to the starch polymers, amylase, and amylpectin, which form a new α-1,4-glucosidic linkage in ADP-glucose is, and thus the newly formed glucosidic linkage is retained. The starch synthase is therefore characterized as retaining GT-B glycosyltransferase (117). More than one form of starch synthase is present in many plant tissues. This finding has been summarized in several publications (1–4, 118), and the starch synthases are encoded by more than just one gene. Some starch synthases are bound to the starch granule and are designated as starch granule-bound starch synthases. They may be solubilized by α-amylase digestion of the granule, whereas others designated as soluble starch synthases can be found in the soluble portion of the plastid fraction.

Starch synthase or glycogen synthase activity can be measured by transfer of [14C]glucose from ADP-glucose into an appropriate primer such as amylopectin or rabbit glycogen and then precipitation of the labeled polymer with ethanol (119).

Characterization of the starch synthases

A phylogenetic tree based on the various deduced amino acid sequences of the plant starch synthases and green algae, Chlamydomonas, have identified five subfamilies of starch synthases (2). These subfamilies are known as granule-bound starch synthase (GBSS), starch synthase 1 (SS1), starch synthase 2 (SS2), starch synthase 3 (SS3), and starch synthase 4 (SS4). Additional data indicates that the SS2 class may have diverged even more to classes SS2a and SS2b (120). Also, the GBSS family may have also diverged into another class, GBSS1, GBSS1b or GBSS2 (121–123). The starch synthases are reviewed in detail in Reference 124.

Branching enzyme

As with starch synthases, many isozymes have been found for the Branching Enzyme, and they have been characterized in a number of plants.

Maize endosperm has three branching enzyme (BE) isoforms (24, 25) BE I, IIa, and IIb from maize kernels (3, 125, 126). Molecular weights were 82 KDa for BEI and 80 KDa for BE's IIa and IIb. Table 6 summarizes the in vitro properties of the various maize endosperm BE isoforms from the studies of Takeda et al. (127) and Guan et al. (24, 25). BEI had the highest activity in branching amylose, and its ratio of branching amylpectin was about 3% observed with amylose. The BEIIa and BEIIb isozymes branched amylopectin at twice the rate they branched amylose, and they catalyzed branching of amylopectin at 2.5 to six times the rate observed with BEI. BEI catalyzes the transfer of longer branched chains, and BEIIa and IIb catalyze the transfer of shorter chains (127). This finding may suggest that BEI produces a slightly branched polysaccharide that serves as a substrate for enzyme complexes of starch synthases and BEII isoforms to synthesize amylopectin. BE II isoforms may also play a predominant role in forming the short chains present in amylopectin. Moreover, BEI may be more involved in producing the more interior B-chains of the amylopectin, whereas BEIIa and BEIIb would be involved in forming the exterior (A) chains.

Potato branching enzyme thus shows a high degree of similarity to maize BEI and to maize BEII isoforms, however to a lesser extent (128, 129). As observed with the maize-branching enzymes, potato BEI was more active on amylose than BEII, and BEII was more active on amylopectin than BEI (130–132). Branching isozymes from rice (21, 133), wheat (134, 135), and barley (136, 137) have also been characterized. The effects on starch structure in plants that have mutations in the starch-branching enzymes have been reviewed (2–4, 118, 124).

References

Starch Biosynthesis in Plants


55. Ventriglia T, Luzón M, Ribeiro-Pedro M, Kuhn ML,
54. Smith-White BJ, Preiss J. Comparison of Proteins of ADP-
52. Morell MK, Bloom M, Knowles V, Preiss J. Subunit structure
50. Copeland L, Preiss J. Purification of spinach leaf ADP-glucose
47. Ballicora MA, Freuauf JB. Fu Y, Schürmann P, Preiss J. Ac-
45. Ballicora MA, Fu Y, Nesbitt NM, Preiss J. ADP-glucose py-
44. Fu, Y, Ballicora, MA and Preiss, J. Mutagenesis of the glucose-
41. Iglesias AA, Barry GF, Meyer C, Bloksberg L, Nakata PA,

Starch Biosynthesis In Plants

Starch Biosynthesis in Plants

80. Doan DN, Rudi H, Olsen OA. The allosterically unregulated isoform of ADP-glucose pyrophosphorylase from barley endosperm is the most likely source of ADP-glucose incorporated into endosperm starch. Plant Physiol. 1999;121:965–975.
A book with many good reviews on source-sink relationships and factors affecting source-sink interactions.


Other Reviews on Starch Metabolism


Chemistry of Lipid Domains

Richard M. Epand, McMaster University Health Sciences Centre, Hamilton, Ontario, Canada
doi: 10.1002/9780470048672.wecb283

Biological Background

To begin with, biological membranes are not uniform with respect to membrane sidedness. Both the protein and the lipid composition of these membranes exhibit transbilayer asymmetry. In addition, in polarized mammalian cells, a segregation of both protein and lipid molecules exists between the apical and basolateral surfaces (1–4). Both transbilayer asymmetry and segregation in polarized cells occur over a long time period with slow exchange between domains. In addition, smaller, more transient domains are present along the plane of the membrane. The roles of these transient lateral domains are likely to be the clustering of molecules in the small volume of a membrane domain to allow for their efficient interaction, which may be particularly important for signal transduction processes in which a series of chemical reactions can lead to amplification of a signal. In addition, the nonuniform distribution of lipid components will result in the physical properties of the membrane varying between a domain and the bulk of the membrane, which, in turn, will modulate the activity of membrane proteins.

Chemistry

We will discuss the chemical composition, size, and lifetime of domains in biological and model membranes. This area is currently under very active investigation, and we are likely to see further developments in the future. We will focus particularly on two types of membrane domains for which there is more evidence for their existence in biological membranes and more characterization of their properties. We can classify...
these domains as either cholesterol-rich domains or anionic lipid clusters.

**Cholesterol-Rich Domains**

Cholesterol is a major lipid component of mammalian cell plasma membranes, accounting for approximately 35% of the total lipid of the membrane. Cholesterol has a chemical structure that is very different from the major polar lipid constituents, which is a consequence of the fused ring system of cholesterol that gives this lipid less conformational flexibility than the straight chains of the polar lipids. It is thus not surprising that cholesterol does not mix well in membranes and that cholesterol segregates as crystals at around 50–60 mol% in bilayers of several lipids and at a much lower molar fraction of cholesterol in bilayers comprising lipids with unsaturated acyl chains (5).

The lateral distribution of molecules in a membrane will be the result of differences in the interaction energies between different groups of molecules. As the chemical structure of cholesterol is very different from that of polar lipids, differences will likely exist in the stability of complexes of peptides or proteins with cholesterol versus those with phospholipids. This property will result in cholesterol being distributed nonuniformly in the membrane at thermodynamic equilibrium. In addition, cholesterol will have different extents of interaction with saturated versus unsaturated polar lipids, a well-studied example of which is in model liposomes is the fluid–fluid immiscibility resulting in the formation of domains in mixtures of dioleoylphosphatidylcholine (DOPC), sphingomyelin, and cholesterol. In these mixtures, a saturated acyl chain of sphingomyelin has more favorable interactions with cholesterol than do the unsaturated acyl chains. If a phosphatidylcholine with saturated acyl chains, dioleoylphosphatidylcholine (DOPC), replaces palmitoyl sphingomyelin (PSM) in mixtures with DOPC and cholesterol, the two phase diagrams are essentially identical (6). The size of the domains formed in such a lipid mixture is of the order of micrometers in diameter (7, 8). One domain is highly enriched in sphingomyelin, whereas the other contains phosphatidylcholine as the major phospholipid. As indicated above, the cholesterol partitioning is preferentially into the more ordered sphingomyelin domain, although the segregation of cholesterol between the domains is only partial (9). Interestingly, substituting DOPC with 1-palmitoyl-2-oleoyl phosphatidylcholine (POPC) into these ternary lipid mixtures with cholesterol results in a large difference between the phase diagrams of POPC/cholesterol/PSM and POPC/cholesterol/DPPC with no miscibility transition observed with the latter mixture (6). In addition to differing in chemical composition, the two immiscible domains have different physical properties. The more ordered domain is the one enriched in sphingomyelin. This phase is still a fluid phase, but it has increased order and has been termed a liquid-ordered phase. Specific properties of the liquid-ordered phase place it between a liquid-disordered phase and an ordered phase. The rate of lateral diffusion of the lipid in the liquid-ordered phase is similar to that in a liquid-disordered phase, making it a liquid phase. In contrast, the acyl chain order parameters are closer to that of an ordered or solid phase, which is a consequence of the neighboring cholesterol inhibiting trans to gauche isomerization of the carbon–carbon single bonds. The coexistence of a liquid-ordered and liquid-disordered domain in a mixture of phospholipids and cholesterol will depend on the chemical nature of the phospholipids, the relative proportion of each of the lipid components, and the temperature. The phase diagram of several three-component systems (two phospholipids and cholesterol) in excess water as a function of temperature has been determined.

Although the characterization of membrane domains that depend on the presence of cholesterol have been well studied and largely understood in model systems, the relevance of this behavior to domains in biological membranes is controversial. However, one type of cholesterol-rich domain exists in the plasma membrane of certain mammalian cells for which there is considerable evidence (10): the caveolae. Caveolae have a lipid composition with an enrichment of cholesterol and sphingomyelin (11) and would therefore be expected to form a liquid-ordered domain that is segregated from the surrounding membrane in a liquid-disordered phase. Caveolae have a distinctive morphological structure of a flask-shaped invagination of approximately 65 nm in diameter that can readily be identified with electron microscopy. This characteristic shape is a consequence of this domain containing a high content of the membrane protein caveolin. Caveolae can be isolated after breaking up the cell membrane, without the use of any detergents (12). Caveolae domains are large, stable, and enriched in cholesterol and sphingomyelin. Their lipid composition and physical state are similar to what is expected for membrane rafts, but the size and lifetime of rafts in biological membranes are much less (13).

The term “raft” has been extensively used in the literature to specify a particular kind of domain in a biological membrane, but its use has often evaded a precise definition. We suggest that a raft be defined as a non-caveolae, cholesterol-rich domain in a biological membrane. As this domain is one with a phase boundary separating it from its surroundings, it would be in a liquid-ordered state. A raft domain would be similar to caveolae except for the distinguishing morphology and presence of the protein caveolin in the latter case. Such a raft domain could be visualized by several imaging methods described below, or its presence could be ascertained with spectroscopic methods. This definition of the physical presence of a cholesterol-rich cluster of molecules in a biological membrane is different from rafts that are defined on the basis of detergent insolubility. The membrane fraction that is insoluble in cold detergent and is in a low density fraction upon ultracentrifugation has been suggested to be similar to raft domains in biological membranes; but it is an extrapolation that is increasingly questioned. This fraction should be referred to as the low density, detergent-resistant fraction and not necessarily representing a raft. Detergent resistance is a phenomenological definition, whereas the term raft should be restricted to a physical structure, a domain, in a membrane.

Several additional features of domains exist in biological membranes that are not present in model systems. Two fundamental features of biological membranes that are not present in model membranes are the presence of proteins and the transbilayer asymmetry. In particular, proteins are a major component of plasma membranes, comprising approximately 50% of
weight of the material. They are also important for domain formation, which is most clear in the case of caveolae, whose presence in cells is closely tied to the expression of caveolin (14). These structures can even be induced to form in cells normally lacking caveolae by inducing the expression of caveolin (15), and caveolae are lost in mice lacking caveolin 1 and 3 (16, 17). However, this example is only one that is specific and well-documented of a protein that affects domain formation. In general, any protein that interacts more favorably with some lipids than with others will result in domain formation. It would be anticipated that, as a consequence of cholesterol increasing the packing density of the membrane, many proteins would be excluded from cholesterol-rich regions (18). The preferential interaction of proteins with cholesterol-depleted domains will result in cholesterol concentrating in a different region of the membrane, resulting in the formation of cholesterol-rich domains. An example of such behavior has been shown with an amphipathic helical peptide (19).

Electrostatically Driven Domain Formation

Another mechanism for domain formation is through the clustering of anionic lipids by cations or by polycationic protein segments. There has been particular interest in the role of proteins with cationic clusters in causing the formation of domains enriched in the polyanionic phospholipid phosphatidylinositol 4,5-bisphosphate (PIP2) (20). In particular, three proteins have been identified that contain cationic clusters, sequester PIP2, and are largely devoid of structure. These proteins are the myristoylated alanine-rich C-kinase substrate (MARCKS), growth-associated protein 43 (GAP43), and neuronal acetyl membrane protein (NAP-22 or CAP23 for the form found in chicken) (21). These proteins have been termed “pipmodulins” because of their ability to modulate the availability of free PIP2. In the case of the MARCKS protein, the electrostatically driven sequestering of PIP2 by the cationic cluster of residues in this protein can be reversed by phosphorylation of Ser residues in this cluster by protein kinase C (22). Another modulation of the sequestering of PIP2 by MARCKS results from the competition between Calcium-calmodulin and the anionic lipid surface for the cluster of basic residues in MARCKS (23, 24).

There has been a suggestion that electrostatically driven formation of PIP2 domains have a relationship to membrane rafts. It is known that several proteins that sequester PIP2 locate in raft domains (25, 26). In addition, the domain localization of PIP2 is dependent on the presence of cholesterol (27-29). However, the conclusion that PIP2 is localized in raft domains (25, 26). In addition, the domain localization of PIP2 is dependent on the presence of cholesterol (27-29). However, the conclusion that PIP2 is localized in raft domains is close to being overturned by the data from the cationic MARCKS cluster (21). These proteins have been termed “pipmodulins” because of their ability to modulate the availability of free PIP2. In the case of the MARCKS protein, the electrostatically driven sequestering of PIP2 by the cationic cluster of residues in this protein can be reversed by phosphorylation of Ser residues in this cluster by protein kinase C (22). Another modulation of the sequestering of PIP2 by MARCKS results from the competition between Calcium-calmodulin and the anionic lipid surface for the cluster of basic residues in MARCKS (23, 24).

Fractionation

Cholesterol-rich domains of biological membranes have been isolated on the basis of their insolubility in 1% Triton at 4°C (36). This method has come into question because of the possibility that the Triton itself causes rearrangement of membrane components (37, 38). As indicated above, this fraction should not be referred to as a raft. However, the cholesterol-rich domain in the form of caveolae can be isolated without use of detergent (39, 40), providing stronger evidence that this cholesterol-rich domain exists in a biological membrane before extraction.

Fluorescence Methods

Fluorescence microscopy has been used extensively to study the formation of membrane domains. It has been particularly successful in determining the phase behavior of model membranes in the form of giant unilamellar vesicles (41). One of the limitations of imaging with light microscopy is that only relatively larger domains can be visualized because the method is limited by the wavelength of light. However, higher resolution can be attained with a near-field scanning optical microscope (NSOM) (42). In addition, recently, the method of stimulated emission depletion has been applied to biological systems that will allow imaging of smaller domains (43). Fluorescent-based methods have also been used to measure the interaction between molecules in small domains of nanometer, rather than micron, size. Fluorescence quenching and fluorescence resonance energy transfer methods have been used to demonstrate the presence of small domains in both model as well as biological membranes (44-46). However, application of these methods to biological membranes has not led to definitive results (45). Recent modeling studies have indicated that the lateral mobility of proteins in membranes is sensitive to the protein dimensions, including protein-lipid interactions (47). Measurements of the lateral diffusion of proteins in the apical plasma membrane of epithelial cells have indicated that certain proteins can be highly mobile (48), whereas others are less mobile (49). These differences in mobility have been correlated with the lateral distribution of proteins within the membrane, providing further evidence for the existence of lipid rafts (50).
cells using fluorescence recovery after photobleaching gives evidence for the presence of at least two coexisting liquid phases (2). A further method for the measurement of lateral diffusion in membranes that holds promise for the future is fluorescence correlation spectroscopy (48).

Atomic Force Microscopy (AFM)

AFM can readily be applied to bilayers deposited on a solid support. The method has the advantage that it can give nanometer resolution and that it is sensitive to measurements of height. It has been used for studying phase separation in lipid membranes (49). One also has to consider possible interactions between the solid support and the bilayer, particularly with regard to measuring rates of lateral diffusion. One can also use AFM in combination with fluorescence microscopy or with total internal reflection fluorescence (TIRF) microscopy to locate labeled molecules in the membrane. However, one must be cautious about the fluorescent probe affecting the domain localization (32). It is also possible to use force measurements in AFM to determine the extent of the electrical double layer (30) and thereby examine the clustering of charged lipids in a bilayer.

Future Directions

Important questions still need to be resolved regarding the nature of the interactions among membrane components that can lead to fluid–fluid immiscibility. However, to extend these model system studies to biological membranes, a deeper understanding of the role of proteins in determining domain formation is required. This task is complicated because there are a large variety of proteins in biological membranes and many motifs by which they can interact with surrounding molecules. Domains can be formed if the proteins interact preferentially with certain membrane components, provided these interactions are sufficiently strong to overcome the entropy of mixing. In addition, a more accurate and complete characterization of domains in biological membranes is required and is an active area of research. These domains are heterogeneous and, therefore, no single description of their size, shape, lifetime, or transbilayer properties exists. Each type of membrane domain will have to be studied in its own right. This complex task will require the coordinated effort of membrane biophysicists, cell biologists, and others. It will likely lead to a better understanding of the coordinated behavior of membrane proteins and the functioning of certain domain-related biological processes such as signal transduction and certain cases of endocytosis and viral fusion.

References

22. Ohnori S, Saka N, Shiral T, Yamanoto K, Miyamoto E, Shimizu N, Saito N. Importance of protein kinase C targeting for the phos-


Further Reading


Further Reading


See Also
Membrane Assembly and Stability
Signal Cascades, Protein Interaction Networks in Lipid Rafts
Synthetic Lipids to Study Biological Function
Essential Fatty Acids, Physiology and Clinical Significance of

Undurti N. Das, UND Life Sciences, Shaker Heights, Ohio
doi: 10.1002/9780470048672.wecb282

Essential fatty acids (EFAs)—linoleic acid (LA) and α-linolenic acid (ALA)—are essential for the brain growth and development of humans. EFAs are readily available in the diet, and hence their deficiency is not common. But, to provide their full benefit, EFAs have to be metabolized to their long-chain metabolites. EFAs form precursors to various prostaglandins (PGs), thromboxanes (TXs), leukotrienes (LTs), lipoxins (LXs), resolvins, neuroprotectins, isoprostanes, and hydroxy- and hydroperoxyeicosa-tetraenoates. Certain PGs, TXs, and LTs have pro-inflammatory actions, whereas LXs, resolvins, and neuroprotectins are anti-inflammatory in nature and are critical for wound healing, the resolution of inflammation, and the repair of tissues. EFAs and some of their long-chain metabolites inhibit the activities of angiotensin-converting and HMG-CoA reductase enzymes and cholesteryl ester transfer protein (CETP), enhance acetylcholine levels in the brain, increase the synthesis of endothelial nitric oxide, augment diuresis, enhance insulin action, and could regulate telomerase activity. Thus, EFAs and their metabolites may function as an endogenous “polypill.” In addition, EFAs and their long-chain metabolites react with nitric oxide (NO) to yield respective nitroalkene derivatives that exert cell-signaling actions via ligation and activation of peroxisome proliferator-activated receptors (PPARs). Thus, EFAs and their derivatives have varied biologic actions that may have relevance to their involvement in several physiologic processes and clinical conditions.

Metabolism of Essential Fatty Acids

EFAs also are polyunsaturated fatty acids (PUFAs) because they contain two or more double bonds. PUFAs are fatty acids, some of which have at least two carbon-to-carbon double bonds in a hydrophobic hydrocarbon chain. At least four independent families of PUFA exist, depending on the parent fatty acid from which they are synthesized. They include:

The “ω-3” series, which is derived from α-linolenic acid (ALA, 18:3, ω-3).

The “ω-6” series, which is derived from cis-linoleic acid (LA, 18:2, ω-6).

The “ω-9” series, which is derived from oleic acid (OA, 18:1, ω-9).

The “ω-7” series, which is derived from palmitoleic acid (PA, 16:1, ω-7).

Essential fatty acids (EFAs) are essential for the survival of humans and other mammals; they cannot be synthesized in the body and, hence, have to be obtained in our diet and, thus, are essential (1–4). EFAs are an important constituent of cell membranes and confer on membranes properties of fluidity; thus, they determine and influence the behavior of membrane-bound enzymes and receptors. Two types of naturally occurring EFAs exist in the body: the ω-6 series derived from linoleic acid (L.A, 18:2) and the ω-3 series derived from α-linolenic acid (ALA, 18:3, ω-3). Both the ω-6 and the ω-3 series are metabolized by the same set of enzymes to their respective long-chain metabolites. Although some functions of EFAs require their conversion to eicosanoids and other products, in most instances the fatty acids themselves are active. The longer-chain metabolites of LA and ALA regulate membrane function and are of major importance in the brain, retina, liver, kidney, adrenal glands, and gonads.

Metabolism of Essential Fatty Acids

EFAs also are polyunsaturated fatty acids (PUFAs) because they contain two or more double bonds. PUFAs are fatty acids, some of which have at least two carbon-to-carbon double bonds in a hydrophobic hydrocarbon chain. At least four independent families of PUFAs exist, depending on the parent fatty acid from which they are synthesized. They include:

The “ω-3” series, which is derived from α-linolenic acid (ALA, 18:3, ω-3).

The “ω-6” series, which is derived from cis-linoleic acid (LA, 18:2, ω-6).

The “ω-9” series, which is derived from oleic acid (OA, 18:1, ω-9).

The “ω-7” series, which is derived from palmitoleic acid (PA, 16:1, ω-7).
LA is converted to γ-linolenic acid (GLA, 18:3, n-6) by the action of the enzyme Δ6 desaturase (d-6-d), and GLA is elongated to form dihomo-GLA (DGLA, 20:3, n-6), the precursor of the 3 series of prostaglandins (PGs). DGLA also can be converted to arachidonic acid (AA, 20:4, n-6) by the action of the enzyme Δ5 desaturase (d-5-d). AA forms the precursor of the 2 series of prostaglandins, thromboxanes, and the 4 series of leukotrienes. ALA is converted to eicosapentaenoic acid (EPA, 20:5, n-3) by d-6-d and d-5-d. EPA forms the precursor of the 3 series of prostaglandins and the 5 series of leukotrienes. LA, GLA, DGLA, AA, ALA, EPA and docosahexaenoic acid (DHA, 22:6, n-3) are all PUFAs, but only LA and ALA are EFAs (see Fig. 1 for the metabolism of EFAs). AA and EPA give rise to their respective hydroxy acids, which, in turn, are converted to their respective leukotrienes (LTs). In addition, AA, EPA, and DHA form the precursor to anti-inflammatory compounds such as lipoxins and resolvins. PGs, LTs, lipoxins (LXs), and resolvins are highly active, modulate inflammation, and are involved in several pathophysiological processes, such as atherosclerosis, bronchial asthma, inflammatory bowel disease, and other inflammatory conditions (5–8). In general, for the sake of brevity, the term "EFAs" is used to refer to all unsaturated fatty acids: LA, GLA, DGLA, AA, ALA, EPA, and DHA; and the term polyunsaturated fatty acids (PUFAs) refers to GLA, DGLA, AA, EPA, and DHA. Although the terms EFAs and PUFAs are used interchangeably for the sake of convenience, it should be understood that all EFAs are PUFAs but all PUFAs are not EFAs. Many functions of EFAs are also brought about by PUFAs, and EFA-deficiency states can be corrected largely by PUFAs; that led to the suggestion that PUFAs are "functional EFAs." Hence, in general, many authors use the terms EFAs and PUFAs interchangeably. This convention is followed in the current discussion also.

EFAs/PUFAs play a significant role in the pathology of clinical conditions such as collagen vascular diseases, hypertension, diabetes mellitus, metabolic syndrome X, pсорiasis, eczema, atopic dermatitis, coronary heart disease, atherosclerosis, and cancer (1–8). This role is in addition to the role of PGs and LTs in these conditions. For instance, in ulcerative colitis, the inflammatory events are initiated and perpetuated by PGs and LTs produced from AA, whereas when significant amounts of EPA and DHA are given, the inflammatory process is abrogated to a large extent. This beneficial action of EPA/DHA when supplemented from external sources has been attributed to the displacement AA from the cell membrane phospholipid pool and formation of less pro-inflammatory PGs and LTs from them, and anti-inflammatory molecules LXs and resolvins and hence the favorable response. If the molecular mechanism(s) by which various stimuli can induce preferentially the release of AA, EPA, and/or DHA and convert them to their respective pro- and anti-inflammatory products, then it is possible to develop methods to treat various inflammatory conditions based on this knowledge. Such an understanding and knowledge is expected to lead to strategies that will help to enhance preferentially the formation of less pro-inflammatory molecules from EPA/DHA and to stimulate the synthesis and release of anti-inflammatory molecules such as LXs and resolvins that could lead to suppression of inflammation and wound healing. LXs and resolvins resolve inflammation by suppressing leukocyte infiltration and clearing cellular debris from the site of inflammation. In view of this action, it is important to know the molecular regulation of their formation and release in various cells and tissues and various diseases that is expected ultimately to develop better methods of managing various inflammatory conditions. Thus, PUFAs form precursors to both pro- and anti-inflammatory molecules, and the balance between these mutually antagonistic compounds could determine the final outcome of the disease process.

**Dietary Sources of EFAs**

The EFAs LA and ALA are present in human diet in abundant amounts, and, hence, EFA deficiency is uncommon. In certain specific conditions, such as total parenteral nutrition (TPN) and severe malabsorption, occasionally EFA deficiency could be seen. The current TPN solutions contain adequate amounts of EFAs. The manifestations of an EFA deficiency include inappropriate water loss through the skin, dehydration, scalp dermatitis, alopecia, and depigmentation of hair (9, 10). EFAs are distributed widely in normal human diet. The main dietary sources of EFAs are as follows.

**Dietary Sources of EFAs**

- **Human breast milk** is rich in all types of PUFAs (11); this explains why breast-fed children are healthier compared with bottle-fed children. LA and ALA are present in significant amounts in dairy products, in organ meats such as liver, and in many vegetable oils such as sunflower, safflower, corn, and...
soy. GLA is present in evening primrose oil at concentrations of 7-14% of total fatty acids, in borage seed oil at 20-27%, and in black currant seed oil at 15-20%. GLA also is found in some fungal sources (1-4, 12). DGLA is found in liver, testes, adrenals, and kidneys. AA is present in meat, egg yolks, some seaweeds, and some shrimps. The average daily intake of AA is estimated to be in the region of 100-200 mg/day, more than enough to account for the total daily production of various PGs. EPA and DHA are present mainly in marine fish. Cow milk contains very small amounts of GLA, DGLA, and AA.

EFA deficiencies are unattractive because of the presence of two or more double bonds in their structure: therefore, essential fatty acid deficiency often occurs during food processing and the hydrogenation of oils. When exposed to high temperatures and during hydrogenation, EPA and DHA are denatured and converted into trans fats that are harmful to the body (13-15). Human diet was rich in ω-3 fatty acids in the early humans. But, with the progress in industrialization and the development of fast foods, the content of ω-3 fatty acids in the human diet dwindled, whereas that of ω-6 fatty acids increased. Thus, the ratio of ω-3 to ω-6 fatty acids in the diet should be about 1 or > 1, whereas this ratio now is believed to be about 3 and around 10:1 to 20:25:1 (16). It is recommended that the ratio between ω-3 to ω-6 fatty acids in the diet of early humans was ω-3, whereas this ratio now is believed to be about 2-3 and preferably 2-3:1. It is believed that the fall in the intake of ω-3 fatty acids EPA and DHA in the last 50 years is responsible for the increasing incidence of modern diseases such as atherosclerosis, CHD, hypertension, metabolic syndrome X, obesity, collagen vascular diseases, and, possibly, cancer. This reason explains why trans fats, saturated fats, and cholesterol have pro-inflammatory actions, whereas EFAs and PUFA s possess anti-inflammatory properties. This reason explains why trans fats, saturated fats, and cholesterol are more likely to inhibit the formation of LXs, resolvins, PGI2 (prostacyclin), and other beneficial eicosanoids that prevent platelet aggregation, leukocyte chemotaxis, and activation of pro-inflammatory cytokines. In contrast, trans fats, saturated fats, and cholesterol may activate leukocytes directly, induce the generation of free radicals, and enhance the production and release of pro-inflammatory cytokines that facilitate atherosclerosis (14, 15, 25-29). The direct action of trans fats, saturated fats, and cholesterol interferes with EFA metabolism and promotes inflammation, atherosclerosis, and coronary heart disease (14, 15, 23, 24). Thus, trans fats, saturated fats, and cholesterol have been described as being anti-inflammatory, whereas EFAs and PUFA s possess anti-inflammatory properties.

Based on these studies, it is proposed that EFAs, especially EPA and DHA, from dietary LA and ALA, and AA. This interference, in turn, could lead to the decreased formation of LXs, resolvins, PGI2, and PGI3. It is likely that trans fats, saturated fats, and cholesterol may affect the formation of LXs, resolvins, PGI2, and PGI3. It is likely that trans fats, saturated fats, and cholesterol have pro-inflammatory actions. In contrast, trans fats, saturated fats, and cholesterol may interfere with EFA metabolism and promote inflammation, atherosclerosis, and coronary heart disease (14, 15, 23, 24). Thus, trans fats, saturated fats, and cholesterol have been described as being anti-inflammatory, whereas EFAs and PUFA s possess anti-inflammatory properties.
free radical-induced damage. This finding implies that for endothelial cells to be healthy, they need adequate amounts of EFAs, especially AA, EPA, and DHA so that they can generate physiologic amounts of eNO not only to prevent pathologic platelet aggregation and atherosclerosis but also to protect themselves from the cytotoxic actions of free radicals. Furthermore, NO reacts with PUFA s to yield their respective nitroalkene derivatives that can be detected in plasma. These nitroalkene derivatives, termed nitrolipids, produce vascular relaxation, inhibit neutrophil degranulation and superoxide formation, inhibit platelet activation, and show anti-atherosclerotic properties (3, 4, 31–34). Thus, a close interaction seems to exist between EFAs and their products and trans fats, saturated fats, and cholesterol with regard to the ability of endothelial cells to produce PG1, PG2, NO, and other anti-atherosclerotic and beneficial molecules.

**Actions of EFAs/PUFAs Relevant to Various Clinical Conditions**

**Cell membrane fluidity and modification of receptor properties**

Cell membrane fluidity is determined by its lipid composition. The cell membrane is rendered more rigid if increased amounts of saturated fatty acids and cholesterol are incorporated into the membrane phospholipids. In contrast, increased incorporation of unsaturated fatty acids into the membrane will make it more fluid. This increase in membrane fluidity has been shown to increase the number of receptors and their affinity to their respective hormones or growth factors. For instance, an increase in the rigidity of the cell membrane reduces the number of insulin receptors and their affinity to insulin that, in turn, could cause insulin resistance. Alternatively, an increase in cell membrane fluidity because of an increase in the unsaturated fatty acid content in the membrane phospholipids increases the number of insulin receptors and their affinity to insulin and, thus, decreases insulin resistance (35–42).

This action of EFAs/PUFAs has relevance to the growth and development of the brain during the perinatal and postnatal periods, periods of brain growth, is important for proper development of neurons, synapse formation, and cognitive function. The growth and development of the brain during the perinatal period and adolescence is dependent on the availability of ω-3 and ω-6 fatty acids (43–45). Hence, decreased availability of ω-3 and ω-6 fatty acids during this critical period of growth may impair brain growth and the development of appropriate synaptic connections that, in turn, may lead to the development of neuropsychologic conditions such as dementia, depression, schizophrenia, Alzheimer's disease, and neurodegenerative diseases (Huntington's disease, Parkinson's disease, and spinocerebellar degeneration). The alterations in the neuronal cell membrane fluidity also might influence the cognitive function. This suggestion is supported by the recent observation that EPA/PUFAs activate syntaxin, a plasma membrane protein that has an important role in the growth of neurites. For proper neuronal development and increase in cell membrane surface area, the growth of neurite processes from the cell body is critical (46). Nerve growth cones are highly enriched with AA-releasing phospholipases, which have been implicated in neurite outgrowth (47, 48). Cell membrane expansion occurs through the fusion of transport organelles with plasma membrane, and syntaxin 3, a plasma membrane protein that plays a significant role in the growth of neurites, has been shown to be a direct target for AA, DHA, and other PUFAs (49). AA, DHA, and other PUFAs but not saturated and monounsaturated fatty acids activate syntaxin 3. Of all the fatty acids tested, AA and DHA were found to be the most potent compared with LA and ALA. Even syntaxin 3 that is involved specifically in fast calcium-triggered exocytosis of neurotransmitters and in neurite outgrowth. It is interesting that SNAP25 (synaptosomal-associated protein of 25 kDa), a syntaxin partner implicated in neurite outgrowth, interacted with syntaxin 3 only in the presence of AA that allowed the formation of the binary syntaxin 3-SNAP 25 complex. AA stimulated syntaxin 3 to form the ternary SNARE (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) complex, which is needed for the fusion of plasmalemmal precursor vesicles into the cell surface membrane that leads to membrane fusion. These results clearly demonstrated that AA and DHA change the ω-helical syntaxin structure to expose the SNARE motif for immediate SNAP 25 engagement and, thus, facilitate neurite outgrowth. Hence, the availability of optimal amounts of EFAs/PUFAs during the perinatal and postnatal periods, periods of brain growth, is important for proper development of neurons, synapse formation, and cognitive function.

**Second Messenger Actions of EFAs and their Metabolites**

Several hormones and growth factors activate phospholipase A2 (PLA2) that, in turn, induces the release of DGLA, AA, EPA, and DHA from the cell membrane lipid pool. These fatty acids are used for the formation of various eicosanoids and bring about their actions. The inhibition of PLA2 interferes with the action of several growth factors and cytokines. For example, the tumoroidal action of TNF-α is dependent on its ability to induce PLA2, and inhibitors of PLA2 completely blocked its TNF-α antitumor action (51–54). Activation of PLA2 leads to the release of various PUFAs including AA. PUFAs enhance the activity of protein kinase C (PKC), a well-known second messenger, activate macrophages and polymorphonuclear leukocytes (PMNs), and increase free radical generation (1–4, 53, 56). The effects of various types of PUFAs on the activation of macrophages and PMNs and their ability to generate free radicals depends on the type of EFA s/PUFAs used, on the degree of unsaturation, concentration(s) of the fatty acid, and on the conditions under which the experiments are performed. Thus, the effects of various PUFAs on macrophages and PMNs are complex and have to be interpreted with caution taking into consideration all of the variables involved.
Anti-Inf ective Properties of EFAs

Previously, I proposed that EFA s/PUFAs such as GLA, DGLA, AA, EPA, and DHA might behave as endogenous antibacterials, antifungal, antiviral, and immunomodulating agents (57, 58). LA and ALA have bacteriostatic effect on both gram-positive and gram-negative bacteria (58, 59). Staphylococcus aureus are killed rapidly by ALA, and a variety of bacteria were found to be sensitive to the growth inhibitory actions of LA and ALA in vitro (60, 61). Hydrolyzed linseed oil, which contains 52% ALA, and pure ALA were found to be capable of killing methicillin-resistant S. aureus (62). A nimal herpes, influenza, Sendai, and Sindbis viruses could be inactivated by LA and AA within minutes of contact (63). Furthermore, both prostaglandin E1 (PGE1) and PGA, derived from DGLA, AA, and EPA, have the ability to inhibit viral replication and behave as antiviral compounds (64, 65). It was reported that the remission of myosis fungoides, a rare skin disease of viral etiology, is possible with the oral administration of LA as safflower oil (which contains 76% LA) in dogs that correlated with an increase in the plasma levels of LA and AA (66).

Patients with Plasmodium falciparum showed suppressed lympho-proliferative responses to malaria antigens. This suppressed lymphocyte response has been attributed to the increased production of various prostaglandins (PGs) by monocytes/macrophages. This finding is supported by the observation that enhanced lymphocyte responses to several malaria antigens were enhanced particularly by indomethacin, a cyclo-oxygenase inhibitor, which suggests that malaria-specific T-cells are especially sensitive to the effects of PGs (67). In a related study, it was noted that during intra-erythrocytic development of the parasite, the phospholipid composition of the IEPM (infected erythrocyte plasma membrane) contained more phosphatidylcholine (38.7% versus 31.7%) and phosphatidylinositol (21.2% versus 0.8%) and less sphingomyelin (14.6% versus 28.0%) than normal uninfected erythrocytes. The percentage of PUFAs in normal erythrocyte phospholipids (39.4%) was much higher than in phospholipids from purified parasites (23.3%) or IEPM (24.1%). Large increases in palmitic acid (from 21.88% to 31.21%) and in oleic acid (from 14.64% to 24.60%) and significant decreases in AA (from 17.36% to 7.85%) and in DHA (from 4.34% to 1.89%) occurred because of the infection. The fatty acid profiles of individual phospholipid classes from IEPM resembled the fatty acid profiles of parasite phospholipids rather than those of uninfected erythrocytes, which suggests that these alterations are because of the parasite-directed metabolism of erythrocyte lipids (68). These results are interesting because it has been shown that iloprost, a synthetic prostacyclin (PGI2) analog, not only prevented the development of cerebral malaria in mice but also suppressed malaria antigen-induced tumor necrosis factor (TNF) production (69). PGI2 and PGF2α are derived from AA and EPA, respectively. Furthermore, AA, LA, EPA, and DHA inhibited the growth of P. falciparum in vitro, whereas oleic or docosanoic acids were ineffective. Parasite killing was increased significantly when oxidized forms of LA, AA, EPA, and DHA were used. Antioxidants greatly reduced the fatty acid-induced killing. Mice infected with P. berghei and treated for 4 days with DHA showed marked reduction in parasitemia (70). In this context, it is interesting to note that the cytotoxic effect of the fatty acids was very rapid: The full inhibition of nucleic acids and protein syntheses was observed in less than 30 minutes without showing any effects on RBCs such as the hemolysis of infected cells up to 500 µg/ml; also no effect on the lipid peroxidation, ATP levels, transport through the parasite-induced permeability pathways, or on the phagocytosis of the infected cells was noted (71). Furthermore, mice fed vitamin E-deficient diets that contain ω-3 fatty acids survive infection with lethal P. yoelii. Experiments performed in nu/nu mice (which lack alpha-beta T-cell-receptor-positive T cells and do not produce antimalarial antibody) and nu/nu mice revealed that animals fed casein-based diets that contain 4% menhaden oil with vitamin E supplementation for 4 weeks before infection with lethal P. yoelii developed fulminating parasitemias and quickly died, whereas both nu/nu and nu/nu mice fed diets deficient in vitamin E controlled their parasitemias for the first 18 days of infection. Therefore, the nu/nu mice become anemic and died, whereas the nu/nu mice produced antimalarial antibodies and survived. In an extension of this study, when scid/scid/bg/bg mice (which lack B cells and alpha-beta and gamma-delta T cells and have reduced NK-cell activity) were fed the experimental diet for 6 weeks and then infected with the less-virulent J774.A1 strain of P. yoelii, the animals fed vitamin E-containing diets quickly died, whereas those fed the vitamin E-deficient diet survived without developing detectable parasitemias. These results suggest that under pro-oxidant dietary conditions, mice could control and even survive malaria in the absence of malaria-primed T cells and the antimalarial antibody (72), which implies that fatty acids need to be oxidized to produce their antimalarial action and eliminate malaria parasite. Our studies showed that in patients with P. falciparum malaria, the levels of lipid peroxides (a marker of free radical generation), nitric oxide (a potent free radical with immunomodulatory actions), and concentrations of LA and ALA are low, whereas those of EPA are high. Because the ability of the fatty acids to kill P. falciparum is dependent on their capacity to stimulate free radical generation in neutrophils and macrophages and EPA is more potent than LA in killing the parasite, these results imply that decreased levels of lipid peroxides and nitric oxide may contribute to the susceptibility of P. falciparum infection (73). In addition, it was reported that mice infected with P. berghei ANKA (which developed cerebral malaria) showed increased phosphatidylase A2 mRNA expression in the spleen and cyclooxygenase 1 (COX1) and COX2 expression in the brain and had higher serum LTβ intact levels than the control mice; also it was reported that aspirin-treated infected mice had higher serum LTβ and lower serum unmetabolized infected mice, which suggests that PGs are protective whereas LTs are detrimental in cerebral malaria (74). These results, coupled with the observations that elevated PGE2 in healthy malaria-exposed children may protect against malaria (75), that some PGs actually may reverse chloroquine-resistance of the malarial parasite (76), and that increased production of TXs may be involved in the pathogenesis of some complications associated with malaria suggest that EFA s/PUFAs and their eicosanoid metabolites play a significant role in the protection and the pathogenesis of malaria.
These observations suggest that EFAs/PUFAs and their products have antibacterial, antifungal, antiviral, and antiparasitic actions. Both lymphocytes and macrophages contain significant amounts of PUFAs and are capable of releasing them on appropriate stimulation. In addition, PUFAs stimulate NADPH-dependent superoxide production by macrophages, neutrophils, and lymphocytes that are capable of killing the invading microorganisms. In view of these evidences, it is reasonable to believe that an increased intake of LA, ALA, EPA, and DHA protects against and/or reduces the risk of various infections. Recent studies showed that AA, EPA, and DHA could give rise to anti-inflammatory compounds such as lipoxins (LXs) and resolvins that are essential to limit and resolve inflammation (3, 4). These studies imply that a deficiency of LXs and resolvins could lead to the perpetuation of inflammation and tissue damage. In the light of these facts, it will be interesting to study whether a subclinical deficiency of PUFAs, decreased formation of LXs and resolvins, occurs in subjects who develop the various types of infections and their complications. Because, PUFAs can inactivate enveloped viruses including influenza, it is probably worthwhile to study the effect of various fatty acids on the bird flu virus, specifically, whether increased intake of these fatty acids could reduce the risk of flu.

It is interesting to note that an analog of myristic acid (14:0) showed selective toxicity to African Trypanosomes (77), PUFAs also can kill Helicobacter pylori (78), and the actions of some antibiotics could be potentiated by PUFAs (79), which lends support to the concept that PUFAs may function as endogenous anti-infective-like molecules (80, 81).

**Actions that Could Qualify Pufas to be an Endogenous “Polypill”**

Coronary heart disease (CHD), stroke, peripheral vascular disease, and underlying atherosclerosis are the common cardiovascular diseases (CVD) that are responsible for considerable morbidity and mortality both in the developed and developing countries. Several studies revealed that smoking cessation, β-blockers, antipalpebral agents in the form of low-dose aspirin and other drugs, inhibitors of angiotensin-converting enzyme (ACE), and lipid-lowering agents such as statins that inhibit the activity of HMG-CoA reductase, each reduce the risk of vascular events to a moderate but important degree (82–90). In addition, observational studies suggested lower rates of fractures with statins, higher rates of obstructive airways disease at lower cholesterol concentrations, lower rates of cataracts, and lower rates of dementia with MRC/BHF-HPS study (90). The results of the MRC/BHF-HPS study led to the suggestion that about one-third to three-quarters of future vascular events could be prevented (see Table 1). This suggestion indicates that a combination of aspirin, β-blockers, lipid-lowering (by 1–5 mmol) and the use of statins, and ACE inhibitors could reduce cardiovascular diseases by over 70–80% (91). This concept led to the suggestion that a combination pill (popularly called a “polypill”) consisting of atorvastatin 10 mg or simvastatin 40 mg; and aspirin 75 mg could reduce CHD events by 88% (95% confidence interval 84% to 91%) and stroke by 80% (71% to 87%) (see Table 1 for a possible formulation of one such polypill). It was suggested that one third of the people taking such a combination pill from age 55 years of age would benefit, gaining on an average about 13 years of life free from an CHD event or stroke (92).

**Pufas Inhibit ACE and HMG-CoA Reductase Activities and Augment Endothelial Nitric Oxide Generation**

PUFAs inhibited leukocyte ACE activity (93) and enhanced endothelial nitric oxide (eNOS) generation (94). This finding implies that when tissue concentrations of PUFAs are low, the activity of ACE will be high, which results in increased formation of angiotensin-II and a simultaneous decrease in eNO. It was reported that transgenic rats overexpressing human renin and angiotensinogen genes (dTGR) that develop hypertension, inflammation, and renal failure showed specific renal P450-dependent AA metabolism changes that led to decreased formation of epoxy-eicosatrienoic acids (5,6-, 8,9-, 11,12- and 14,15-EETs) and hydroxyeicosatetraenoic acids (19- and 20-HETEs). Both EETs and HETEs inhibit IL-6 and TNF-α-induced activation of endothelial nitric oxide synthase (eNOS). These observations suggest that EFAs/PUFAs and their products have antibacterial, antifungal, antiviral, and antiparasitic actions. Both lymphocytes and macrophages contain significant amounts of PUFAs and are capable of releasing them on appropriate stimulation. In addition, PUFAs stimulate NADPH-dependent superoxide production by macrophages, neutrophils, and lymphocytes that are capable of killing the invading microorganisms. In view of these evidences, it is reasonable to believe that an increased intake of LA, ALA, EPA, and DHA protects against and/or reduces the risk of various infections. Recent studies showed that AA, EPA, and DHA could give rise to anti-inflammatory compounds such as lipoxins (LXs) and resolvins that are essential to limit and resolve inflammation (3, 4). These studies imply that a deficiency of LXs and resolvins could lead to the perpetuation of inflammation and tissue damage. In the light of these facts, it will be interesting to study whether a subclinical deficiency of PUFAs, decreased formation of LXs and resolvins, occurs in subjects who develop the various types of infections and their complications. Because, PUFAs can inactivate enveloped viruses including influenza, it is probably worthwhile to study the effect of various fatty acids on the bird flu virus, specifically, whether increased intake of these fatty acids could reduce the risk of flu.

It is interesting to note that an analog of myristic acid (14:0) showed selective toxicity to African Trypanosomes (77), PUFAs also can kill Helicobacter pylori (78), and the actions of some antibiotics could be potentiated by PUFAs (79), which lends support to the concept that PUFAs may function as endogenous anti-infective-like molecules (80, 81).

**Actions that Could Qualify Pufas to be an Endogenous “Polypill”**

Coronary heart disease (CHD), stroke, peripheral vascular disease, and underlying atherosclerosis are the common cardiovascular diseases (CVD) that are responsible for considerable morbidity and mortality both in the developed and developing countries. Several studies revealed that smoking cessation, β-blockers, antipalpebral agents in the form of low-dose aspirin and other drugs, inhibitors of angiotensin-converting enzyme (ACE), and lipid-lowering agents such as statins that inhibit the activity of HMG-CoA reductase, each reduce the risk of vascular events to a moderate but important degree (82–90). In addition, observational studies suggested lower rates of fractures with statins, higher rates of obstructive airways disease at lower cholesterol concentrations, lower rates of cataracts, and lower rates of dementia with MRC/BHF-HPS study (90). The results of the MRC/BHF-HPS study led to the suggestion that about one-third to three-quarters of future vascular events could be prevented (see Table 1). This suggestion indicates that a combination of aspirin, β-blockers, lipid-lowering (by 1–5 mmol) and the use of statins, and ACE inhibitors could reduce cardiovascular diseases by over 70–80% (91). This concept led to the suggestion that a combination pill (popularly called a “polypill”) consisting of atorvastatin 10 mg or simvastatin 40 mg; and aspirin 75 mg could reduce CHD events by 88% (95% confidence interval 84% to 91%) and stroke by 80% (71% to 87%) (see Table 1 for a possible formulation of one such polypill). It was suggested that one third of the people taking such a combination pill from age 55 years of age would benefit, gaining on an average about 13 years of life free from an CHD event or stroke (92).

**Pufas Inhibit ACE and HMG-CoA Reductase Activities and Augment Endothelial Nitric Oxide Generation**

PUFAs inhibited leukocyte ACE activity (93) and enhanced endothelial nitric oxide (eNOS) generation (94). This finding implies that when tissue concentrations of PUFAs are low, the activity of ACE will be high, which results in increased formation of angiotensin-II and a simultaneous decrease in eNO. It was reported that transgenic rats overexpressing human renin and angiotensinogen genes (dTGR) that develop hypertension, inflammation, and renal failure showed specific renal P450-dependent AA metabolism changes that led to decreased formation of epoxy-eicosatrienoic acids (5,6-, 8,9-, 11,12- and 14,15-EETs) and hydroxyeicosatetraenoic acids (19- and 20-HETEs). Both EETs and HETEs inhibit IL-6 and TNF-α-induced activation of endothelial nitric oxide synthase (eNOS). These observations suggest that EFAs/PUFAs and their products have antibacterial, antifungal, antiviral, and antiparasitic actions. Both lymphocytes and macrophages contain significant amounts of PUFAs and are capable of releasing them on appropriate stimulation. In addition, PUFAs stimulate NADPH-dependent superoxide production by macrophages, neutrophils, and lymphocytes that are capable of killing the invading microorganisms. In view of these evidences, it is reasonable to believe that an increased intake of LA, ALA, EPA, and DHA protects against and/or reduces the risk of various infections.

Recent studies showed that AA, EPA, and DHA could give rise to anti-inflammatory compounds such as lipoxins (LXs) and resolvins that are essential to limit and resolve inflammation (3, 4). These studies imply that a deficiency of LXs and resolvins could lead to the perpetuation of inflammation and tissue damage. In the light of these facts, it will be interesting to study whether a subclinical deficiency of PUFAs, decreased formation of LXs and resolvins, occurs in subjects who develop the various types of infections and their complications. Because, PUFAs can inactivate enveloped viruses including influenza, it is probably worthwhile to study the effect of various fatty acids on the bird flu virus, specifically, whether increased intake of these fatty acids could reduce the risk of flu.

It is interesting to note that an analog of myristic acid (14:0) showed selective toxicity to African Trypanosomes (77), PUFAs also can kill Helicobacter pylori (78), and the actions of some antibiotics could be potentiated by PUFAs (79), which lends support to the concept that PUFAs may function as endogenous anti-infective-like molecules (80, 81).

**Actions that Could Qualify Pufas to be an Endogenous “Polypill”**

Coronary heart disease (CHD), stroke, peripheral vascular disease, and underlying atherosclerosis are the common cardiovascular diseases (CVD) that are responsible for considerable morbidity and mortality both in the developed and developing countries. Several studies revealed that smoking cessation, β-blockers, antipalpebral agents in the form of low-dose aspirin and other drugs, inhibitors of angiotensin-converting enzyme (ACE), and lipid-lowering agents such as statins that inhibit the activity of HMG-CoA reductase, each reduce the risk of vascular events to a moderate but important degree (82–90). In addition, observational studies suggested lower rates of fractures with statins, higher rates of obstructive airways disease at lower cholesterol concentrations, lower rates of cataracts, and lower rates of dementia with MRC/BHF-HPS study (90). The results of the MRC/BHF-HPS study led to the suggestion that about two-thirds to three-quarters of future vascular events could be prevented (see Table 1). This suggestion indicates that a combination of aspirin, β-blockers, lipid-lowering (by 1–5 mmol) and the use of statins, and ACE inhibitors could reduce cardiovascular diseases by over 70–80% (91). This concept led to the suggestion that a combination pill (popularly called a “polypill”) consisting of atorvastatin 10 mg or simvastatin 40 mg; three blood pressure lowering drugs such as a thiazide, a
NF-κB and prevent vascular inflammation (95), which suggests that AA and other PUFAs not only regulate ACE activity and Ang-II levels in the tissues but also possess anti-inflammatory properties by generating anti-inflammatory metabolites.

In the presence of aspirin, AA, EPA, and DHA are converted to form epi-lipoxins, lipoxins, and resolvins that, in turn, enhance the formation of eNO (3, 4, 96). Lipoxins possess potent anti-inflammatory actions (reviewed in References 3 and 4). In addition, NO not only blocks the interaction between leukocytes and the vascular endothelium during inflammation but also stimulates the formation of PGi2, a potent vasodilator and platelet anti-aggregator, from AA (97, 98).

PUFAs are potent inhibitors of the HMG-CoA reductase enzyme and similar to statins are useful in the treatment of hyperlipidemias (99–102). Statins enhance plasma AA levels and decrease the ratio of EPA to AA significantly (100). This finding suggests that PUFAs mediate many actions of statins (103) and that this could be one mechanism by which they lower cholesterol levels. Statins and PUFAs have many overlap actions such as the inhibition of IL-6 and TNF-α production and NF-κB activation plus the ability to enhance eNO production; thus, both possess anti-inflammatory actions and both are useful in atherosclerosis, coronary heart disease, osteoporosis, stroke, Alzheimer’s disease, and inflammatory conditions such as lupus and cancer (3, 4, 94, 104–121). These similar and overlap actions strongly indicate that the molecular mechanisms of actions of statins and PUFAs are similar, if not identical. Furthermore, when a combination of statins and PUFAs are given together, a synergistic beneficial effect was seen in patients with combined hyperlipemia (122).

Both PUFAs and statins by inhibiting the HMG-CoA reductase enzyme reduce the formation of isoprenoid that contains compounds that are formed from mevalonate (because the HMG-CoA reductase enzyme reduces the formation of mevalonate). These isoprenoid precursors are necessary for the posttranslational lipid modification (prenylation) and, hence, the function of Ras and other small GTPases. Small GTPases, the prenylated products of the mevalonate pathway, have negative regulatory control on the expression of BMPs (bone morphogenetic proteins). BMPs are essential for neuronal growth, proliferation, and differentiation (123) and also for bone growth (105, 114). Thus, PUFAs modulate brain growth and development and neuronal differentiation. This finding explains why PUFAs are useful in the prevention and treatment of dementia and Alzheimer’s disease (115–120). Similar to PUFAs, statins also enhance the concentrations of BMPs in brain and bone and thus could be of benefit in the treatment of Alzheimer’s disease and osteoporosis (105, 109, 113). Yet another action of PUFAs and statins that contributes to their beneficial actions is their ability to enhance eNO (30, 124), a pleiotropic molecule that has many biologic actions including its ability to function as a neurotransmitter (125) and prevent osteoporosis (126, 127). But unlike statins that cannot be given during pregnancy, PUFAs can be consumed confidently during pregnancy, lactation, and infancy. In fact, PUFAs are recommended during pregnancy, lactation, and infancy to improve brain growth and development (1–4, 43, 44).

**Pufas and Cholesteryl Ester Transfer Protein (CETP) Activity**

Several studies suggested that HDL-cholesterol (high-density lipoprotein-cholesterol, HDL-C) is an independent risk factor for CHD. Higher plasma LDL-C is associated with a decreased incidence of CHD (128). This finding led to the suggestion that therapeutic strategies that raise HDL-C could be of benefit in preventing CHD with the hope that raising HDL-C will increase the movement of cholesterol from the periphery back to the liver (the so-called reverse cholesterol transport or RCT pathway) and protection from CHD will follow.

CETP is a hydrophobic plasma glycoprotein, mainly synthesized in the liver, that possesses the unique ability to facilitate the transfer of cholesteryl ester (CE). CETP circulates in the blood, bound predominantly to HDL. CETP mediates the transfer of cholesteryl esters from HDL to VLDL and LDL in exchange for triglycerides. CETP also promotes the transformation of HDL2 to HDL3, an action that could promote reverse cholesterol transport. CETP inhibition produces an increase in HDL by markedly delaying the catabolism of apoA-I and A-II (129), an action that increases reverse cholesterol transport. These actions of CETP suggest that CETP inhibition could prevent atherosclerosis (130–132).

In a study performed in healthy, normolipidemic men, it was observed that a lipid-lowering diet rich in monounsaturated fatty acid (oleic acid) decreased CETP concentrations to a significant degree (133). In HepG2 cells, it was noted that 0.5 mM of AA, EPA, and DHA reduced the levels of CETP mRNA by more than 50% of the control levels with a corresponding significant decrease in the CETP mass (134). This finding is supported by the observation that in type 2 diabetic subjects, CETP activity was correlated significantly with the HDL-C to apoA-I ratio and to the LDL-C to HDL-C ratio. In addition, a significant negative correlation was found between plasma CETP activity and monounsaturated fatty acid content of plasma phospholipids or free PUFAs, especially with ω-3 fatty acids, which suggests that PUFAs suppress CETP activity (135).

Torcetrapib, a small molecule inhibitor of CETP, is very effective at raising HDL-C and apolipoprotein A-I and decreasing levels of LDL-C and triglycerides. In fact, it was observed that the administration of torcetrapib with atorvastatin led to the initiation of clinical trials. In patients with familial hypercholesterolemia, torcetrapib with atorvastatin as compared with atorvastatin alone did not result in the reduction of the progression of atherosclerosis as measured by carotid arterial-wall thickness, despite a significant increase in HDL-C levels and a decrease in levels of LDL-C and triglycerides. In another study that involved the use of torcetrapib alone or in combination with other statins, torcetrapib did not reduce CHD. Higher plasma HDL-C is associated with a decreased incidence of CHD (128). This finding led to the suggestion that therapeutic strategies that raise HDL-C could be of benefit in preventing CHD with the hope that raising HDL-C will increase the movement of cholesterol from the periphery back to the liver (the so-called reverse cholesterol transport or RCT pathway) and protection from CHD will follow.

CETP is a hydrophobic plasma glycoprotein, mainly synthesized in the liver, that possesses the unique ability to facilitate the transfer of cholesteryl ester (CE). CETP circulates in the blood, bound predominantly to HDL. CETP mediates the transfer of cholesteryl esters from HDL to VLDL and LDL in exchange for triglycerides. CETP also promotes the transformation of HDL2 to HDL3, an action that could promote reverse cholesterol transport. CETP inhibition produces an increase in HDL by markedly delaying the catabolism of apoA-I and A-II (129), an action that increases reverse cholesterol transport. These actions of CETP suggest that CETP inhibition could prevent atherosclerosis (130–132).

In a study performed in healthy, normolipidemic men, it was observed that a lipid-lowering diet rich in monounsaturated fatty acid (oleic acid) decreased CETP concentrations to a significant degree (133). In HepG2 cells, it was noted that 0.5 mM of AA, EPA, and DHA reduced the levels of CETP mRNA by more than 50% of the control levels with a corresponding significant decrease in the CETP mass (134). This finding is supported by the observation that in type 2 diabetic subjects, CETP activity was correlated significantly with the HDL-C to apoA-I ratio and to the LDL-C to HDL-C ratio. In addition, a significant negative correlation was found between plasma CETP activity and monounsaturated fatty acid content of plasma phospholipids or free PUFAs, especially with ω-3 fatty acids, which suggests that PUFAs suppress CETP activity (135).

Torcetrapib, a small molecule inhibitor of CETP, is very effective at raising HDL-C and apolipoprotein A-I and decreasing levels of LDL-C and triglycerides. In fact, it was observed that the administration of torcetrapib with atorvastatin led to the initiation of clinical trials. In patients with familial hypercholesterolemia, torcetrapib with atorvastatin as compared with atorvastatin alone did not result in the reduction of the progression of atherosclerosis as measured by carotid arterial-wall thickness, despite a significant increase in HDL-C levels and a decrease in levels of LDL-C and triglycerides. In another study that involved the use of torcetrapib alone or in combination with other statins, torcetrapib did not reduce CHD. Higher plasma HDL-C is associated with a decreased incidence of CHD (128). This finding led to the suggestion that therapeutic strategies that raise HDL-C could be of benefit in preventing CHD with the hope that raising HDL-C will increase the movement of cholesterol from the periphery back to the liver (the so-called reverse cholesterol transport or RCT pathway) and protection from CHD will follow.

CETP is a hydrophobic plasma glycoprotein, mainly synthesized in the liver, that possesses the unique ability to facilitate the transfer of cholesteryl ester (CE). CETP circulates in the blood, bound predominantly to HDL. CETP mediates the transfer of cholesteryl esters from HDL to VLDL and LDL in exchange for triglycerides. CETP also promotes the transformation of HDL2 to HDL3, an action that could promote reverse cholesterol transport. CETP inhibition produces an increase in HDL by markedly delaying the catabolism of apoA-I and A-II (129), an action that increases reverse cholesterol transport. These actions of CETP suggest that CETP inhibition could prevent atherosclerosis (130–132).

In a study performed in healthy, normolipidemic men, it was observed that a lipid-lowering diet rich in monounsaturated fatty acid (oleic acid) decreased CETP concentrations to a significant degree (133). In HepG2 cells, it was noted that 0.5 mM of AA, EPA, and DHA reduced the levels of CETP mRNA by more than 50% of the control levels with a corresponding significant decrease in the CETP mass (134). This finding is supported by the observation that in type 2 diabetic subjects, CETP activity was correlated significantly with the HDL-C to apoA-I ratio and to the LDL-C to HDL-C ratio. In addition, a significant negative correlation was found between plasma CETP activity and monounsaturated fatty acid content of plasma phospholipids or free PUFAs, especially with ω-3 fatty acids, which suggests that PUFAs suppress CETP activity (135).

Torcetrapib, a small molecule inhibitor of CETP, is very effective at raising HDL-C and apolipoprotein A-I and decreasing levels of LDL-C and triglycerides. In fact, it was observed that the administration of torcetrapib with atorvastatin led to the initiation of clinical trials. In patients with familial hypercholesterolemia, torcetrapib with atorvastatin as compared with atorvastatin alone did not result in the reduction of the progression of atherosclerosis as measured by carotid arterial-wall thickness, despite a significant increase in HDL-C levels and a decrease in levels of LDL-C and triglycerides. In another study that involved the use of torcetrapib alone or in combination with other statins, torcetrapib did not reduce CHD. Higher plasma HDL-C is associated with a decreased incidence of CHD (128). This finding led to the suggestion that therapeutic strategies that raise HDL-C could be of benefit in preventing CHD with the hope that raising HDL-C will increase the movement of cholesterol from the periphery back to the liver (the so-called reverse cholesterol transport or RCT pathway) and protection from CHD will follow.
with atorvastatin, an increase in blood pressure with no significant decrease in the progression of coronary atherosclerosis was noted (138).

These results with torcetrapib and atorvastatin suggest that the simultaneous inhibition of CETP and the HMG-CoA reductase enzyme leads to an elevation of plasma HDL-C and a decrease in LDL-C, triglycerides, and cholesterol but that it does not arrest the progression of atherosclerosis. In contrast, PUFAs, especially ω-3 EPA and DHA, not only inhibit CETP and the HMG-CoA reductase enzyme and lower plasma triglycerides, cholesterol, and LDL-C with little or no change in HDL-C but also are effective in arresting atherosclerosis and preventing CHD (139-151). In contrast to the results with torcetrapib and atorvastatin, Yokoyama et al. (152) reported that a combination of ethyl EPA and 10 mg of pravastatin or 5 mg of simvastatin prevented major coronary events and especially nonfatal coronary events in Japanese hypercholesterolemic patients with a mean period of follow up of 4.6 years. It is interesting to note that the benefits were in addition to statin treatment and that fish oil was found to be safe and well tolerated. These results once again confirm that EPA and DHA are of benefit in the prevention and treatment of cardiovascular diseases. Thus, PUFAs seem to be superior to CETP and statins in the prevention of CVD despite the fact that they do not necessarily increase plasma HDL-C levels.

Effects on Platelets and Other Hemostatic Indices

Both EPA and DHA, when given orally, are incorporated rapidly into platelets and compete with AA for the 2-acyl position of membrane phospholipid and as a substrate for the cyclooxygenase (CO) and lipoxygenase (LO) enzymes. As a result, when stimulated, such platelets produce less amounts of TXA2 and more of PGI2 that is less potent in inducing platelet aggregation and thrombosis (153). Increased intake of fish oil rich in EPA and DHA produces a lower platelet count, less platelet aggregation, a longer bleeding time, higher urinary fish oil rich in EPA and DHA produces a lower platelet count, from CAD/CVD. (PAI-1), they are effective still in preventing overall mortality and on the activity of plasminogen activity inhibitor-type-1 have a very significant effect on blood lipids, on fibrinolysis, ω-3 EPA and DHA as an “endogenous polypill,” then it is imperative that they (PUFAs) should show beneficial actions similar to those observed with conventional, synthetic diuretics. These beneficial actions of PUFAs can be attributed to the formation of beneficial PGA, PGE2, PGJ2, PGI2, and recently identified resolvins and protectins and a decrease in the production of TXA2 and LTs (165). In this context, it is interesting to note that diuretic furosemide enhances endothelial synthesis and the release of bradykinin and related kinins that, in turn, stimulates endothelial PGI2 formation via B2 kinin receptor activation (166); also, COX-2 derived PGs interact with the renin-angiotensin system to regulate renal function (167).
**Pufas and the Parasympathetic Nervous System**

An autonomic function can be assessed by the measurement of heart rate variability (HRV) and the evaluation of baroreflex sensitivity (BRS). HRV reflects the physiologic levels of tonic autonomic regulation, whereas BRS indicates the capacity of reflex autonomic regulation. Both low HRV and low BRS are associated with increased cardiovascular risk. Vagal stimulation by a release of acetylcholine (ACh) and adrenergic stimulation mediated by norepinephrine and epinephrine regulate the autonomic function and thus the variations in HRV and BRS. Several studies revealed that ω-3 fatty acids reduce the risk of sudden death by preventing life-threatening cardiac arrhythmias and by significantly increasing HRV (168). Furthermore, a direct positive correlation was noted between the content of DHA in cell membranes and the HRV index, which suggests an anti-arrhythmic effect of the ω-3 fatty acids (169). Because increased parasympathetic tone is responsible for an increase in the ventricular fibrillation threshold and protects against ventricular arrhythmias, it is likely that EPA/DHA supplementation enhances the parasympathetic tone. This suggestion is supported by the observation that EPA/DHA supplementation increases hippocampal ACh levels, the principal neurotransmitter of parasympathetic nerves (170). Hence, it is likely that EPA/DHA supplementation increases the brain ACh levels that lead to an increase in the parasympathetic tone and so to an increase in HRV and protection from ventricular arrhythmias.

Vagus nerve stimulation also inhibits TNF synthesis in the liver, and ACh, the principal vagal neurotransmitter, significantly attenuated the release of the proinflammatory cytokines TNF-α, IL-16, IL-18, and IL-22 but not the anti-inflammatory cytokine IL-10 by stimulated macrophages in vitro and in vivo (171–174). Because EPA/DHA enhances brain ACh levels (170) and evidence suggests that even AA augments ACh release (175), it is possible that Pufas enhance parasympathetic tone and, thus, increase HRV and prevent ventricular arrhythmias. These results imply that an inverse correlation could exist between the TNF levels and the parasympathetic tone. The higher the TNF levels are, the lower parasympathetic tone is and vice versa. Also, the higher the parasympathetic tone, the higher the brain ACh levels, and so the higher the plasma, cardiac, and brain EPA/DHA/AAs levels are. Because normally a balance is maintained between the parasympathetic and sympathetic tones, it is reasonable to suggest that whenever the parasympathetic tone (vagal tone) is enhanced sympathetic tone is reduced (akin to the blocking of β-receptors as it occurs in instances of the use of β-blockers). Thus, indirectly Pufas may function like β-blockers.

Thus, Pufas, especially an optimal combination of EPA, DHA and possibly, GLA, DGLA, and AA, show all the qualities of the suggested “polydrug,” viz they show aspirin-like action, inhibit the activities of HMG-CoA and ACE enzymes, possess diuretic and antihypertensive actions, and indirectly show β-blocker-like action.

**Pufas Possess Anti-Inflammatory Actions**

AA, EPA, DHA, GLA, DGLA, LXs and resolvins suppress IL-1, IL-2, IL-6, and TNF-α production by T cells (110–112, 149, 176–180). This claim suggests that EFA/Pufas and their metabolites function as endogenous anti-inflammatory molecules and regulate immune response and thus are likely to be of benefit in obesity, insulin resistance, atherosclerosis, metabolic syndrome X, type 2 diabetes mellitus, CHD, depression, and Alzheimer’s disease that are considered as diseases of low-grade systemic inflammation (1–8, 24, 120). Some beneficial actions of Pufas in various inflammatory conditions are because of the formation of anti-inflammatory compounds such as lipoxins, resolvins, and neuroprotection D1.

**Role of Efas/Pufas in Some Clinical Conditions**

It is evident from the preceding discussion that Pufas/Pufas and their metabolites are useful in many clinical conditions as outlined below.

**Inflammatory conditions**

Pufas and their products have the ability to modulate inflammation. The amount and type of Pufa(s) released and their products formed in response to inflammatory stimuli depend on the cell membrane phospholipid fatty acid content and the activity of the CO and LO enzymes. Because Pufas are obtained from diet, it suggests that dietary content of Pufas is one factor that modulates the degree of inflammation. Increased dietary intake of GLA, DGLA, and EPA/DHA substantially decreases inflammatory response. This beneficial action can be ascribed to the decreased formation of pro-inflammatory eicosanoids and cytokines and to an increase in the production of beneficial molecules such as PGE1, PG I2, PG I3, aNO, LXs, resolvins, and NPD1. When the cell membrane lipid pool is rich in GLA/DGLA/EPA/DHA and contains appropriate amounts of AA, there could occur specific activation of sPLA2 and cPLA2 (soluble and cytosolic phospholipase A2, respectively) in response to an injury/inflammatory stimuli that leads to the formation of increased amounts of LXs, PGD2, and 15deoxy-Δ12,14PGJ2, which results in the prevention and the resolution of inflammation. Several studies showed that oral or parenteral supplementation of GLA/EPA/DHA is of benefit to patients with insulin resistance, metabolic syndrome X, CHD, rheumatoid arthritis, lupus, psoriasis, sepsis, inflammatory bowel disease, nephritis, bronchial asthma, dermatitis, and other inflammatory conditions (reviewed in References 1–8, 24, 92, 116, 151, 152, 157).
Atherosclerosis

Healthy endothelial cells release adequate amounts of NO, PG12, and PGE1, to prevent the aggregation of platelets so that atherosclerosis is prevented. An increased production of pro-inflammatory cytokines and free radicals occurs because of sheer stress, hyperglycemia, clinical or subclinical infections, low-grade systemic inflammation as seen in type 2 diabetes mellitus, hypertension, hyperlipidemia, and metabolic syndrome X. EPA/DHA/ALA/DGLA inhibit free radical generation, suppress IL-6 and TNF-α synthesis and secretion, enhance eN0 synthesis, and, thus, prevent oxidant stress (reviewed in References 149, 150 and 183). Endothelial cells that line atherosclerosis-free blood vessel walls have abundant concentrations of the essential fatty acid linoleate, whereas fatty streaks (an early stage of atherosclerosis) are deficient in EFAs (183-186). An EFA deficiency promotes respiratory uncoupling (187, 188) and atherosclerosis (3, 183, 189). Bernal-Mizrachi et al. (184) showed that oxidative stress increases ROS generation and decreases NO formation. These evidences suggest that endothelial cell deficiency of PUFAs increases the production of pro-inflammatory cytokines and free radicals that results in the development of insulin resistance, a decrease in plasma and cellular HDL concentrations, and a decrease in the formation of eNO, PGE1, PG12, PGA2, LXs, resolvins, and ND1 that ultimately may promote atherosclerosis. Providing adequate amounts of various PUFAs can restore normalcy.

Metabolic Syndrome X

Metabolic syndrome X is a low-grade systemic inflammatory condition in which plasma levels of C-reactive protein (CRP), TNF-α, and IL-6 are elevated. A negative correlation exists between plasma TNF-α and HDL cholesterol, glycosylated hemoglobin, and serum insulin concentrations. EPA, DHA, and AA, inhibit TNF-α and IL-6 production (111, 112), enhance eN0 generation (30), inhibit HM-G-CoA reduce (103) and ACE enzyme activities (83), function as endogenous ligands for PPARs (reviewed in References 3 and 4), modulate leptin gene expression (190, 191), enhance the production of adiponectin (192), and decrease insulin resistance (193). This finding may explain why PUFAs are useful to protect against CHD, prevent the progression of atherosclerosis, and decrease blood pressure.

Brain Growth and Development and Cognitive Functions

Several studies showed that AA, EPA, and DHA are essential not only for brain growth and development but also to modulate the synthesis, release, and action of various neuropeptides. Because the brain is rich in AA, EPA, and DHA, one important function of these fatty acids in the brain could be to ensure the presence of an adequate number of insulin receptors; the insulin receptor number depends on the amounts of PUFAs incorporated in the cell membrane phospholipids (35-42). Thus a defect in the metabolism of PUFAs or not having adequate amounts of PUFAs incorporated into the neuronal cell membranes during fetal development and infancy may cause a defect in the expression or function of insulin receptors in the brain. This defect may lead to the development of type 2 diabetes as seen in NIRKO mice (194). Furthermore, systemic injections of either glucose or insulin in ad libitum fed rats resulted in an increase in extracellular acetylcholine in the amygdala (195). Aetylcholine modulates dopamine release that, in turn, regulates appetite (196). As already discussed above, ACh inhibits the production of pro-inflammatory cytokines in the brain and thus, ACh3, exists the neurons from the cytotoxic actions of TNF-α.

For proper neuronal development and increase in cell membrane surface area, growth of neurite processes from the cell body is critical (197). Nerve growth cones are highly enriched with AA-releasing phospholipases, which have been implicated in neurite outgrowth (198, 199). Cell membrane expansion occurs through the fusion of transport organelles with plasma membrane (200), and syntaxin 3, a plasma membrane protein that has an important role in the growth of neurites, is a direct target for AA, DHA, and other PUFAs (201). It was reported that AA, DHA, and other PUFAs activate syntaxin 3. Even syntaxin 1 that is specifically involved in fast calcium-triggered exocytosis of neurotransmitters is sensitive to AA (201). These results imply that AA, EPA, and DHA are involved both in the exocytosis of neurotransmitters and in neurite outgrowth. SNA P25 (synaptosomal-associated protein of 25 kDa), a syntaxin partner implicated in neurite outgrowth, interacted with syntaxin 3 only in the presence of AA that allowed the formation of the binary syntaxin 3-SNAP 25 complex (soluble N-ethylmaleimide-sensitive factor attachment protein receptor), which is needed for the fusion of plasmalemmal precursor vesicles into the cell surface membrane that leads to membrane fusion (202). These results clearly demonstrated that AA and DHA change the ω3-helical syntaxin structure to expose a SNARE motif for immediate SNAP 25 engagement and, thus, facilitate neurite outgrowth.

Puskas and his colleagues (203-205) noted that during brain growth and development, feeding with ω-3 DHA/ALA diets altered the expression of genes involved in synaptic plasticity, cytoskeleton, signal transduction, ion channel formation, energy metabolism, and regulatory proteins. These results imply that perinatal supplementation of PUFAs may play a critical role in the pathobiology of several adult diseases including metabolic syndrome X and Alzheimer’s disease (3, 4, 43, 44, 206-208).

Alzheimer’s Disease

Fish and fish oil components, EPA, and DHA are of benefit in Alzheimer’s disease (209-211). A reduction in dietary DHA in an Alzheimer’s mouse model showed a loss of postsynaptic protein teins associated with increased oxidation and showed increased caspase-cleaved actin, which was localized in dendrites; however, DHA-restricted mice when given DHA were protected against dendritic pathology and behavioral deficits and showed increased antiapoptotic B A D phosphorylation, which implies that DHA could be useful in preventing Alzheimer’s disease
in which synaptic loss is critical (212, 233). DHA attenuated amyloid-β secretion accompanied by the formation of neuroprotectin D1 (NPD1), a DHA-derived 10,17-docosatriene (214). In Alzheimer’s hippocampal cornu ammonis region, DHA and NPD1 were reduced, including the expression of enzymes involved in NPD1 synthesis, cytosolic phospholipase A2, and 15-lipoxygenase. NPD1 repressed amyloid-β-induced activation of pro-inflammatory genes and upregulated the antiapoptotic genes encoding Bcl-2, Bcl-xl, and Bcl-2-Related-Omics (Bcl-2) (211), which indicates its (NPDI) anti-inflammatory nature. Soluble amyloid precursor protein NPD1 synthesis from DHA (214).

Presenilin, a major component of γ-secretase, generates amyloid-β. Overexpression of phospholipase D1 decreases the catalytic activity of γ-secretase. PUFAs (especially AA and DHA) enhance acetylcholine release in the brain and, thus, bring about some of their beneficial effects in Alzheimer’s. Furthermore, EPA and DHA have the ability to enhance NO generation (30), suppress production of pro-inflammatory cytokines (110–113), and enhance brain acetylcholine levels (170), a neurotransmitter whose levels are decreased in Alzheimer’s disease (217). Thus, PUFAs modulate neural function, including neurotransmission, membrane fluidity, ion channel, enzyme regulation, and gene expression, as well as prevent inflammation; thus, they bring about their beneficial actions in Alzheimer’s disease.

Fetal-Alcohol Syndrome

Ethanol exposure during brain development induces neurodevelopmental defects referred to as fetal-alcohol syndrome (FAS) that is characterized by hyperactivity, learning and memory deficits, mental retardation, psychosis, depression, and schizophrenia. Ethanol-induced neurotoxicity, oxidative stress, induction of apoptosis, excitoactivity, interference with the action of growth factors, and EPA metabolism all could be responsible for the development of FAS. Neurons are susceptible to ethanol-induced apoptotic cell death during synapticogenesis during a brain growth spurt, which occurs during the third trimester of pregnancy and in the perinatal period. Recent studies showed that nicotine/nicotine enhanced the neuronal survival following free radical exposure and oxidative stress (218, 219). Nicotinamide is a cofactor in the metabolism of EFAs, which could explain the beneficial action of nicotinamide in FAS (220).

Depression

Depression is more likely to occur in individuals whose intake of PUFAs, especially of n-3 fatty acids, is lower (221). Pro-inflammatory cytokines might cause depressive illness (222). A significant decrease of w-3 fatty acids in plasma and/or in the membranes of red blood cells in subjects with depression has been reported (223–225). Because w-3 fatty acids suppress the production of IL-1β, IL-2, IL-6, and TNF-α, this finding suggests that these fatty acids could play a role in depression (222). In addition, antidepressants act like inhibitors of cyclo-oxygenase (222). Double-blind placebo-controlled and other studies (226–228) revealed that an addition of the w-3 fatty acids EPA and DHA was associated with a longer period of remission among depressed patients. Thus, epidemiological, experimental, and clinical data favor the idea that PUFAs play a role in the pathogenesis and/or the treatment of depression.

Schizophrenia and Huntington’s Disease

Patients with schizophrenia have increased concentrations of pro-inflammatory cytokines both in the systemic circulation and cerebrospinal fluid and showed decreased EPA and DHA in the plasma phospholipid. Clinical trials showed that supplementation of efflux DHA is of significant benefit to these patients (228). Huntington’s disease is an inherited neurodegenerative disorder because of a mutation in exon 1 of the Huntingtin gene that encodes a stretch of polyglutamine (poly Q) residues close to the N-terminus of the Huntington protein. Aggregated poly Q residues are toxic to the neuronal cells. Transgenic R6/1 mice that develop motor abnormalities of Huntington’s disease showed increased survival rates and decreased neurologic deficits when supplemented with ethyl EPA (229), which suggests that unsaturated fatty acids may prevent or arrest poly Q aggregation. These results suggest that PUFAs are useful in various neurological diseases. Understanding the molecular mechanisms of action of EPA/DHA as to why DHA is useful in Alzheimer’s disease whereas ethyl EPA is of benefit in Huntington’s disease and schizophrenia may throw more light on the pathobiology of these diseases.

Conclusions

It is evident from the preceding discussion that EFAs and their metabolites such as GLA, AA, EPA, DHA, eicosanoids, LXs, resolvins, NPD1, and nitrolypins have many actions and participate in several disease processes (Figure 2). In this context, it is important to note that certain PUFAs, such as GLA, have both antimutagenic and anticancer actions that have been discussed in detail elsewhere (230). Mutagens and carcinogens block ∆5 and ∆6 desaturases in normal cells much before their conversion into malignant cells. Pretreatment or simultaneous treatment with GLA completely prevented DNA damage induced by mutagens and carcinogens, which implies that GLA and other PUFAs could function as endogenous antimutagenic and anticancer molecules. GLA has selective tumorcidal action and is effective against human malignant glioma and other cancers (5, 231).

NO reacts with PUFAs to yield their respective nitroalkene derivatives that can be detected in plasma. These nitroalkene derivatives of various EFAs induce vascular relaxation, inhibit neutrophil degranulation and superoxide formation, inhibit platelet activation, and have endogenous PPAR-γ ligand activity.
and decay in the blood to release NO. This finding suggests that EFA sPUFAs not only form precursors to various eicosanoids, resolvins, LXs, and NPD1 but also react with various other molecules to form novel compounds that have significant biologic activity.

The major question is how these simple fatty acids can have so many biologic—and at times diametrically opposite—actions. One reason could be their ability to give rise to many metabolites that have specific biologic actions. Furthermore, these fatty acids when incorporated into the cell membrane alter its properties including fluidity that, in turn, modulates the number and affinity of various receptors to their respective growth factors, hormones, peptides, and proteins. Yet another action is their ability to form complexes with other biologically active molecules as is seen with NO to form various nitroalkene derivatives. Formation of such complexes between EFAs and other biologically active molecules could impart specific and distinct properties to these newly formed entities that in turn may show varied biologic actions. Deciphering the formation of such complexes is not only interesting but also challenging because such complexes may form the basis of understanding certain less well understood physiologic and pathologic processes.

A through structurally EFAs are simple, they form precursors to a variety of compounds with many biologic actions. We are yet to understand the molecular triggers that facilitate the formation of specific biologically active molecules in various cells and tissues. Such an understanding may lead to the development of methods to enhance selectively the formation of the desired lipid(s) to obtain a specific function or action. In view of their varied actions, EFA sPUFAs and their products could form the basis for the development of many drugs.

References


Essential Fatty Acids, Physiology and Clinical Significance of

Essential Fatty Acids, Physiology and Clinical Significance of


Essential Fatty Acids, Physiology and Clinical Significance of


183. Das UN. Defect in the activity of Δ6 and Δ5 desaturases may be a factor in the initiation and progression of atherosclerosis. Prostaglandins Leukot. Essent. Fatty Acids. In press.


Lipid bilayers, Properties of

Lipid bilayers are the structural basis of biological membranes; they provide excellent structures for microencapsulation and are interesting objects of study as thermotropic and lyotropic smectic liquid crystals. This article introduces the fundamental principles that lead to the formation of lipid bilayers and briefly lists some amphiphiles that form lipid bilayers in biological systems. The temperature-dependent phase behavior of bilayers formed from a single lipid species is discussed briefly followed by a discussion of bilayer dimensions and hydration. The article then introduces the reader to phase behavior of bilayers formed from mixtures of lipids and the consequences of phase coexistence for component compartmentalization and percolation in lipid bilayers. Special emphasis is placed on the phase behavior of bilayers that contain sterols. The article introduces the reader to some physical properties of lipid bilayers, which include the hydration force, elasto-mechanical properties, and the resistance of bilayers to lateral stretching and/or compression and bending. Brief discussions of molecular dynamics in lipid bilayers from trans/gauche isomerism and rotation of lipid chains in bilayers to rotational, wobbling, and translational diffusion of lipids in bilayers is followed by an introduction to the dynamics of insertion, desorption, and transverse translocation of lipids in bilayers. Lipid bilayers are then discussed in terms of their permeability to aqueous solutes and some models for bilayer permeation are mentioned briefly. Finally, the reader is introduced to the electrical properties of lipid bilayers with brief discussions of the dipolar potential, the surface electrostatic potential, and the transmembrane potential. The article ends with a brief discussion of lipid compositional asymmetry across the lipid bilayer in biological membranes and its possible causes. Key references are cited wherever relevant.

Lipid bilayers are formed by many amphiphilic molecules in the presence of water. Their interest derives not only from the fact that they are a major, if not the only, organizing principle of biological membranes (1), but also because they tend to form closed (usually) spherical structures (liposomes or lipid bilayer vesicles) in which inner and outer aqueous spaces are separated by the lipid bilayers (2) which thereby provides a means of encapsulation (3, 4).

Amphiphilic molecules (or amphiphiles), in general, tend to aggregate in aqueous solution above some critical concentration (the critical micellar concentration or CM C) because of the hydrophobic effect (5). The aggregate structure secludes the apolar portions of these molecules from the aqueous medium and exposes the polar portions to water at the surface. The form of the aggregate is dictated by packing constraints (6) given by the optimal area per amphiphile,\(^1\) A\(_0\); at the interface of the aggregate and the aqueous phase, the mean volume of the amphiphiles, V; and their critical length,\(^2\) l\(_c\). Lipid bilayers are formed when the packing parameter p = V /A\(_0\) l\(_c\) = 1. Thus, a lipid bilayer is a self-aggregated sheet of amphiphiles (usually two molecules thick) in which the polar portions of the constituent molecules are exposed to water at the two surfaces of the sheet and the apolar portions are secluded from water in the volume between these two surfaces. Partial or total interdigitation of the apolar portions of the amphiphiles from the two monolayers of the bilayer is possible under certain conditions. The surfaces of the bilayer in contact with the aqueous

\[1\]The “optimal area per amphiphile” is defined as the mean area per amphiphile at the surface of the aggregate when the free energy per amphiphile in the aggregate is a minimum.

\[2\]The critical length is the mean length of the amphiphiles in an aggregate along the normal to the aggregate/aqueous interface.
Lipid Bilayers, Properties of

phase may be charged depending on the chemical identity of the polar portion of the amphiphiles and the conditions (pH, ionic strength) of the aqueous phase; the bilayer interior always is an apolar environment. The chemical nature of the apolar portions (usually long aliphatic chains) and the fact that these chains are anchored to the polar head groups of the amphiphiles located at the bilayer-water interface, make the lipid bilayer a highly anisotropic structure. Figure 1 (7, 8) shows some typical lipid bilayers.

This article discusses lipid bilayers formed by lipids of interest to biological membranes. The physics of these lipid bilayers have been the subject of excellent monographs (9, 10) and bilayers have been discussed in the context of their colloidal properties (11). Compilations of data on the physical properties of lipid bilayers may be found in References 12 and 13. Most generic physico-chemical properties are common to all bilayers. Much of our knowledge on the properties of lipid bilayers is the result of studies on bilayers formed in the laboratory by hydration of chemically defined diacyl lipids, phosphoglycerolipids, or sphingolipids. The aggregates formed by these lipids in water are lyotropic and thermotropic in character (i.e., the structure depends on the molar fraction of water and the temperature (14–16)). These so-called “model” lipid bilayers form spontaneously on hydration of the amphiphiles. Usually, they are studied as oriented single bilayers or on stacked multibilayers on solid supports, multilamellar liposomes or vesicles (MLV), or unilamellar vesicles of different diameters: small unilamellar vesicles with a diameter of about 20 nm (17); large unilamellar vesicles with a diameter of about 100 nm (18); and giant unilamellar vesicles with a diameter in the µm range, usually 10-50 µm (19, 20). MLVs are composed of several concentric lipid bilayer vesicles each separated from the one inside and the one outside by a thin layer of water. This form of lipid bilayer is, therefore, a smectic liquid crystalline phase.

Chemical Composition

The main lipid constituents of bilayers in biological membranes are derivatives of sn-1,2- (or sn-2,3-) diacylglycerol (phosphatic acids, phosphatidylethanolamines, phosphatidylcholines, phosphatidylserines, phosphatidylglycerols, phosphatidylinositol, and glycosylated diacylglycerols), derivatives of sphingosine (sphingomyelin, ceramide, and glycosylated derivatives of ceramides), and steroids (cholesterol in mammalian membranes, β-sitosterol, campesterol, and stigmasterol in plants, and ergosterol in eukaryotic microorganisms such as fungi). Cardiolipins (1,3-diphosphatidylglycerol) are important constituents of bacterial membranes and the membranes of mitochondria and chloroplasts. The aliphatic chains of glycolipids may sometimes be linked to the glycerol by ether linkages as in the case of plasmalogens (important constituents of some mammalian membranes) and some bacterial glycolipids. The membranes of extremophiles contain variable amounts of lipids in which long-chain (usually branched) α,ω-aliphatic diols are attached by ether linkage to two polyols (usually glycerol) to form bipolar lipid molecules that span the entire lipid bilayer. The aliphatic chains of naturally occurring lipids are usually from 14 to 24 carbon atoms long, and may be fully saturated or unsaturated with from 1 to 4 (usually 3) double bonds. The aliphatic chains of bacterial lipids are often branched and may include cyclic (3- to 6-membered ring) structures. Figure 2 shows the chemical structures of some lipids of biological importance.

Physical Properties

Phase behavior and phase transitions

At low temperatures, in lipid bilayers prepared from a single lipid species, the acyl/alkyl chains of the lipids in the bilayer are characterized by a high trans/gauche configurational ratio.

---

![Figure 1](image1.png) **Figure 1** Different types of lipid bilayers. In the Lc and Lc′ phase bilayers, the polar head groups are shown as squares with arrows to indicate that a crystalline order exists in the arrangement of the head groups. In the P phase, the saw-tooth ripple is indicated by the parallel lines, and only a few lipids are drawn in this structure to show the way in which the lipids are arranged in the different domains of this structure. The reader is referred to References (7) and (8) for the structural details of the P phase.
Lipid Bilayers, Properties of

(a) Diacylglycerolipids (including Glycerophospholipids)

\[ R = \text{H, Diacylglycerol}, \]
\[ R = \text{Glycoside, Diacylglycerol-glycolipid} \]
\[ R = -\text{PO}_2\text{H}, \text{Phosphatidic acid} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH}_2)_2\text{NH}_2, \text{Phosphatidylethanolamine} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH}_2)_2\text{N}(\text{CH}_3)_2, \text{Phosphatidylcholine} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH})\text{COOH-NH}_2, \text{Phosphatidylserine} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH})\text{CH(OH)}\text{CH}_2\text{OH}, \text{Phosphatidylglycerol} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH})\text{NH}_2, \text{Phosphatidylserine} \]
\[ R = -\text{PO}(\text{OH})\text{O}(\text{CH})\text{CH}_2\text{OH}, \text{Phosphatidylinositol} \]

(b) Sphingolipids

\[ X = \text{H}, Y = \text{H, Sphingosine} \]
\[ X = \text{H}, Y = \text{OC(O)(CH}_2)_n\text{CH}_3, \text{Ceramide} \]
\[ X = -\text{PO}(\text{OH})\text{O}(\text{CH}_2)_2\text{N}(\text{CH}_3)_2\text{CO}-, \text{Sphingomyelin} \]
\[ X = \text{Glycoside, Y} = \text{CH}_2-\text{CH(OH)}\text{CH}_2\text{OH, Glycosphingolipid} \]

Figure 2. Chemical structures of the more common lipids found in biological membranes: a) lipids derived from Diacylglycerols; b) lipids derived from sphingosine; c) the more common sterols; d) some unusual lipids.

The chains, viewed along their long axes, are packed parallel to each other in a more or less hexagonal lattice with a lattice spacing of about 0.4 nm, and the bilayers are characterized by a high degree of conformational, rotational, and translational order. These bilayers are said to be in the "gel" phase. In gel phase, the acyl/alkyl chain axes may be oriented perpendicular to the bilayer plane (the \( L_\beta \) phase) or be slightly tilted relative to this plane (the \( L_\beta' \) phase). The tilt is understood to be the result of the in-plane area of the head groups being slightly larger than the in-plane area of two acyl chains. Depending on the chemical identities of the lipids and the temperature, more than one ordered phase has been characterized in lipid bilayers. In the so-called "sub-gel" or \( L_c \) phase (or \( L_c' \) when the chains are tilted relative to the bilayer normal) phase, which is observed at very low temperatures, the acyl chain order is at least equal to but may be greater than in the \( L_\beta \) (or \( L_\beta' \)) phase and may include a crystalline ordering of the lipid molecules. The "rippled" or \( P_{\beta'} \) phase, which is observed in some ordered bilayers at higher temperatures, shows a periodic saw-tooth ripple in the bilayer plane. This rippled pattern has two types of domains, the two faces of the saw-tooth pattern, with distinctly different packing of the lipids, the transition from one domain to the next possibly including some disordered phase lipid (7, 8). The acyl chain order in the \( P_{\beta'} \) phase is lower than that in the \( L_\beta \) (or \( L_\beta' \)) and \( L_c \) (or \( L_c' \)) phases. Transitions between the various phases in ordered lipid bilayers occur at characteristic temperatures that depend on the chemical identity of the lipids and their acyl chain lengths. When raising the temperature above a characteristic temperature, \( T_m \), which depends on the type of lipid that constitutes the bilayer, the ordered or gel phase is converted into a so-called "liquid-crystalline," "fluid," or \( L_\alpha \) phase in which the acyl chain configuration is characterized by a low trans/gauche configurational ratio and the chains are still packed in a more or less hexagonal lattice but with a low coherence length and
Lipid Bilayers, Properties of

(c) Sterols

Cholesterol

(-Sitosterol

Campesterol

Stigmasterol

Ergosterol

(d) Unusual Lipids

Plasmalogens

R = –O−PO(OH)−O−(CH2)2−N(CH3)3, Choline Plasmalogens

Cardiolipin

Bipolar lipid

Figure 2 (Continued)

Lipid bilayers in the fluid phase have low conformational, rotational, and translational order. The chain-melting transition at \( T_m \) is observed in all lipid bilayers regardless of the chemical identity of the lipid that constitutes the bilayer. The different transitions in the gel phase do not occur in all lipid bilayers, in which transitions depend on the lipid species of which the bilayer is constituted. When heating some lipid bilayers in the \( L_\alpha \) phase even more (e.g., those formed from some phosphatidylethanolamines), the hydrated lipid aggregate may be converted into one of several other phases that include inverted hexagonal and cubic phases. In bilayers formed from a given class of lipids, the characteristic temperatures at which the various phase transitions occur are dependent on the acyl chain lengths of the lipids and their degree of saturation.

The dependence of the “chain-melting” or main phase transition temperature in a lipid bilayer, \( T_m \), on the nature of the lipid head group as well as the length of the acyl chains and the degree and type (cis– or trans–) of unsaturation of the acyl chains has been studied exhaustively for some diacyl phospholipid classes (21). Bulkier head groups result in lower values of \( T_m \) for equivalent acyl chains. For head group homologs with two identical acyl chains, \( T_m \) increases monotonically, albeit in a nonlinear manner, with increasing acyl chain length. Usually, naturally occurring lipids have one saturated and one unsaturated acyl chain with one or more carbon-carbon double bonds, the latter usually in the cis– configuration. When more than one double bond occurs in the same acyl chain, these are usually separated from each other by a methylene group. The unsaturated chain is usually attached to the \( sn-2 \) position of glycerol. Lipids with unsaturated acyl chains in the \( sn-2 \) position have \( T_m \) values that are considerably lower than their fully saturated homologs, the reduction depends on the position and the number of unsaturated bonds. Introduction of a single double bond results in a very large drop in \( T_m \), the second unsaturated bond causes a smaller reduction in \( T_m \), and little or no reduction occurs in \( T_m \) on further addition of double bonds in an acyl chain. In the case of lipids with a single double bond in the \( sn-2 \) acyl chain, the largest effect on \( T_m \) is observed when the position of the double bond is roughly in the middle of the acyl chain.
As might be expected, the lipid bilayer phase transitions from a more ordered to a less ordered state are endothermic in nature and can be followed by differential scanning calorimetry (15). Bilayers prepared from dipalmitoylphosphatidylcholine (DPPC), for example, show at least three endothermic transitions: 1) the so-called subtransition (from the L_{α}′ to the L_{β}′ phase), which occurs at about 35°C with a transition enthalpy of 6.5 kJ mol\(^{-1}\); 2) the so-called pre-transition (from the L_{α}′ to the P3 phase), which occurs at 35°C with a transition enthalpy of 6.5 kJ mol\(^{-1}\); and 3) the so-called main (or chain-melting) transition, which occurs at 41.5°C with a transition enthalpy of 36.5 kJ mol\(^{-1}\). The enthalpies of the sub-transitions (when these occur) have been found to be more or less the same for bilayers prepared from different lipids irrespective of acyl chain lengths or head group structure. The same seems to be true for the enthalpies of the pre-transition. The enthalpy of the main or chain-melting transition, however, is very dependent on the acyl chain lengths and the degree as well as the type (cis- or trans-) of unsaturation of the lipid that forms the bilayer. It is similar for similar chain lengths in bilayers formed from lipids with different head groups. The phase transitions are also accompanied by nonmonotonic changes in lipid bilayer volume.

**Molecular and supramolecular dimensions in lipid bilayers**

The exact dimensions of a phospholipid bilayer membrane in terms of the in-plane area and the height of the lipid molecules as well as the thickness of the water layer that is associated with them is dependent on the chemical identity of the phospholipid head group, the length and the degree of saturation of the acyl chains, and the degree of hydration. This information may be obtained from a combination of small-angle X-ray diffraction by MLV or oriented multi-bilayer samples of phospholipids in excess water, electron and/or neutron density profiles across lipid bilayers, and atomic level molecular dynamics simulations of hydrated lipid bilayers. NMR studies on selectively deuterated phospholipids have also been important in elucidating acyl chain and lipid head group conformations. Our current understanding of the structure of phospholipid bilayers in terms of atomic/molecular detail has been summarized critically in an excellent review (23). The structural details of lipid bilayers in ordered (or gel) phases have been easier to define than in fluid (or liquid crystalline) phases that are more relevant for biological membranes. The reasons for this difficulty have to do with the poor diffraction patterns that result from long wavelength fluctuations or undulations in liquid crystalline phase bilayers, which destroy crystalline long-range order in the multilayer and local molecular fluctuations within each bilayer of the liquid crystalline lipid bilayer stacks. In L_{α} phase phosphatidylcholine bilayers, the choline head group has a probability distribution function that places it very close to the probability distribution function for the phosphate group (i.e., the choline head group lies roughly parallel to the bilayer plane and only slightly above the level of the phosphate group). The glycerol backbone of the lipid is oriented almost perpendicular to the membrane plane and the sn-1 acyl chain proceeds vertically into the bilayer, whereas the sn-2 acyl chain is kinked at the second carbon.

This structure makes the effective length of the sn-2 chain slightly (about 1.5 methylene groups) shorter than the sn-1 acyl chain. Table 1 lists some structural parameters for several phosphatidylcholine bilayers and one phosphatidyl-ethanolamine bilayer in the L_{α} phase. For comparison, the same structural parameters are listed for two of these bilayers [DPPC and dilauroylphosphatidyl-ethanolamine (DLPE)] in the gel phase as well. The most obvious structural differences among the L_{α} phase bilayer membranes listed in Table 1 can be attributed to the identity of the head group—the number of water molecules associated with the lipid bilayer is significantly different for the phosphatidylcholine (or phosphatidyl-ethanolamine) bilayers [DLPE] than when they are saturated (DPPC and DMPC). As might be expected, when gel phase bilayers are compared with L_{α} phase bilayers formed from the same lipid chemical species, the hydrophobic core thickness is significantly larger and the area per lipid is significantly lower in the gel phase (fully extended lipid chains). The number of water molecules associated with the lipid bilayers is also significantly reduced in the gel phase compared with the L_{α} phase (13 compared with 30 in DPPC and 6 compared with 9 in DLPE).

**Phase coexistence in lipid bilayers**

The properties of lipid bilayers formed from mixtures of lipids are very relevant to the understanding of the lipid bilayers that form the basis of biological membranes. Detailed studies have been performed on bilayers formed from binary lipid mixtures, and some reports in the recent literature describe phase diagrams of lipid bilayers prepared from ternary mixtures that include cholesterol. Figure 4 (24–31) shows some phase diagrams of lipid bilayers formed from binary and ternary mixtures of lipids. The general observation is that lipids in a bilayer are not very
miscible with each other in the gel phase. In binary mixtures of phosphatidylcholines, for example, differences in chemical identity of the head group and differences in chain length of four carbon atoms or more lead to gel phase immiscibility. Also, lipids with the capacity to form interdigitated gel phases do not mix well with lipids that do not have this capacity, even though their polar head groups and molecular masses are identical. Lipids generally seem to be mutually miscible in the fluid phase, although at least one report discusses immiscibility in the fluid phase of a bilayer prepared from a binary lipid mixture in which the polar head groups of the two chemical constituents of the bilayer were different.

Lipid bilayers that contain sterols are particularly interesting from the biological perspective because the plasma membranes of the cells of most eukaryotic organisms contain very large molar fractions of sterols (typically 30-50 mol%). In recent years, this aspect has received much attention because of the “raft” hypothesis (for a review on “rafts” see Reference 29). The condensing effect of cholesterol on fatty acids and phosphatidylcholine in monolayers has been known for a long time (32) and has been studied in detail. It was generally accepted that the condensing effect of cholesterol was a result of cholesterol–acyl chain interactions that forced the acyl chains in the lipid-expanded monolayer film to assume a conformation with higher trans/gauche configurational ratio. Studies on bilayers that contain high cholesterol concentrations confirmed these conclusions for bilayers as well; they showed that cholesterol significantly increased the trans/gauche ratio (conformational order) in the acyl chains of the lipids to values expected for rotating all-trans chains and were significantly higher than those typical for the cholesterol-free L1 phase. Translational diffusion in the cholesterol-rich bilayers was also shown to be slower compared with cholesterol-free bilayers. The reduction of the long-range translational diffusion coefficient has been reported to be from about 2-3 fold (28, 33) up to about 10-fold (34). In 1987, (pen et al. (35) integrated this information into a thermodynamic and a microscopic model for the interaction of cholesterol with the phospholipid molecules. They proposed that cholesterol forms an Ld phase with lipids in which the lipid chains are ordered conformationally. This phase, which was denominated the liquid ordered (Lo or L1 phase), is distinct from the Ld phase with conformationally disordered chains (which was denominated the Ld or Ld phase) observed in cholesterol-free lipid bilayers at temperatures above Tm. The formation of the Ld liquid-ordered phase is the consequence of the flat and rigid cholesterol structure that maximizes its interaction with the lipid acyl chains in an L1 phase bilayer by forcing these into a predominantly all-trans conformation but, at the same time, cholesterol cannot be incorporated into the in-plane crystalline lattice of a gel phase bilayer because of its size and shape. A compromise solution is to retain the all-trans configuration of the chains and simultaneously maintain the translational order of the fluid phase. Thus, in lipid bilayers that contain cholesterol, two fluid phases could possibly coexist in the lipid bilayer (see phase diagrams for ternary lipid mixtures in Fig. 4) depending on the molar fraction of cholesterol.

Phase coexistence in lipid bilayers may be an important physical property for membranes of cells. When two phases coexist in a bilayer, depending upon the relative mass fractions of the phases and the shapes of their domains, one of the phases is the percolative (physically continuous) and the other is nonpercolative (physically discontinuous or dispersed as isolated domains). Changes in the physico-chemical properties of the membrane (lateral pressure, temperature, and chemical composition are the most relevant for biological membranes) result in interconversion between the two phases— one phase grows at the expense of the other. In phase-separated systems of this type, a critical mass ratio of phases called the percolation threshold, at which the previously continuous phase becomes discontinuous and the previously discontinuous phase becomes continuous, becomes

| Table 1 Structural parameters for some fully hydrated phospholipid bilayers (taken from Reference 23) |
|-----------------|-----|-----|-----|-----|-----|-----|-----|
| Lipid           | DPC | DPPE| DMPC| DOPC| EPC | DLPE| DLPE |
| Temperature     | 20°C| 50°C| 30°C| 30°C| 30°C| 20°C| 35°C|
| V (nm³)         | 1.144| 1.232| 1.103| 1.125| 1.103| 1.261| 0.863| 0.907|
| D (nm)          | 6.70 | 6.70 | 6.27 | 6.31 | 6.63 | 5.06 | 5.24 | 4.88 |
| A (nm²)         | 0.479| 0.460| 0.596| 0.725| 0.694| 0.410| 0.512| 0.454|
| 2D (nm)         | 2.84 | 2.85 | 2.62 | 2.71 | 2.71 | 2.00 | 2.58 | 2.00 |
| Dₒ (nm)         | 4.42 | 3.83 | 3.60 | 3.69 | 3.69 | 3.98 | 3.56 | 3.56 |
| Dₒ' (nm)        | 5.24 | 4.65 | 4.42 | 4.51 | 4.51 | 4.70 | 4.28 | 4.28 |
| nₒ              | 12.6 | 30.1 | 25.6 | 32.8 | 34.7 | 5.9  | 8.8  | 8.8  |
| nₒ'             | 3.7  | 8.6  | 7.2  | 11.1 | 10.2 | 2.0  | 4.7  | 4.7  |

V, lipid molecular volume; D, lamellar repeat spacing; A, average interfacial area per lipid; 2D, thickness of the hydrocarbon core of the lipid bilayer; Dₒ, head group peak to head group peak distance in the section density profile of a bilayer; Dₒ', steric thickness of the bilayer; this thickness includes some water intercalated in the polar head group of the lipid; nₒ, number of water molecules associated with each lipid in the bilayer; nₒ', number of water molecules that are “bound” or intercalated in the polar region of the lipid bilayer.
Lipid Bilayers, Properties of

**Figure 4** Some phase diagrams for lipid bilayers in excess water prepared from binary and ternary lipid mixtures. a) Multibilayer lipid vesicles prepared from binary mixtures of DMPC and DPPC (24); b) Multibilayer lipid vesicles prepared from binary mixtures of DMPC and DSPC [adapted by Reference (25); from data for perdeuterated lipids published by Knoll et al. (26)]; c) Multibilayer lipid vesicles prepared from binary mixtures of diC17:0PC and C22:0C12:0PC (27); d) Multibilayer lipid vesicles prepared from binary mixtures of DMPC and cholesterol (28); e) Multibilayer lipid vesicles prepared from ternary mixtures of palmitoyl sphingomyelin, POPC, and cholesterol [adapted by Reference (29), from data published by De Almeida et al. (30)]; lipid bilayers prepared from ternary mixtures of DSPC, DOPC, and cholesterol (31).

Phase separation in a membrane may lead to segregation of membrane components, which includes proteins (in-plane compartmentalization) based on their preferred solubility in one or the other of the coexisting phases; this segregation may have important consequences for in-plane bimolecular reactions that occur in these membranes (37). Crossing a percolation threshold (changes in the phase mass ratio may be induced by osmotic stress, temperature changes, or changes in chemical composition) can connect previously disconnected domains and their constituents. The previously continuous domain, and its constituents, becomes simultaneously discontinuous.

In systems with phase coexistence, the free energy of the system is discontinuous at the interface between the phases. This discontinuity results in an interfacial tension that drives the system toward a minimization of the interface. In three-dimensional systems, the interfacial tension acts on the area (surface tension); in two-dimensional systems, it acts along the line (line tension) that separates the phases. In both cases, if the interfacial tension is large enough, then it may be expected that...
the coexisting phases separate into macroscopic domains with the smallest possible separating surface or line between them (a so-called bulk phase separation) at equilibrium. Interfacial tension may be reduced by surface-active agents, which are amphiphilic molecules that locate in the interfacial region and interact more favorably on one side with one of the phases and on the other side with the other phase. Such a reduction in interfacial tension is observed in thermodynamically stable emulsions and in metastable dispersions of immiscible phases. In the lipid bilayers of biological membranes, it is conceivable that some lipids and proteins may act as the amphiphiles that reduce interfacial line tension between domains of coexisting phases in the lipid bilayer.

Although the question of whether phase separations do occur in the lipid bilayers of biological membranes is hotly debated, it is generally accepted that these membranes do show spatial and temporal heterogeneities (domains) in lipid bilayer order. These heterogeneities may have several origins. First, membrane physiology involves processes of vesicle fusion with and budding from a given membrane. These processes will result in localized heterogeneities in chemical composition and physical properties in the lipid bilayer, which dissipate in a diffusion-limited manner within a short time. Second, protein aggregation by some force external to the lipid bilayer (reorganization of the cytoskeleton with its attached proteins; extra-membranal cross-linking of proteins, etc.) may cause a coalescence of the boundary lipid shells around the proteins, which results in a localized change in lipid order that is associated with the protein aggregate. Because the boundary layer has a coherence length of about two lipid shells around the protein surface, this heterogeneity may be expected to dissipate with the disintegration of the protein aggregate. Third, assuming that the system is in a steady state, the lipid bilayer of biological membranes may be a multiphase system because of its compositional complexity and the mutual immiscibility of its chemical constituents. Domains of these phases would be thermodynamically stable. As discussed above, if the line tension around the phase domains is small enough to be insufficient to drive domain growth, then phase coexistence in the bilayer does not necessarily imply a bulk phase separation or that the small dispersed domains of a given phase grow in time. They may, however, be dragged together by forces external to the membrane to form larger domains or platforms.

As discussed at the beginning of this article, the formation of lamellar aggregates such as bilayers by hydrated lipids is dependent on the shape of the lipid molecules, which are characterized by a critical packing parameter, \( p \), whose value must be above equal to 1. When \( p \approx 1 \), spherical micellar aggregates are formed, and for \( p > 1 \) the aggregate formed is an inverted micelle or an inverted hexagonal phase. Both aggregates are characterized by very large curvatures (curvature \( \approx 1/R \), where \( R \) is the radius of curvature; see Fig. 5b for definitions) (38), the normal micellar aggregate has a positive and the inverted aggregate a negative curvature. Some physiological events such as membrane fusion, membrane fission, and the drawing out of very narrow tubular membrane structures require the localized formation of membrane structures with high curvatures. Thus, to be responsive to these physiological events, the lipid bilayers in biological membranes must contain a certain amount of lipids that are readily available for and that can form these highly curved structures. At the same time, the presence of these lipids should not cause a significant perturbation of the bilayer membrane and its physical properties. This balance is maintained by biochemical mechanisms that change membrane lipid composition according to need (39).

**The hydration force**

The characteristics of the water associated with the polar lipid bilayer surfaces are of particular interest because they become very important in processes of physiological significance such as membrane fusion, and they may play a role in the mechanisms of association of proteins and small molecules with lipid bilayers and biological membranes. A very small fraction of the lipid bilayer-associated water molecules are actually immobilized, and a larger fraction (about 30% of the total bilayer-associated water in multilamellar PC bilayers in the L\(_c\) phase) has a probability distribution function that is more or less coincident with the probability distribution function for the polar head group of the lipids. Nevertheless, the pressure, \( P \), which must be exerted to remove the bilayer-associated water, is large and varies (from \(-0.5\) to \(-500 \text{ N cm}^{-2}\)) with the thickness of the inter-bilayer water space as:

\[
P = P_0 \exp\left(-\frac{\alpha}{\lambda}\right)
\]

where \( d_w \) and \( \lambda \) are the inter-bilayer water thickness in a multibilayer stack, and \( \lambda \) is a characteristic length (\( \approx 0.2 \text{ nm} \) in L\(_c\) phase PC bilayers). The resistance to removal of the inter-bilayer water has been called the “hydration force”; it is the consequence of contradictory interaction forces: A van der Waals interaction that brings bilayers together is counterpoised by repulsive interactions that result from long-range undulation fluctuations of the fluid lipid layers, static repulsion of the head groups, and water dipoles with a preferred orientation at the bilayer-water interface (surface hydration). This phenomenon has been discussed by Parsegian and coworkers in excellent reviews (40, 41).

**Elasto-mechanical properties of lipid bilayers**

Lipid bilayers fall into the category of materials that have come to be known as “soft matter.” These materials are condensed phases that possess many characteristics of liquids but are simultaneously structured. The intermolecular interactions that cause lipid molecules to self-aggregate to fluid bilayers in the presence of water also impart a degree of “toughness” to these bilayers compared with conventional liquids. For detailed discussions of these properties of lipid bilayers, the reader is referred to References 10, 38, and 42. Toughness implies resistance to forces that shear, break, or bend the bilayer (Fig. 5). Bending without breaking and resistance to shear stresses, for instance, are essential properties of the membranes of cells (particularly erythrocytes and platelets) that are forced in the blood stream to squeeze through very narrow capillaries. This property is also essential to cell membranes (in particular the apical surface membranes of epithelial and endothelial cells).
Lipid Bilayers, Properties of

Stretching Bending Shearing
(a) Forces that result in elasto-mechanical deformations in bilayers

Definitions of the curvatures in lipid monolayers and bilayers

Curvature radii
Spontaneous curvatures
Positive curvature
Negative curvature

Figure 5  a) Schematic view of the mechanical stresses and consequent elasto-mechanical deformations that can occur in lipid bilayers. b) (left panel): Definition of the curvatures in a bent sheet (38); (right panel): Illustration of the spontaneous curvatures in a lipid monolayer.

that line the lumen of tubes in the body through which relatively viscous fluids flow (for example, the endothelial cells of blood vessels, the epithelial cells of the gastrointestinal tract, the kidney and the urinary tract, the respiratory tract, etc.).

The lipid bilayer membrane in living cells is a fluid membrane and, therefore, has no shear rigidity. However, within the cell and subjacent to the membrane lies an intricate network of the cytoskeleton that is attached with some regularity to the lipid bilayer that constitutes the cell membrane via proteins that are anchored in the bilayer. The shear rigidity of cell membranes is thus provided in a large measure by the cell cytoskeleton.

Resistance of an infinitely thin sheet to stretching (or in-plane compression) and bending stresses are expressed in terms of the respective elasto-mechanical moduli. The area compressibility modulus, $K_A$, is defined via the energy required per unit area $E_{KA}$ to produce an area increment $\Delta A$ in a sheet of reference area $A_0$: $E_{KA} = \frac{1}{2} K_A (\frac{\Delta A}{A_0})^2$.

The resistance to bending is expressed in terms of two moduli: the mean curvature modulus, $\kappa_C$, and the Gaussian curvature modulus, $\kappa_G$, which are both defined via the energy, $E_C$, required for bending the sheet (43): $E_C = \left[ \frac{\kappa_C}{2 R_1^2} + \frac{\kappa_G}{2 R_2^2} + \frac{K_C}{R_1 R_2} \right] A$, where $A$ is the area of membrane, $R_1$ and $R_2$ are the principal radii of curvature, and $R_0$ is the radius of spontaneous curvature of the sheet (see Fig. 5 for definitions). In the case of lipid bilayers, the spontaneous curvature is the result of the packing parameter, $p$, for the lipids that constitute the two monolayers of the lipid bilayer (see above for definition) not being exactly equal to 1. For bilaterally symmetrical bilayers, the spontaneous curvature ($C_0 = R_0 - 1$) is zero because the two monolayers curve in opposite directions and the force resulting from the hydrophobic effect symmetrically forces them together. For asymmetric bilayers, however, this may not be the case. The mean curvature modulus of a bilayer, $\kappa_C$, and the area compressibility modulus, $K_A$, are related to each other ($K_C = K_A d^2$, for the simplest case of an infinitely thin sheet), the exact relationship is dependent on the coupling between the monolayers and whether the lateral pressure distribution is uniform across the bilayer (38). $\kappa_C$ has a quadratic dependence on the thickness of the bilayer.

Marsh (38) has listed several measured values of the elastic moduli of lipid bilayers. Typically, the area compressibility moduli are in the range of 200-250 mN m$^{-1}$ for fully hydrated symmetric bilayers in the $L_\alpha$ phase prepared from phosphatidylcholines, and they are not very dependent on the degree of saturation of the acyl chains. Cholesterol has a significant effect on the area compressibility modulus of a bilayer. Thus, $K_A$ for bilayers of 1-stearoyl-2-oleoylphosphatidylcholine (SOPC) increases from 235 mN m$^{-1}$ in the absence of cholesterol ($L_\alpha$ liquid-disordered phase) to 640 mN m$^{-1}$ in bilayers with 40 mol% cholesterol ($L_{eta'}$-rich ordered phase).
made from an equimolar mixture of SOPC and cholesterol (Lα, liquid-ordered phase). The mean curvature elastic moduli have been reported to be in the range of $10^{-12}$ for Lα phase phosphatidylcholine bilayers, but the exact value is dependent on the method used for their determination.

Molecular dynamics, diffusion, trans-bilayer translocation, and lipid exchange in bilayers

The dynamics of molecules in a lipid bilayer is of fundamental importance in the role that the lipid bilayer plays in a biological membrane. Dynamics in a lipid bilayer spans a frequency scale from $10^{12}$ s$^{-1}$ (for the vibrational dynamics of single bonds in the lipid molecules) through $10^{-10}$ s$^{-1}$ for trans-membrane translocation ("flip-flop") of a lipid molecule. From a functional perspective, the most important dynamic modes that have to be considered are the trans-gauche configurational isomerism in lipid molecules and the rotational and translational directional dynamics of these in the bilayer. These dynamic modes determine how fast the lipid bilayer responds to changes in volume and pressure (including localized changes resulting from conformational transitions of proteins associated with the bilayer of biological membranes) and the rate at which lipids exchange between different environments in a bilayer. A rather important dynamic mode is the rate at which lipid molecules spontaneously desorb from and insert into lipid bilayers, because these rates determine the lower limit (noncatalyzed rates) at which lipids can exchange between the different (presumed noncontiguous) membrane environments in cells.

Changes in the configuration (trans-gauche isomerism) of single bonds in lipid molecules result in rotational motions in the lipid acyl chains and head groups; an increased configurational entropy is the cause of phase transitions in lipid bilayers. The global lipid conformation is a result of the combination of configurations about single bonds in the acyl chains and the head group. The headgroup and acyl chains may be viewed as anchored positionally at the glycerophosphate group at the aqueous interface, and they have the freedom to execute angular excursions relative to the bilayer normal. The headgroup is, therefore, capable of motion in a cone whose vertex lies at the carbon atom linked to the phosphate group and has some preferred orientation, which is known to be more or less parallel to the bilayer surface in bilayers prepared from phosphatidylcholines and from phosphatidylethanolamines. Each acyl chain is also capable of a conical excursion with the vertex close to the carbonyl group. The angular excursion executed by each carbon atom in the chain varies with the position of the carbon atom in the chain. In general, carbon atoms closer to the glycerophosphate anchor (approximately up to carbon atom 10) execute a smaller angular excursion than those farther away from it. The excursion angle is usually expressed in terms of the C-H bond order parameters measured by H{H-NMR on lipid bilayers prepared from deuterated lipid molecules (44, 45). Frequencies for trans-gauche isomerization depend, as might be expected, on the physical state of the lipid bilayer, and are found to be $10^{-11}$ s$^{-1}$ for the head group and $10^{-9}$ s$^{-1}$ in the chains in the crystalline (Lc) phase and increase to $10^{8}$ s$^{-1}$ in the fluid (Lα) phase (7).

In the fluid (Lα) phase, the constituents of a lipid bilayer have diffusional freedom that may be described in terms of the diffusion coefficients for wobbling motions of the lipid long axis (D⊥), rotation of the lipids around their long axis (D∥), and the translational diffusion in the plane of the lipid bilayer (D,) and the transmembrane direction (D.). Deuterium NMR studies (46) and Molecular Dynamics simulations (47) of pure dimyristoylphosphatidylcholine (DMPC) bilayers at 40-45°C give values of D⊥ in the range of $10^{6}$ to $10^{7}$ s$^{-1}$ and values of D∥ of $10^{-3}$ s$^{-1}$. In Lα, liquid-ordered phase bilayers prepared from equimolar binary mixtures of DMPC and cholesterol, D⊥ increases about threefold, whereas D∥ decreases over 10-fold (48), which is consistent with a higher order and reduced chain entanglement in the liquid-ordered compared with the liquid-disordered phase. The translational diffusion coefficient (D,) has been measured using a wide variety of techniques. In Lα, liquid-disordered phase lipid bilayers prepared from pure phosphatidylcholines, D⊥ from $10^{-10}$ to $10^{-11}$ m$^{2}$ s$^{-1}$ depending on the lipid and the temperature (49), which corresponds to a hopping frequency ($ν_h$) of $10^{7}$ to $10^{8}$ s$^{-1}$ assuming the area per lipid molecule in the bilayers to be $0.5$ nm$^{2}$. In gel phase (Lβx-1), phase phosphatidylcholine bilayer D⊥ has been reported to be from 3 to 9 orders of magnitude slower ($10^{-15}$ s$^{-1}$) than in the fluid phase (for a review see Reference 50).

In Lα, liquid-disordered phase membranes prepared from binary mixtures of a phosphatidylcholine and cholesterol, the value of D⊥ has been reported to be from $2-3$ times lower to about $10$ times lower than it is in Lα liquid-disordered phase membranes prepared from the phosphatidylcholine alone (28, 33, 34).

The theoretical description of translational diffusion in a lipid bilayer depends on the size of the diffusing particle. Theoretical descriptions based on fluid hydrodynamic theory (51, 52) have been shown to be applicable to particles whose radius in the plane of the bilayer is significantly larger than the radius of the lipid molecules that constitute the bilayer, in which case the diffusion coefficient may be given by:

$$D_i = \frac{kT}{4\pi \eta b} \left( \frac{v_w}{\eta_w} \right)^{-1} \left[ \frac{R}{h} \right]^{-1}$$

where $k$ is the Boltzmann constant, $T$ is the temperature, $h$ is the thickness of the bilayer sheet and height of the embedded diffusing particle considered to be a cylinder of radius $R$, $\eta$ are the viscosities of the bilayer fluid and the bounding water, respectively, with $\eta_w < \eta$, and $\gamma = 0.5772$ is Euler's constant. The diffusion of lipids or particles with a radius that is equal to or smaller than that of the lipid molecules in a lipid bilayer is not described adequately by the continuum fluid hydrodynamic model; a theoretical model based on free volume theory has been proposed for this purpose according to which D⊥ may be given by (48, 50):

$$D_i = \frac{kT}{4\pi \eta b} \left[ e^{-\gamma} / (\gamma \eta b + 1) \right]^{1/2}$$

where γ is a factor that accounts for overlapping free volumes ($0.5 \leq \gamma \leq 1$). $v_w$ is the van der Waals molecular volume of a
lipid in the bilayer, \( v^* \) is a critical free volume below which the diffusing particle does not move into an adjacent free volume, \( v_{fp} \) is the free volume at the phase transition temperature, \( T_m \), and \( v \) is the coefficient of volume expansion of the fluid lipid bilayer. In this model, particles that span the entire thickness of the bilayer feel the translational frictional drag, \( 1 \), at both the bilayer-water interfaces. Particles that span only a single monolayer feel the frictional drag at one of the bilayer-water interfaces and at the bilayer midplane. There are no dissipative interactions within the bilayer itself. The free volume, \( v_m \), of the fluid lipid bilayer is the difference between the volume per lipid at temperature \( T \) and the van der Waals volume of the lipid:

\[
v_m = v - v = v_0 + 6[1 - (T_m - T)]
\]

The viscous drag felt at the bilayer-water interface can be large when the bounding fluid viscosity approaches the viscosity of the membrane (53) or when the fluid lipid bilayer is associated with a rigid substrate (54).

The lipids in a lipid bilayer may translocate across the bilayer from one monolayer to the apposed monolayer. This transmembrane translocation process, which is also known as "flip-flop," is slow for lipids with large polar head groups such as glycero-lipids and sphingoglycosphospholipids but can be fast in the case of lipids with very small polar moieties such as cholesterol. Typical first-order rate constants for transmembrane translocation of a phospholipid-like molecule in \( L_\alpha \) liquid-disordered phase bilayers prepared from 1-palmitoyl-2-oleoylphosphatidylcholine (POPC) are \(-10^{-8} \text{ s}^{-1}\) and may be about 10-fold slower in \( L_\alpha \) liquid-ordered phase bilayers prepared from binary mixtures of the same phosphatidylcholine and cholesterol but is much slower (\(-10^{-3} \text{ s}^{-1}\)) in \( L_\alpha \) liquid-ordered phase membranes prepared from sphingomyelin and cholesterol (55). The activation free energy for the process, which corresponds to the energy necessary to put the translocating lipid molecule at the bilayer mid-plane, is \(-100 \text{ kJ mol}^{-1}\). In contrast, the rate constant for transmembrane translocation of cholesterol may be \(-1 \text{ s}^{-1}\) (56).

In any system that contains noncontiguous lipid bilayers, such as a living cell, the lipid components of these bilayers can exchange through a process that involves noncatalyzed desorption of single lipid molecules from one membrane and insertion into another. This lipid exchange involves three processes: 1) a first-order desorption of lipid molecules from the donor bilayer, 2) a diffusion-limited second-order encounter of the lipids in the aqueous phase with the surface of the acceptor bilayer to form an encounter complex, and 3) a first-order insertion of the surface located lipid molecules into the acceptor bilayer. The rate constants for the first and last processes have been measured for lipid-derived probes (57). Desorption rate constants are on the order of \(10^{-9} \text{ s}^{-1}\) from \( L_\alpha \) liquid-disordered phase bilayers prepared from POPC and are about 10-fold slower when the bilayer is an \( L_\alpha \) liquid-ordered phase prepared from sphingomyelin and cholesterol. Insertion rate constants are on the order of \(10^{-6} \text{ s}^{-1}\) in \( L_\alpha \) phase bilayer membranes prepared from POPC, with or without cholesterol, and about 10-fold slower when the acceptor membrane is an \( L_\alpha \) liquid-ordered phase prepared from a binary mixture of sphingomyelin and cholesterol. Thus, the noncatalyzed exchange of lipids between noncontiguous lipid bilayers in a cell is a slow process that is limited by the rate of desorption from the donor lipid bilayers.

**Lipid bilayers as permeability barriers**

Because of its apolar interior, the lipid bilayer is a barrier to diffusional equilibration of solutes between the two aqueous compartments that it separates. The ability of most small solute molecules (50 < molecular weight < 300) to cross the bilayer is directly proportional to their ability to partition into hexadecane or olive oil from an aqueous solution (58), which is an observation first made by Overton (59) and is often referred to as "Overton's Law." Permeation of lipid bilayers by small polar molecules and ions seems to occur via one or a combination of both of two mechanisms depending on the nature of the permeants and the nature of the bilayers. First, a "solubility-diffusion" mechanism treats the bilayer as a slab of liquid hydrocarbon sandwiched between two bulk aqueous compartments. The permeant must partition into the bilayer slab from one of the aqueous compartments, diffuse across it, and leave by dissolving into the second aqueous compartment. In this case, the permeability coefficient, \( P \), is given by:

\[
P = \frac{K v D_m}{d}
\]

where \( K \) is the partition coefficient of the permeant between the bilayer and aqueous phases, \( D_m \) is its translational diffusion coefficient in the bilayer phase and \( d \) is the bilayer thickness. Second, a pore mechanism assumes the formation of transient water-filled pores (often referred to as "water wires") across the bilayer because of density fluctuations in it. The permeant (in a partially or fully hydrated state) is then assumed to diffuse through the bilayer via these transient pores. In this case, the permeation coefficient is proportional to the probability of formation of pores that span the entire bilayer, the mean pore radius, and the translational diffusion coefficient of the permeant in water. The expected dependence of the permeation coefficient on bilayer thickness is steeper in the case of the pore mechanism than it is in the solubility-diffusion mechanism, which thereby provides the basis for a critical test of which mechanism is applicable to the permeation of a given permeant particle (60).

Uncharged, small polar molecules such as water, glycerol, and urea permeate bilayers, regardless of their thickness, via the solubility-diffusion mechanism. Protons seem to permeate bilayers predominantly via the pore mechanism for bilayers with an apolar layer thickness of up to \(-3 \text{ nm}\) (phosphatidylcholine bilayers with acyl chains up to about 16 to 18 carbon atoms); the solubility-diffusion mechanism becomes predominant for thicker bilayers. Potassium and halide ions permeate bilayers almost completely via a solubility-diffusion mechanism when the bilayers are thicker than \(-2 \text{ nm}\) (60, 61). Typical permeability coefficients, for permeation of bilayers prepared from phosphatidylcholines in the \( L_\alpha \) liquid-disordered phase with a 3 nm hydrophobic thickness, are \(-10^{-3} \text{ m s}^{-1}\) for water permeation, \(-10^{-9} \text{ m s}^{-1}\) for the permeation of protons, \(-3 \times 10^{-11} \text{ m s}^{-1}\) for potassium ions, and \(-6 \times 10^{-12} \text{ m s}^{-1}\) for chloride ions.
Electrical properties of lipid bilayers

Lipid bilayers have electrical properties that are not just a result of the charges on the lipid molecules. The association of any charged particle with a lipid bilayer and/or the permeation of lipid bilayers by any charged particle are strongly conditioned by the electrical properties of the bilayer. There are three types of electrical potentials associated with lipid bilayers suspended in aqueous electrolyte solutions: 1) a dipole potential, 2) a surface potential, and 3) a transmembrane potential.

The dipole potential (discussed in Reference 62) results from the anisotropic orientation of the lipids in a bilayer and from orientation of water dipoles at the lipid bilayer-water interface. The dipole potential of an L_2 liquid-disordered phase phosphatidylcholine bilayer is ~250 mV, which is positive inside the bilayer. The anisotropic orientation of the lipid molecules in a bilayer results in a “surface dipole moment” whose main contributions come from the dipole moments of the terminal methyl groups and the carbonyl groups of the lipid acyl chains, particularly the carbonyl group of the acyl chain attached to the sn-2 position of glycerol. The dipole moments of the lipid head groups and the chain methylene groups are assumed to contribute nothing because of their orientation parallel to the surface. A very significant, if not the major, contribution to the dipole potential may have its origin in the orientation of bound water dipoles at the membrane-water interface (63). As observed earlier, the orientation of water molecules at the bilayer surface is an important cause of the hydration force. The dipole potential of a lipid bilayer, which is dependent on lipid orientation, is different for different lipid phases. More ordered lipid bilayers might be generally expected to have a larger dipole potential. Differences in the dipole potential of different coexisting phases in a bilayer are responsible for determining the size, shape, and arrangement of domains in lipid monolayers (and possibly bilayers as well) and may make the coalescence of dispersed domains of the same phase to a macroscopically separated phase in the bilayer a very slow process (64). It may even determine the transverse superposition of domains of the same (or similar) phase across the lipid bilayer with phase coexistence (65).

The surface potential of a bilayer is a result of having charged lipids in this bilayer (for reviews see References 63 and 66). In fluid phase bilayers, rapid translational diffusion of the lipids allows the surface charge associated with the lipids to be considered a smeared charge and the electrostatic potential at the surface of the bilayer, \( \psi_s \), is well described by the Gouy-Chapman equation:

\[
\sinh(2e\zeta_0 T/2kT) = A_0/(c\zeta_0)^{1/2}
\]

where \( \zeta \) is the valence of the counter ion in the electrolyte solution, \( e \) is the electronic charge, \( k \) is the Boltzmann constant, \( T \) is the temperature, \( \sigma \) is the surface charge density, \( c \) is the number of ions of a z+2 electrolyte in the aqueous solution, \( A_0 = (\kappa_0 z e_{\text{water}} T)^{-1} \) is a constant, \( e_0 \) is the permittivity of free space, and \( \kappa_0 \) is the dielectric constant of water (~80). The electrostatic potential varies with distance, \( x \), from the charged surface as:

\[
\psi_s = \frac{kT}{e} \ln \left[ \frac{1 - \exp(-\alpha x)}{1 + \exp(-\alpha x)} \right]
\]

where \( \alpha = (\exp(e\zeta_0 T/2kT) - 1)/(e\exp(e\zeta_0 T/2kT) + 1) \) and \( \alpha^2 = ([\kappa_0 z e_{\text{water}} T]/2z^2 e^2 e_0)^{1/2} \) is the Debye length. For small surface potentials, the Gouy-Chapman equation for the surface potential reduces to:

\[
\psi_s = \varphi_0/z \exp(-\alpha x)
\]

and the expression for the distance-dependent potential reduces to:

\[
\psi_s = \varphi_0 \exp(-\alpha x)
\]

The surface potential results in a concentration of the counter-ions and a depletion of the co-ions in the aqueous phase close to the charged surface, which results in the so-called electrical double layer. The ion concentrations in the aqueous phase at a given distance \( x \) from the charged surface are given by:

\[
c_+ = c \exp(-2\zeta_0 \psi_s/kT)
\]

It is important to note that each monolayer that constitutes a lipid bilayer has its own surface potential, and whether the surface potential is equal on both sides of the lipid bilayer will depend on the transverse compositional symmetry of the bilayer. Lipid bilayers in biological membranes are usually compositionally asymmetric, which is a fact that could originate a potential difference (generally given as \( \Delta \Phi \) in the literature) between the two membrane surfaces that may be expected to play a role in ion permeation of the lipid bilayer and, therefore, be important in cell membrane physiology. One consequence of the surface potential is that protons are concentrated (or depleted in the case of a positively charged surface) in the electrical double layer with the resulting consequence that the apparent \( pK_a \) of ionizable groups that suffer acid-base equilibria close to the surface of the bilayer is different from what it would be in bulk aqueous solution. The value of this apparent \( pK_a \) is dependent on the ionic strength of the bulk aqueous electrolyte solution and approaches the value of the \( pK_a \) in bulk solution as the ionic strength is raised.

The transmembrane potential is the difference in the electrical potentials of the two bulk aqueous phases separated by the bilayer (for a detailed treatment see Reference 67). The lipid bilayer may be viewed as an electrical capacitor in which the electrical potential works on the flux of the ions in a direction that is contrary to that of the diffusion potential. At equilibrium, the electrical potential and the diffusion potential are exactly balanced so that no net flux of ions occurs. The resulting transmembrane potential with respect to a given ion, \( \psi_{\text{membrane}} \), is given by the Nernst equation:

\[
\Delta \psi_{\text{membrane}} = \psi_0 - \psi_1 = \frac{RT}{F} \ln \left[ \frac{[\text{co-ion}]}{[\text{counter-ion}]} \right]
\]
where $R$ is the gas constant, $F$ is the Faraday constant, $C_i$ are the concentrations of the ion $N_i$, and the subscripts (1) and (2) represent the electrolyte solutions on the two sides of the bilayer. A biological membrane is bathed in electrolyte solutions that have different ions that permeate the bilayer, each at its own rate. Thus, if the bilayer is bathed in aqueous solutions of NaCl, then the concentrations of both Na$^+$ and Cl$^-$ on both sides of the bilayer and the respective permeation coefficients, $P_{Na^+}$ and $P_{Cl^-}$, must be taken into account to account for the global transmembrane potential, which is done in the Goldman-Hodgkin-Katz equation:

$$
\Delta \Psi = \frac{RT}{F} \left( \frac{P_{Na^+}c_{Na^+}}{c_{Na^+}} + \frac{P_{Cl^-}c_{Cl^-}}{c_{Cl^-}} \right)
$$

where the subscripts (1) and (2) refer, as before, to the electrolyte solutions on the two sides of the bilayer.

In living cells, ion channels and energy-dependent pumping mechanisms (with more or less specificity) facilitate or catalyze the permeation of certain ions in one direction or the other across the lipid bilayers that constitute the cell membranes. This mechanism makes the permeation of certain ions several orders of magnitude faster than the permeation of others; which thereby accentuates the transmembrane potentials discussed above. These processes are particularly important in the creation and the transmission of the action potential in nerves.

Lipid bilayer asymmetry in biological membranes and its possible causes

Most lipid bilayers prepared by hydration of lipids in the laboratory (the so-called model systems) are bilaterally symmetric at equilibrium (i.e., the chemical composition with regard to the lipids of each monolayer that constitutes these bilayers is identical). That is not the case in the lipid bilayers of biological membranes that are known to be compositionally asymmetric in all examined cases. A important contribution to the transverse asymmetry observed in biological membranes comes from the vectorial nature of lipid biosynthesis in cell organelles such as the endoplasmic reticulum and the Golgi compartments. Other contributions certainly come from the relatively slow transmembrane translocation (see above) of lipids in lipid bilayers, the asymmetric chemical composition of the aqueous compartments on the two sides of the membrane, and from the lipid shape-dependent (see discussion on the packing parameter above) spontaneous curvature of lipid monolayers. When all contributions are insufficient, the living cell seems to maintain lipid asymmetry in its membranes enzymatically using the so-called “flipases” that may or may not be dependent on metabolic energy (68, 69). Transverse translocation of most polar lipids across a lipid bilayer involves an intermediate “transesterification state,” in which the polar portion of the lipid molecule lies in the highly apolar bilayer midplane, with an activation free energy ($\Delta G^{\circ}$) on the order of 100 kJ mol$^{-1}$. It would seem, therefore, that transverse asymmetry is of paramount importance in cell membrane structure and physiology. It will be an important challenge from the perspective of Cell Biology and Biochemistry to understand why. From the perspective of lipid bilayer Physics and Chemistry it will be a nontrivial challenge to make relatively stable compositionally asymmetric lipid bilayers.

References

Lipid Bilayers, Properties of


Lipid Bilayers, Properties of


Further Reading


See Also

GENERAL/INTRODUCTION
Analytical Chemistry in Biology
Computational Chemistry in Biology
Lipid Micelle Organization
Membrane Architecture Function Relationships
Membrane Assembly in Living Systems
Membrane Fusion, Mechanisms of
Membranes: Topics in Chemical Biology
Modelling and Computation: Lipid Assemblies
Nature: A Model System for Chemistry
Organic Chemistry in Biology
Physical Chemistry in Biology
Self-Assembling Systems, Applications of
Self-Assembling Systems, Synthetic Applications of
Self Assembly, Bienergetics of
Self Assembly, Computation and Modeling of Therapeutics, Lipid Membrane Applications
Vesicle Formation, Dynamics of Water, Properties of

CHEMICAL COMPOSITION
Chromatography of Lipids
Lipids from Whole Cells and Tissues, Extraction of Fluorescence Techniques: Lipids
GC-MS of Lipids

Glycolipids, Synthesis of Imaging Techniques: Lipids Labeling Techniques: Lipids Lipidomics Lipids, Semi-synthetic Steroid and Triterpene Biosynthesis

PHYSICAL PROPERTIES: PHASE BEHAVIOR AND PHASE TRANSITIONS
Calorimetry: Overview of Applications in Chemical Biology Lipids, Phase Transitions of

PHYSICAL PROPERTIES: MOLECULAR AND SUPRAMOLECULAR DIMENSIONS IN LIPID BILAYERS
Lipid Orientation, Determination of NMR of Lipids NMR: Overview of Application in Chemical Biology

PHYSICAL PROPERTIES: PHASE COEXISTENCE IN LIPID BILAYERS
Lipid Domains, Chemistry of Lipid Rafts

PHYSICAL PROPERTIES: ELASTO-MECHANICAL FORCES IN LIPID BILAYERS
Cell Membranes, Dynamics of Membranes, Fluidity of

PHYSICAL PROPERTIES: MOLECULAR DYNAMICS, DIFFUSION, TRANS-BILAYER TRANSLLOCATION, AND LIPID EXCHANGE IN BILAYERS
Cell Membranes, Dynamics of Lipid Homeostasis, Chemistry of Membrane Trafficking

PHYSICAL PROPERTIES: LIPID BILAYERS AS PERMEABILITY BARRIERS
Ion Transport Water Channels

PHYSICAL PROPERTIES: ELECTRICAL PROPERTIES OF LIPID BILAYERS
Brain Development, Neurochemistry of Electrochemical gradients

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Lipids, Chemical Diversity of

Irina A. Guschina and John L. Harwood, School of Biosciences, Cardiff University, Wales, United Kingdom
doi: 10.1002/9780470048672.wecb292

Lipids represent some of the most complex biologic molecules, and their diversity is crucial for their cellular functions. The wide range of chemical and physical properties of different lipids determines a variety of roles for these compounds in biologic processes. According to their structures, lipids can be divided into two main groups: the nonpolar lipids (acylglycerols, sterols, free fatty acids, hydrocarbons, alcohols, wax, and steryl esters) and the polar lipids (phosphoglycerides, glycosylglycerides, and sphingolipids). Triacylglycerols act as energy stores and metabolic fuels. Polar lipids and sterols are important structural components of cell membranes where they may have many diverse functions. Arranged as bilayers, they establish permeability barriers for cells and organelles and provide a microenvironment for membrane-associated proteins as well as participate directly in metabolism and a multitude of membrane fusion events (1). Waxes as surface coverings are integral to water balance and protect organisms from noxious environmental conditions. In addition to a structural role, lipids serve as important intermediates in cell signaling pathways (e.g., sphingolipids, inositol lipids, phosphatidic acid, lysophospholipids, oxidative products) and play a role in mediating cellular responses to the environment (2). In this article, we will summarize our knowledge on the structural diversity of lipids that provide information on the various lipid classes and the fatty acids that occur naturally.

Glycerolipids

Acylglycerols (glycerides)

The basic structure of glycerolipids is a glycerol (propane-1,2,3-triol) backbone to which the hydrophobic acyl groups (fatty acids) are esterified. Monoacylglycerols and diacylglycerols are fatty acid monoesters and diesters of glycerol, respectively. These partial glycerides are important intermediates in metabolism. Moreover, sn-1,2-diacylglycerols are important in animal tissues where they function as signaling molecules in many cellular processes. Triacylglycerols are fatty acid triesters of glycerol. They are accumulated as storage products and are the major components of most natural fats and oils.

Glycerol-derived ethers

Alkyl ethers and alk-1-enyl ethers (or plasmalogens) are two general types of glycerol ethers. Mono- and dialkyl ethers have been reported to occur naturally. Examples include butyl alcohol (sn-1-Octadecylglycerol) and chymyl alcohol (sn-1,0-hexadecyl alcohol) as monoalkyl ethers and diphytanoylglycerol (sn-2,3-di-0-tetramethylhexadecylglycerol) as a diacyl ether. Although trialkyl ethers have not yet been found in nature, they have been synthesized with various combinations of saturated and unsaturated chains. In addition, monacyl and diacyl derivatives of alkyl ethers are reported for some natural sources. Similar to alkyl forms, the mono- and dialk-1-enyl ethers of glycerol have been identified in natural samples, whereas no trialk-1-enyl ethers have been isolated so far. A cyclated (mono- and diacyl derivatives) alk-1-enyl ethers (or neutral plasmalogens) have been also reported (2).

Glycerophospholipids

The structure of glycerophospholipids is characterized by a 1,2-diacyl-3-phospho-sn-glycerol, or phosphatidyl moiety, and a variable headgroup linked to the phosphate. Major phosphoglycerides found in nature are phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylglycerol (PG), phosphatidylinosline, phosphatidylinositol (PI), and diphaspho-
lipids with a high proportion (up to 70%) in the outer monolayer. PG is an only phospholipid present in significant quantities in (3).

release of N-acylethanolamines after phospholipase D action involved into the endocannabinoid signalling system through may be present in significant amounts. These compounds are

phospho-(3 to contain acylphosphatidylglycerol (1,2-diacyl-

sn)

3-

trans-

Fig. 1

PE (trivial name "cephalin") also has a net neutral charge (Fig. 1). PE is widespread and usually the second most abundant phospholipid in animal and plant membranes. It is also the main lipid component of microbial membranes. In animal tissues, phosphatidylethanolamine may exist in diacyl, alkyacyl and alkkenylacyl forms. Moreover, animal phosphatidylethanolamine usually contains higher levels of arachidonic and docosahexaenoic acids in comparison with the other zwitterionic phospholipid, PC. The partly methylated derivatives of PE (phosphatidyln-methyl-ethanolamine, phosphatidyl-N-dimethyl-ethanolamine) are found in small amounts in many organisms. These derivatives are metabolic intermediates in the conversion of PE into PC. In some tissues, N-acylated derivatives of PE may be present in significant amounts. These compounds are involved into the endocannabinoid signalling system through release of N-acylphosphatidylethanolamines after phospholipase D action (3).

PG is negatively charged (Fig. 1). It is the major phosphoglyceride in photosynthetic tissues and many bacteria. In plants, PG is an only phospholipid present in significant quantities in the thylakoid membranes where it can comprise 10% of the total lipids with a high proportion (up to 70%) in the outer monolayer. A n-4 fatty acid, δ5-trans-hexadecenoic acid (16:1(9)), is found in all eukaryotic photosynthetic organisms, especially extenifying the sn-2 position of PG (4). It is interesting to note that the stereochemical configuration of a double bond and its Δ2 position are very unusual for natural fatty acids. PG-16:1(3) is thought to play an essential role in photosynthetic membranes.

Although PG is only a minor phospholipid in animal tissues (1-2% of total phospholipids), it can be a second abundant phospholipid in some lung surfactants at up to 11% of the total (5).

For bacteria, PG is present in almost all types. In Esherichia coli, for example, PG accounts for 20% of its membrane lipids. The diacyl form of PG is more common in many bacteria, but some may contain the alkylacyl- and alkkenylacyl derivatives as predominant forms. Several prokaryotic species, including Salmonella typhimurium and E. coli, and parasitic protozoa, such as Trichomonas vaginalis and T. foetus, have been found to contain acylphosphatidylglycerol (1,2-diacyl-sn-glycero-3-phospho-(3-acyl)-sn-glycerol) in appreciable amounts (6). For example, Corynebacterium ammoniagenes contained about 30% of this lipid with oleate as a dominant fatty acid on the head group of glycerol (7). A fully acylated phosphatidylglycerol (bis-phosphatic acid or phosphatidyl(diacylglycerol) and its plasmalogen analogs have been also identified in a marine bacterium MB 45. In some bacteria, an amino acid (lysine, ornithine, arginine, or alanine) may be attached to the 3'-hydroxy of the base glycerol to form an O-amiinoacylphosphatidylglyceroI. Two other unusual derivatives of phosphatidylglycerol have been isolated from the primitive organisms, the Haloar-chea. They were identified as phosphatidylglycerol sulfate and phosphatidylglycerol phosphate methyl ester.

Diposphatidylglycerol (trivial name cardiolipin) has a unique dimeric structure with four acyl groups and two negative charges (Fig. 1). It is common in bacteria, and it can be found in the inner mitochondrial membranes of eukaryotes (i.e., those membranes that generate an electrochemical potential for substrate transport and ATP synthesis).

Phosphatidylserine also has a net negative charge. It is a widespread but minor lipid in eukaryotes, accounting usually for less than 10% of the total phospholipids. Its greatest con-centration has been noted for myelin from brain tissue. Phos-phatidylserine is concentrated in the inner monolayer of the plasma membrane and the other cellular membranes (Fig. 1).

Phosphatidylinositol is another widespread and minor lipid with a negative charge. It is an important lipid, both as a membrane constituent and in various phosphorylated forms as key molecules in signalling processes in eukaryotes (Fig. 1). The phosphatidylinositol phosphates are found with phosphates attached at any position of the inositol ring. However, PI-4-phosphate, PI-4,5-bisphosphate, and their 3-phosphates are the most important. PI also forms part of the glycophosphatidylinositol (GPI) anchor for certain surface proteins.

The acyl structure diversity

Monacyl derivatives of phosphoglycerides, lysy-derivatives, are present in small amounts in most tissues, but their occurrence in large levels usually indicates lipid degradation before or during lipid extraction. Only cereals grains contain lysophosphatidylcholine and lysophosphatidylethanolamine as the ma-jor phospholipids (as starch inclusion compounds). Moreover, lysosphatidyglycerols, especially lysophosphatic acid, are impor-tant signaling molecules (8).

Plasmalogens are monacyl monoalk-1-enyl ether forms of phospholipids. The most common forms of plasmalogens are the choline, ethanolamine, and serine derivatives. They are present in most animal tissues, especially in the mammalian brain. Plasmalogens are structural components of membranes. It has been proposed that plasmalogens protect membranes against oxidative stress (9).

Platelet-activating factor or 1-alkyl-2-acetyl-sn-glycero-3-phosphocholine is an ether analog of phosphatidylcholine. This biologically active lipid is of great current interest. Initially, it was found to affect aggregation of platelets and to induce a hyper-tensive response at very low concentrations. Recent research has revealed its role as a mediator of inflammation and of other physiologic effects on many different types of cells.

A rhodobacteria contain large amounts of unique lipids, which are based on 2,3-dialkyl-3-phosphoglycerides (DPG) and 2,3-dialkyl-4-phosphoglycerides (DAGP). DPG is the major phospholipid and DAGP is the major phosphoglyceride in these organisms. DPG and DAGP are present in large levels usually indicates lipid degradation before or during lipid extraction. Only cereals grains contain lysophosphatidylcholine and lysophosphatidylethanolamine as the ma-jor phospholipids (as starch inclusion compounds). Moreover, lysosphatidyglycerols, especially lysophosphatic acid, are impor-tant signaling molecules (8).

Platelet-activating factor or 1-alkyl-2-acetyl-sn-glycero-3-phosphocholine is an ether analog of phosphatidylcholine. This biologically active lipid is of great current interest. Initially, it was found to affect aggregation of platelets and to induce a hyper-tensive response at very low concentrations. Recent research has revealed its role as a mediator of inflammation and of other physiologic effects on many different types of cells.

A rhodobacteria contain large amounts of unique lipids, which are based on 2,3-dialkyl-3-phosphoglycerides (DPG) and 2,3-dialkyl-4-phosphoglycerides (DAGP). DPG is the major phospholipid and DAGP is the major phosphoglyceride in these organisms. DPG and DAGP are present in large levels usually indicates lipid degradation before or during lipid extraction. Only cereals grains contain lysophosphatidylcholine and lysophosphatidylethanolamine as the ma-jor phospholipids (as starch inclusion compounds). Moreover, lysosphatidyglycerols, especially lysophosphatic acid, are impor-tant signaling molecules (8).

Platelet-activating factor or 1-alkyl-2-acetyl-sn-glycero-3-phosphocholine is an ether analog of phosphatidylcholine. This biologically active lipid is of great current interest. Initially, it was found to affect aggregation of platelets and to induce a hyper-tensive response at very low concentrations. Recent research has revealed its role as a mediator of inflammation and of other physiologic effects on many different types of cells.

A rhodobacteria contain large amounts of unique lipids, which are based on 2,3-dialkyl-3-phosphoglycerides (DPG) and 2,3-dialkyl-4-phosphoglycerides (DAGP). DPG is the major phospholipid and DAGP is the major phosphoglyceride in these organisms. DPG and DAGP are present in large levels usually indicates lipid degradation before or during lipid extraction. Only cereals grains contain lysophosphatidylcholine and lysophosphatidylethanolamine as the ma-jor phospholipids (as starch inclusion compounds). Moreover, lysosphatidyglycerols, especially lysophosphatic acid, are impor-tant signaling molecules (8).

Platelet-activating factor or 1-alkyl-2-acetyl-sn-glycero-3-phosphocholine is an ether analog of phosphatidylcholine. This biologically active lipid is of great current interest. Initially, it was found to affect aggregation of platelets and to induce a hyper-tensive response at very low concentrations. Recent research has revealed its role as a mediator of inflammation and of other physiologic effects on many different types of cells.

A rhodobacteria contain large amounts of unique lipids, which are based on 2,3-dialkyl-3-phosphoglycerides (DPG) and 2,3-dialkyl-4-phosphoglycerides (DAGP). DPG is the major phospholipid and DAGP is the major phosphoglyceride in these organisms. DPG and DAGP are present in large levels usually indicates lipid degradation before or during lipid extraction. Only cereals grains contain lysophosphatidylcholine and lysophosphatidylethanolamine as the ma-jor phospholipids (as starch inclusion compounds). Moreover, lysosphatidyglycerols, especially lysophosphatic acid, are impor-tant signaling molecules (8).

Platelet-activating factor or 1-alkyl-2-acetyl-sn-glycero-3-phosphocholine is an ether analog of phosphatidylcholine. This biologically active lipid is of great current interest. Initially, it was found to affect aggregation of platelets and to induce a hyper-tensive response at very low concentrations. Recent research has revealed its role as a mediator of inflammation and of other physiologic effects on many different types of cells.
Lipids, Chemical Diversity of

- Phosphatidylcholine
- Phosphatidylethanolamine
- Phosphatidylglycerol
- Phosphatidylserine
- Phosphatidylinositol
- Diphosphatidylglycerol (cardiolipin)

Figure 1  Phosphoglycerolipids.
Lipids, Chemical Diversity of

Glycosylglycerides

Glycosylglycerides are characterized by a 1,2-diacyl-sn-glycerol moiety with a monosaccharide attached at the sn-3 position of the glycerol backbone. In photosynthetic bacteria, cyanobacteria, algae, and higher plants, glycosylglycerides (glycolipids) are located predominantly in photosynthetic membranes. The major plastid lipids, galactosylglycerides, are neutral lipids. They contain one or two galactose molecules linked to the sn-3 position of the glycerol that corresponds to 1,2-diacyl-3-O-β-D-galactopyranosyl-sn-glycerol (or monogalactosyldiacylglycerol, MGDG) and 1,2-diacyl-3-O-(1→4)-β-D-galactopyranosyl-sn-glycerol (or digalactosyldiacylglycerol, DGDG) (Fig. 2). MGDG and DGDG are present at 40–55% and 15–35% of the total lipids in thylakoid membranes, respectively. A characteristic feature of these lipids is a very high content of polyunsaturated fatty acids. In addition to these main galactolipids, trigalactosyldiacylglycerols and tetragalactosyldiacylglycerols as well as other homologs with many different sugar combinations (e.g., mannosylglycerols and diglucosyldiacylglycerols with the diglucosyl residue that is linked through a β-1→2 linkage (as in kojibiose) and a galactosyldiacylglycerol have been isolated from some bacterial species. Monomannosyldiacylglycerols have been reported for bacteria, algae, and higher plants (11, 12). For example, in eubacteria, glycosylglycerols with one to three glycosyl units linked to sn-1,2-diacylglycerol were most common, but others with up to five glycosyl units have been also found. Monogalactosyldiacylglycerols and diglucosyldiacylglycerols with the diglucosyl moiety that have an α-1→2 linkage (as in kohioside) and a galactosyldiacylglycerol have been isolated from some Streptococcus species. Monomannosyldiacylglycerols and digalactosyldiacylglycerols have been reported for Micrococcus luteus. In Bacillus megaterium, N-acetylgalactosamine has been found linked to a diacylglycerol. In glycolipids present in some bacteria, a glycerophosphate group is linked to a carbohydrate moiety (11).

Various glycosylglycerols and galactosylglycerides have been also isolated in small amounts from animal tissues (13).

The galactoglycerolipids usually contain a single galactose residue that is linked through a β-glycosidic link between the C-1 of galactose and the C-3 of glycerol. The glycoacylglycerides that have been reported may contain up to eight glucose residues. Moreover, alkylacyl and diacyl lipids as well as sulfated forms have been identified (13).

Another class of glycosylglycerides, which is present in appreciable amounts in both photosynthetic and in nonphotosynthetic species, is the plant sulfolipid sulfoquinovosyldiacylglycerol, or 1,2-diacyl-3-O-(6-deoxy-6-sulfoglucopyranosyl)-sn-glycerol (SQDG) (Fig. 2). This lipid is unusual because of its sulfonic acid linkage. It consists of monogalactosyldiacylglycerol with a sulfonic acid in position 6 of the monosaccharide moiety. The sulfoglycosidic moiety (6-deoxy-6-sulfoglucoside) is described as sulfoquinovosyl. The sulfonic residue carries a full negative charge at physiologic pH. In the unicellular alga Chlamydomonas reinhardtii, an acylated derivative of this sulfolipid, 2-O-acyl-sulfoquinovosyldiacylglycerol has been found. Sulfoquinovosylglycerol has been isolated from the marine red alga Gracilaria verrucosa. For a review of algal lipids, see Reference 12.

In addition to sulfoquinovosyldiacylglycerol, the marine diatom Nitzschia alba has been found to contain a sulfonium analog of phosphatidylcholine, phosphatidylsulfatene (14). This lipid has two methyly groups attached to the sulphur atom, and it substitutes completely for phosphatidylcholine in Nitzschia alba. In other marine diatoms and algae, both lipids have been found (15).

Several lipids conjugated to taurine (2-amino ethylsulfonic acid, H2NCH2CH2SO3H) have been reported for many organisms. So, membrane lipids of certain bacteria and algae contain a taurine linked to diacylglycerol. In a seawater bacterium Hyphomonas jannaschiana, which is unique because of an absence of phospholipids, a taurglycolipid, 1,2-diacylglycerol-3-glucuronylglycerol has been described as sulfoquinovosyl. The sulfonic residue carries a full negative charge at physiologic pH. In the unicellular alga Chlamydomonas reinhardtii, an acylated derivative of this sulfolipid, 2-O-acyl-sulfoquinovosyldiacylglycerol has been found. Sulfoquinovosylglycerol has been isolated from the marine red alga Gracilaria verrucosa. For a review of algal lipids, see Reference 12.

In addition to sulfoquinovosyldiacylglycerol, the marine diatom Nitzschia alba has been found to contain a sulfonium analog of phosphatidylcholine, phosphatidylsulfatene (14). This lipid has two methyly groups attached to the sulphur atom, and it substitutes completely for phosphatidylcholine in Nitzschia alba. In other marine diatoms and algae, both lipids have been found (15).

Several lipids conjugated to taurine (2-amino ethylsulfonic acid, H2NCH2CH2SO3H) have been reported for many organisms. So, membrane lipids of certain bacteria and algae contain a taurine linked to diacylglycerol. In a seawater bacterium Hyphomonas jannaschiana, which is unique because of an absence of phospholipids, a taurglycolipid, 1,2-diacylglycerol-3-glucuronylglycerol has been described as sulfoquinovosyl. The sulfonic residue carries a full negative charge at physiologic pH. In the unicellular alga Chlamydomonas reinhardtii, an acylated derivative of this sulfolipid, 2-O-acyl-sulfoquinovosyldiacylglycerol has been found. Sulfoquinovosylglycerol has been isolated from the marine red alga Gracilaria verrucosa. For a review of algal lipids, see Reference 12.

Various glycosylglycerols and galactosylglycerides have been also isolated in small amounts from animal tissues (13).

The galactoglycerolipids usually contain a single galactose residue that is linked through a β-glycosidic link between the C-1 of galactose and the C-3 of glycerol. The glycoacylglycerides that have been reported may contain up to eight glucose residues. Moreover, alkylacyl and diacyl lipids as well as sulfated forms have been identified (13).

Another class of glycosylglycerides, which is present in appreciable amounts in both photosynthetic and in nonphotosynthetic species, is the plant sulfolipid sulfoquinovosyldiacylglycerol, or 1,2-diacyl-3-O-(6-deoxy-6-sulfoglucopyranosyl)-sn-glycerol (SQDG) (Fig. 2). This lipid is unusual because of its sulfonic acid linkage. It consists of monogalactosyldiacylglycerol with a sulfonic acid in position 6 of the monosaccharide moiety. The sulfoglycosidic moiety (6-deoxy-6-sulfoglucoside) is described as sulfoquinovosyl. The sulfonic residue carries a full negative charge at physiologic pH. In the unicellular alga Chlamydomonas reinhardtii, an acylated derivative of this sulfolipid, 2-O-acyl-sulfoquinovosyldiacylglycerol has been found. Sulfoquinovosylglycerol has been isolated from the marine red alga Gracilaria verrucosa. For a review of algal lipids, see Reference 12.

In addition to sulfoquinovosyldiacylglycerol, the marine diatom Nitzschia alba has been found to contain a sulfonium analog of phosphatidylcholine, phosphatidylsulfatene (14). This lipid has two methyly groups attached to the sulphur atom, and it substitutes completely for phosphatidylcholine in Nitzschia alba. In other marine diatoms and algae, both lipids have been found (15).

Several lipids conjugated to taurine (2-amino ethylsulfonic acid, H2NCH2CH2SO3H) have been reported for many organisms. So, membrane lipids of certain bacteria and algae contain a taurine linked to diacylglycerol. In a seawater bacterium Hyphomonas jannaschiana, which is unique because of an absence of phospholipids, a taurglycolipid, 1,2-diacylglycerol-3-glucuronylglycerol has been described as sulfoquinovosyl. The sulfonic residue carries a full negative charge at physiologic pH. In the unicellular alga Chlamydomonas reinhardtii, an acylated derivative of this sulfolipid, 2-O-acyl-sulfoquinovosyldiacylglycerol has been found. Sulfoquinovosylglycerol has been isolated from the marine red alga Gracilaria verrucosa. For a review of algal lipids, see Reference 12.
has been isolated from extremely halophilic bacteria (11), and a glycolipid sulfate, 2,3,6′-tetraacetyl-α,α-tetrahalose-2′-sulfate has been isolated from Mycobacterium tuberculosis (16). For a review on the isolation, characterization, and biochemistry of taurolipids, see Reference 17.

**Betaine lipids**

Betaine lipids contain a betaine moiety as a polar group linked to the sn-3 position of glycerol by an ether bond. No phosphorus or carbohydrate group exists in betaine lipids. Three types of betaine lipids have been identified: 1,2-diacylglycerol-3-0-4′-(N,N,N-trimethyl)homoserine (DGTS), 1,2-diacylglycerol-3-0-2′-(hydroxymethyl)-(N,N,N-trimethyl)-β-alanine (DGTA), and 1,2-diacylglycerol-3-0-carboxy-(hydroxymethyl)-choline (DGCC) (Fig. 3) (18). These betaine lipids are all zwitterionic at pH 7 because they have a positively charged trimethylammonium group and a negatively charged carboxyl group (Fig. 3). Betaine lipids are not found in higher plants, either gymnosperms or angiosperms, but they are distributed widely in photosynthetic bacteria, algae, ferns, bryophytes, lichens, some fungi, and protozoans. Based on an obvious structural similarity between betaine lipids and phosphatidylcholine and on their taxonomic distribution (namely, their reciprocal relationship in many species), it has been suggested that betaine lipids, especially DGTS, are more primitive lipids evolutionarily that, in the lower plants, play the same functions in membranes that PC does in higher plants and animals (19).

**Diol Lipids**

Diol lipids have been reported only recently when techniques for their structure elucidation were developed. Small quantities of diol lipids, such as diaclylpropane-1,3-diol, diesters of butane-1,3-diol, and butane-1,4-diol are found in mammalian and fish liver, mammalian adipose tissues, egg yolk, corn seed, and yeast (20). From Actinomyces olivaceus, mixed acyl- and alk-1-enyl derivatives of simple diols have been isolated (21). An acylated diol phospholipid has been produced by the yeast Lypomyces starkeyi when grown on propane-1,2-diol (22).

**Sphingolipids**

Sphingolipids are lipids that contain sphingosine (trans-D-erythro-1,3-dihydroxy-2-amino-4-octadecene) or a related CH₂ OOCR' CH R"COO CH₂ O CH₂ CH₂ CH₂ N(CH₃)₃ COO⁻ 1,2-diacylglycerol-3-0-4′-(N,N,N-trimethyl)-homoserine (DGTS) CH₃-DOCOR' R"COO-CH₂-O-CH₂-O-CH₂-O-N(CH₃)₃ COO⁻ 1,2-diacylglycerol-3-0-2′-(hydroxymethyl)-(N,N,N-trimethyl)-β-alanine (DGTA) CH₃-DOCOR' R"COO-CH₂-O-CH₂-O-CH₂-O-N(CH₃)₃ COO⁻ 1,2-diacylglycerol-3-0-carboxy-(hydroxymethyl)-choline (DGCC)

Figure 3 Betaine lipids.

Sphingomyelin

Figure 4 Sphingomyelin.
aminoglycerolipids are a group of very complex lipids with a range of diverse functions. It has recently become apparent that sphingolipids are involved in many human diseases including diabetes, cancers, neurologic syndromes, infections, and diseases of the cardiovascular and respiratory systems (23). Moreover, an increased interest in these molecules results from the involvement of sphingolipids in the segregation/compartmentalization of membranes by self-organizing functional lipid microdomains, known as lipid rafts. Sphingolipids are located exclusively in the outer (exoplasmic) leaflet of the plasma membrane bilayer. In contrast, raft glycosphingolipids largely prefer the inner (cytoplasmic) leaflet, and cholesterol is most likely abundant in both leaflets. The lipid raft model is based on the assumption that sphingomyelin and other sphingolipids together with cholesterol are not distributed evenly in the plasma membrane, but rather they assemble to specific microdomains or raft. They act to compartmentalize and to separate membrane proteins with different biochemical functions. It is believed that up to 50% of the plasma membrane may consist of such rafts, and rafts formation has been shown to be crucial in modulation of signaling events in cells (24).

Sphingophospholipids and glycosphingolipid phospholipids

Sphingomyelin (or ceramide phosphorylcholine) is the most common lipid in this class, and is the phosphorylcholine ester of an N-acylsphingosine (or ceramide) (Fig. 4). It is a major lipid in certain membranes in animal tissues. Although it is usually lower in concentration than phosphatidylcholine, it can comprise about 50% of the lipids in certain tissues. Sphingomyelin replaces phosphatidylcholine in erythrocytes of most mammalian animals entirely. Sphingomyelin is probably absent from microorganisms and plants.

In the plasma of the newborn pig and infant, 3-O-acyl-D-erythro-sphingomyelin has been detected. In this lipid, position 3 of the sphingosine residue is esterified by an additional fatty acid.

In addition, lypo-sphingomyelins and sphingosines phosphorylcholine have been reported to be present in tissues in small amounts. They are involved in some signaling processes in the cell in a manner similar to sphingosine-1-phosphate (25, 26). Sphingosine-1-phosphate has been found in insects, yeasts, and plants; its function as an important cellular messenger has been shown in many studies. It regulates calcium mobilization inside the cell as well as cell growth and proliferation in response to several stress factors. Ceramide, phosphorylinositol, or myo-inositol (1,0)-phosphoryl (1,0)-ceramide, the sphingolipid analog of phosphatidyl myoinositol, is a major component of sphingolipids in yeasts (e.g., Saccharomyces cerevisiae). In these organisms, two other related inositol-containing sphingolipids, mannosyl-Ceramide and mannosylinositolphosphorylceramide, have been also found. Ceramide phosphorylethanolamine, the sphingolipid analogue of phosphatidylethanolamine, has been reported for the first time in the h outfly, Musca domestica. Shortly after, it was identified in the bacteria Bacteroides ruminicola and B. melanogenicus as well as in some protozoa, snails, marine bivalves, insects, chicken liver, and rat liver (27). In recent studies, this lipid has been isolated from three species of oomycete plant pathogens (28) and from several species of Sphingobacterium (29).

Ceramide phosphorylglycerols are known as membrane components in anaerobic bacteria of the genus Bacteroides and in the oral pathogen Porphyromonas gingivalis (30). A sphingoid analog of phosphatidic acid, ceramide-1-phosphate, has recently emerged as potent bioactive agent. Recent studies defined new biologic functions for this lipid related to control of numerous aspects of cell physiology, which include cell survival and mammalian inflammatory responses (26).

The first glycolipids identified in some plants were glycosphingolipids with a phytosphingosine and/or dehydro sphingosine backbone that contains both inositol and sphingosine. They have been named “phytosphingolsphingolipids” and more than 20 molecular forms have been identified with glucuronic acid, glucosamine, and many other carbohydrate moieties. The analysis of such compounds has been discussed thoroughly in (31).

From the bacterium Sphingobacterium spiritivorum, ceramide phosphorylmannose has been recently reported for the first time. A rhizobacteria has been shown to contain various phosphoglycolipids (11).

A another type of glycosphingolipid phospholipids are glycosphin golipids that are phosphorylated even more (i.e., where the ceramide is linked directly to sugar moieties not via phosphate group). The earthworm, Pheretima hilgendorfi, has been shown to contain such lipids, namely cholinephosphoryl-6Gal-1-Cer and cholinephosphoryl-6Gal-1-Der. In addition, triglycosylphosphorylporphospholipids with either a terminal mannose or a galactose linked to phosphorylcholine were also present in this species. Some filamentous fungi contain glycosphin golipids (e.g., phosphocholine-containing glycosyl inositolphosphorylceramides) (30).

Ceramides

Ceramides are formed by attachment of a fatty acid (linked by an amide bond) to the amino group of sphingosine or other related amino alcohol. The most commonly found sphingosyl alcohols are D-erythro-sphinganine (dihydrosphingosine, sphinganine), D-erythro-sphingosine (sphingosine, 4-sphingenine), C2D-dihydrosphingosine (icosahexanin), and 4-hydroxy-D-erythro-sphinganine (phytosphingosine, 4-hydroxydihydrosphingosine). A thorough ceramides are found usually at trace levels in tissues, they play important biologic roles (1). Moreover, they are synthesized as the key intermediates in the biosynthesis of sphingolipids, in which glycosylceramide is a key precursor of complex oligoglycolipids (Table 1).

Cerebrosides

Cerebrosides are glycosides of N-acyl long-chain bases (ceramides). Galactosyl and glucose are the monosaccharides usually found. Galactosylceramide is the principal glycosphin golipid in brain tissue (2). In general, galactosylceramides are present in all nervous tissues, and account for up to 2%
and 2-hydroxynervonic (\(-\text{OH} \ C_{24}:1\)) (2).

\(\alpha\)-OH \ C_{24}:0\), also present. Typical fatty acids reported are behenic (C\(_{22}:0\)), through the amino group. In addition, odd chain fatty acids may 2-hydroxy fatty acids, which are linked to the sphingosine base moiety, plasmal-o-galactosylceramide, has been identified (32).

Ceramide with a long-chain cyclic acetal at the carbohydrate residues linked to a ceramide base) are important components of tissues in small amounts. In the equine brain, a galactosyl-ceramide with 1-

Most compounds are located on the plasma membrane facing into the extracellular space. The structure of carbohydrate units may determine the function of oligoglycosphingolipids, and one of most important of these is in the immune defense system (33). In plants, two series of oligoglycosylceramides with either mannosyl or galactosyl residues are synthesized by elongation of glucosylceramide. They are located in the endoplasmic reticulum, Golgi, tonoplast, and plasma membranes. Their functions in plants are poorly understood (34).

**Gangliosides**

Gangliosides are glycosphingolipids that contain a oligoglycosylceramide to which a sialic acid (\(\alpha\)-D-N-acetylneuraminic acid) residue is attached via glycosidic linkages to one or more of the monosaccharide units (i.e., via the hydroxyl group on position 2) or to another sialic acid residue. The level of these lipids is around 6% of total lipids in the brain tissues. Neuronal membranes contain 10-12 gangliosides of the total lipid content (20-25% of the outer layer). They are also present at low levels in all animal tissues but not found outside of the animal kingdom.

Gangliosides are involved in control of cell growth and differentiation as well as cell interactions (33). They also play a key role in the immune defense systems. Many reactions are mediated through the location of these molecules in membrane rafts.

**Glycosphingolipid sulfates**

Glycosphingolipids with a sulfate ester group attached to the sugar moiety are termed glycosphingolipid sulfates (sulfatides or sulfoglycosphingolipids). The distribution of these compounds in mammalian tissues as well as their fatty acid composition and base composition are usually similar to those of cerebrosides. 3'-Sulfo-galactosylceramide is an example of one of the more abundant glycolipid compounds of brain myelin. It has been also found in many other organs, especially in the

of the dry weight of gray matter and 12% of white matter. Oligodendrocytes contain large amounts of these lipids.

Glucosylceramides are considered as characteristic components of plants but also found in low amounts in animal tissues, such as spleen, nervous tissues, and erythrocytes. In plants, they are abundant in photosynthetic tissues (31). From nonphotosynthetic tissues, monoglycosylceramides that contain a \(\beta\)-D-mannopyranosyl residue have been isolated.

In addition, fucosylceramide has been reported in a colon carcinoma, a xylose-containing ceramide was identified in an avian salt gland, and glycosylceramides with mannose and glucuronic or galacturonic acids as sugar moieties have been isolated from some microorganisms. \(\alpha\)-Acylated (with fatty acids in various positions of the sugar moiety) glucosylceramides and galactosylceramides have been also found in brain tissues in small amounts. In the equine brain, a galactosylceramide with a long-chain cyclic acetal at the carbohydrate moiety, plasmal-o-galactosylceramide, has been identified (32).

Galactosylceramides usually contain large amounts of 2-hydroxy fatty acids, which are linked to the sphingosine base through the amino group. In addition, odd chain fatty acids may also present. Typical fatty acids reported are behenic (C\(_{22}:0\)), \(\alpha\)-Acylated (with fatty acids in various positions of the sugar moiety) glucosylceramides and galactosylceramides have been also found in brain tissues in small amounts. In the equine brain, a galactosylceramide with a long-chain cyclic acetal at the carbohydrate moiety, plasmal-o-galactosylceramide, has been identified (32).

Galactosylceramides usually contain large amounts of 2-hydroxy fatty acids, which are linked to the sphingosine base through the amino group. In addition, odd chain fatty acids may also present. Typical fatty acids reported are behenic (C\(_{22}:0\)), \(\alpha\)-Acylated (with fatty acids in various positions of the sugar moiety) glucosylceramides and galactosylceramides have been also found in brain tissues in small amounts. In the equine brain, a galactosylceramide with a long-chain cyclic acetal at the carbohydrate moiety, plasmal-o-galactosylceramide, has been identified (32).

Galactosylceramides usually contain large amounts of 2-hydroxy fatty acids, which are linked to the sphingosine base through the amino group. In addition, odd chain fatty acids may also present. Typical fatty acids reported are behenic (C\(_{22}:0\)), \(\alpha\)-Acylated (with fatty acids in various positions of the sugar moiety) glucosylceramides and galactosylceramides have been also found in brain tissues in small amounts. In the equine brain, a galactosylceramide with a long-chain cyclic acetal at the carbohydrate moiety, plasmal-o-galactosylceramide, has been identified (32).

Table 1 Classification of oligoglycolipids

<table>
<thead>
<tr>
<th>Root name</th>
<th>Abbreviation</th>
<th>Partial structure*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ganglio</td>
<td>Gg</td>
<td>Gal(\beta)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Lacto</td>
<td>Lc</td>
<td>Gal(\alpha)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Neolacto</td>
<td>Nc</td>
<td>Gal(\beta)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Globo</td>
<td>Gb</td>
<td>Gal(\alpha)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Isogloblo</td>
<td>Gg</td>
<td>Gal(\alpha)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Mollu</td>
<td>Mu</td>
<td>Gal(\alpha)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
<tr>
<td>Arthro</td>
<td>At</td>
<td>Gal(\alpha)-GalNAcGal(\beta)-Gal(\beta)-Cer</td>
</tr>
</tbody>
</table>

*Roman numerals show sugar positions in the root structure.

In general, hundreds of different neutral oligoglycosphin-golipids have been identified, and each mammalian organ has a characteristic pattern of neutral ceramides with kidney, lung, spleen, and blood that contain large amounts. In erythrocytes, for example, megaglycolipids with up to 50 sugar groups are present, but triglycosylceramides to pentaglycosylceramides are usually more abundant.
kidney. In addition to these "classic" sulfatides, a whole series of sulfatides in organisms that range from mycobacteria to mammals have now been characterized. These include sulfolactosyloceramide and other sulfatides derived from oligosaccharides of the globo series and gangliosides isolated from human kidney, and the ganglioside sulfatides isolated from echinoderms (35).

**Phosphonolipids**

Phosphonolipids are analogs of various phospholipids, which consist of aminoethylphosphonic acid residues, have been found in nature. The aminoethylphosphonic acid moiety may be linked through a phosphorus-carbon bond to either a ceramide or diacylglycerol backbone in phosphonolipids. Ceramide aminoethylphosphonate (or ceramide ciliatine) was the first phosphonolipid analog of sphingosylphosphatides, and has been identified in sea anemones (Fig. 5). Lately, it has also been detected in several mollusks, protists, bacteria, and bovine brain tissue together with its N-methyl derivative. In some bacterial species, phosphonolipids with a 1-hydroxy-2-aminoethane residue linked to the phosphorus moiety have been also detected. Phosphonoglycosphingolipids such as 6-O-(aminoethylphosphono)galactosylceramide (Fig. 5) and its N-methyl analog, have been isolated from some marine invertebrates. The role and biochemical functions of these lipids are still unknown.

The phosphonate form of phosphatidylethanolamine, 1,2-diacyl-sn-glycerol-3-(2'-aminoethyl)phosphonate (Fig. 5), is present in several species of protists as well as in bovine tissues and in human aorta. Its alkylacyl and alkenylacyl forms have also been found. From the freshwater cyanobacteria *Aphanizomenon flos-aquae*, 2-acyloxyethylphosphonate (with unusual biosurfactant properties) has been isolated. Phosphonolipids have been reviewed in References 36 and 37.

**Lipid-Soluble Arsenic Compounds (Arsenolipids)**

Lipid-soluble arsenic compounds were first identified from marine organisms over 30 years ago. By now, more than 100 naturally occurring arsenolipids have been detected in a wide range of organisms, including lichens, fungi, freshwater and marine algae, invertebrates, fishes, plants, and animals (38). The level of accumulation of these lipids may vary from 1.5 to 33.8 µg/g dry weight. Some plants and hyper-accumulators contain more than 1 mmol As g⁻¹ dry weight. Many species of bacteria as well as most algae species examined seem to convert arsenic compounds into their methylated derivatives, such as trimethylarsine (39). Structures, distribution, the possible metabolism of arsenolipids, as well as analytical techniques used for their separation, identification, and quantification are thoroughly reviewed in Reference 38.

**Other Esters**

A group termed "other esters" includes a large variety of functionally diverse lipids. Wax esters are a typical example. The term "wax" is used commonly for esters of long-chain fatty acids with long-chain primary alcohols, but sometimes it is used for the entire mixture of lipids that contain waxes. Esters are present in both plants and animals where they form the water-repellent surface coating (e.g., skin surface of animals and the leaf cuticle). Esters of normal alcohols with monobranched or multibranched fatty acids have been found in the preen glands of birds (2). Complex waxes (in which either the fatty acid or the alcohol component or both has a complex structure) have been isolated from some bacteria. Dicesters...
of phthiocerols (C33–C35 branched-chain diols) with myco-

ceric acids (C29–C32 branched-chain acids) are examples of
these compounds that have been reported for Mycobacterium
spp. (40).

Two types of diester waxes are characteristic of animal
skin-surface lipids. In the first type, acyl derivatives of hydroxy
acids formed by reaction between the carboxyl group of a
fatty acid with the hydroxyl group of a hydroxy acid. In the
second type, waxes consist of an alkane α,β-diol in which both
hydroxyls are esterified with fatty acids (41).

Ester forms are found in the commonly occurring sterols
and vitamin alcohols, such as vitamin A, the D vitamins, and
vitamin E, as well as in carotenoids (flowers and green algae)
and terpenoids (rose petals). Various other esters have been
reported in different bacteria and yeasts where they may occur
as glycoside and amino acid derivatives. For more details, see
Reference 2.

Fatty Acids

Fatty acids are aliphatic monocarboxylic acids. Over 1000
fatty acids are known with different chain length, positions,
configurations, and types of unsaturation. A division of various
functional groups along the aliphatic chain brings more struc-
tural and functional diversity to this group.

Saturated fatty acids

In most natural lipids, straight- or normal-chain saturated com-
ponents (even-numbered) account for up to 10-40% of the total
fatty acids. Straight-chain fatty acids with 16 and 18 carbon
atoms are the most common naturally-occurring compounds in
animals and plants although odd-numbered and even-numbered
homologs from 2C to 36C carbon atoms have been reported for
natural samples (2). In general, fatty acids do not exist as free
fatty acids because their high affinity for many proteins.
Moreover, elevated levels of free (nonesterified) fatty acids are
usually artifacts of cell damage that allow lipases to breakdown
the endogenous acyl lipids. Table 2 shows examples of the most
important straight chain saturated fatty acids together with the
information on their distribution.

Table 2 Some naturally occurring straight chain saturated fatty acids

<table>
<thead>
<tr>
<th>Number of C atoms</th>
<th>Systematic name</th>
<th>Common name</th>
<th>Melting point (°C)</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Ethanoic</td>
<td>Acetic</td>
<td>16.7</td>
<td>Esterified to glycerol in ruminant milk fats, some plant triacylglycerols and as alcohol acetates in many plants. As a component of platelet-activating factor.</td>
</tr>
<tr>
<td>3</td>
<td>Propanoic</td>
<td>Propionic</td>
<td>–22.0</td>
<td>In the rumen.</td>
</tr>
<tr>
<td>4</td>
<td>Butanoic</td>
<td>Butyric</td>
<td>–7.9</td>
<td>In milk fat of ruminants, in the rumen.</td>
</tr>
<tr>
<td>6</td>
<td>Hexanoic</td>
<td>Caproic</td>
<td>–8.0</td>
<td>In milk fat.</td>
</tr>
<tr>
<td>8</td>
<td>Octanoic</td>
<td>Caprylic</td>
<td>12.7</td>
<td>Minor compound of most animal and plant fats, major component of milk and some seed triacylglycerols.</td>
</tr>
<tr>
<td>10</td>
<td>Decanoic</td>
<td>Capric</td>
<td>29.6</td>
<td>Minor compound in many tissues. Major acid in milk and some seed triacylglycerols.</td>
</tr>
<tr>
<td>12</td>
<td>Dodecanoic</td>
<td>Lauric</td>
<td>42.2</td>
<td>Widely occurring, major acid in some seed fats, e.g., coconut oil and palm kernel oil.</td>
</tr>
<tr>
<td>14</td>
<td>Tetradecanoic</td>
<td>Myristic</td>
<td>52.1</td>
<td>Widely occurring, major acid in some seed fats, e.g., coconut oil and palm kernel oil.</td>
</tr>
<tr>
<td>16</td>
<td>Hexadecanoic</td>
<td>Palmitic</td>
<td>60.7</td>
<td>The most common saturated fatty acid in animals, plants and microorganisms.</td>
</tr>
<tr>
<td>18</td>
<td>Octadecanoic</td>
<td>Stearic</td>
<td>69.6</td>
<td>Major fatty acid in animal tissues and some fungi, usually minor in plants, major only in a few plant species (e.g., cocoa butter). Usually minor compounds and only occasionally significant.</td>
</tr>
<tr>
<td>20</td>
<td>Eicosanoic</td>
<td>Arachidic</td>
<td>75.4</td>
<td>Usually minor compounds and only occasionally significant.</td>
</tr>
<tr>
<td>22</td>
<td>Docosanoic</td>
<td>Behenic</td>
<td>80.0</td>
<td>Minor component in some seed triacylglycerols and plant waxes.</td>
</tr>
<tr>
<td>24</td>
<td>Tetracosanoic</td>
<td>Lignoceric</td>
<td>84.2</td>
<td>Widely spread as a constituent of sphingolipids, minor acid in some seed triacylglycerols and plant waxes.</td>
</tr>
<tr>
<td>26</td>
<td>Hexacosanoic</td>
<td>Cerotic</td>
<td>87.7</td>
<td>Usually as component of plant and insect Waxes.</td>
</tr>
<tr>
<td>28</td>
<td>Octacosanoic</td>
<td>Mertonic</td>
<td>90.9</td>
<td>Major acid in some plant waxes.</td>
</tr>
</tbody>
</table>
Lipids: Chemical Diversity of branched-chain fatty acids have been found, and branches other than methyl are present in microbial lipids. Two distinct types, which are often found in bacteria, are the iso-methyl branched fatty acids with the branch point on the penultimate carbon (one from the end) and anteiso-methyl branched fatty acids with the branch point on the ante-penultimate carbon atom (two from the end). Fatty acids with these structures and with 10 to more than 30 carbon atoms have been identified in some bacteria and in other microorganisms. However, branch points are also found in other positions. For example, 10-R-methyloctadecanoic acid (or tuberculostearic acid) (Fig. 6) is a characteristic of the tubercle bacillus and related species. A wide range of mono-methyl branched fatty acids occurs in some mammalian tissues with the uropygial (preen) gland of birds as a major source. In membranes, they have a function (as an alternative to double bonds) in increasing the fluidity of the lipid bilayer (1).

From Mycobacteria and related species, fatty acids with \( \beta \)-hydroxy-\( \alpha \)-alkyl branched structure and high molecular weight (60 to 88 carbons or more) have been isolated. These mycolic acids can contain a variety of functional groups, which include double bonds, cyclopropane rings, methoxy groups, epoxy groups, and keto groups depending on bacterium species.

Monoenoic (Monounsaturated) Fatty Acids

Straight- or normal-chain (even-numbered) monoenoic fatty acids (with one double bond) amount to a significant proportion of the total fatty acids in most natural lipids. The double bond is usually of the cis- or Z-configuration, but some fatty acids with trans- or E-double bonds are found. For example, a major fatty acid esterified to phosphatidylglycerol in the photosynthetic membranes of higher plants and algae is trans-3-hexadecenoic acid. Often the cis bond is present at the \( \Delta 9 \) position.

The most common monoenoines are straight-chain compounds with 16 or 18 carbon atoms. cis-9-Octadecenoic acid (the trivial name oleic acid) is one of the most abundant fatty acids in nature. Table 3 shows examples and the occurrence of the most abundant monoenoic fatty acids.

A cis-bond in a fatty acid introduces a kink (a 60° bend) in the acyl chain with a restriction in motion at that point. Very long chain (C22:1 and longer) cis-monoenoic acids have relatively high melting points, whereas the more common C18 monoenoines are usually liquid at room temperature (1, 2).

In addition to the normal ethylenic double bonds, some fatty acids possess acetylenic bonds. They have been found in rare seed oils and some moss species.

![Figure 6](image_url)
Table 4

<table>
<thead>
<tr>
<th>Number of C atoms</th>
<th>Systematic name</th>
<th>Common name</th>
<th>Melting point (°C)</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>cis-9-octadecenoic</td>
<td>Oleic</td>
<td>16</td>
<td>Most abundant in plants and animals in both structural and storage lipids. A iso found in microorganisms.</td>
</tr>
<tr>
<td>18</td>
<td>cis-9- and 12-octadecenoic</td>
<td>Elaidic</td>
<td>44</td>
<td>In rumen fats as a result of hydrogenation of polynsaturated fatty acids.</td>
</tr>
<tr>
<td>18</td>
<td>cis-11-octadecenoic</td>
<td>Vaccenic</td>
<td>15</td>
<td>Common in bacteria, minor acid in most animal and plant lipids.</td>
</tr>
<tr>
<td>20</td>
<td>cis-11-eicosanoic</td>
<td>Gondoic</td>
<td>24</td>
<td>In fish oils and rape seed oil.</td>
</tr>
<tr>
<td>22</td>
<td>cis-13-docosanoic</td>
<td>Erucic</td>
<td>34</td>
<td>Major component of rapeseed oil and oils of Crucifer family, small amounts in animal tissues and fish oil.</td>
</tr>
</tbody>
</table>

Methylene-interrupted polynsaturated acids

Significant amounts of polynsaturated fatty acids (PUFA) with methylene-interrupted (two or more double bonds of the cis-configuration separated by a single methylene group) are found in all higher organisms. In higher plants, the number of double bonds are usually two or three, but in algae and other lower plants as well as in animals there can be up to six (Table 4).

The n-6 and n-3 families are two principal families of polynsaturated fatty acids occurring in nature and derived biosynthetically from linoleic (9-cis, 12-cis-octadecadienoic or C18:2n-6) and α-linolenic (9-cis, 12-cis, 15-cis-octadecatrienoic or C18:3n-3) acids, respectively (Fig. 6). Both fatty acids are synthesized in plants that can insert double bonds at the ∆5, ∆12, and ∆15 positions in a C18 chain but not in animals (they can insert double bonds at the ∆9, but not at ∆12 and ∆15). Therefore, these two acids are essential dietary components. In animals, additional double bonds are inserted between the carboxyl group and the ∆9 position by ∆9 desaturase enzymes, and the chain can also be extended in two carbon units at the carboxyl end of the molecules by elongase enzymes.

In animal biology, long chain metabolites of linoleate and linolenate may be even more important. The functions of arachidonic (C20:4n-6), eicosapentaenoic (C20:5n-3), and docosahexaenoic (C22:6n-3) acids make them essential conditionally (12). They are signaling molecules and are involved in the regulation of gene expression. These C20 fatty acids are also precursors of eicosanoids, which include prostaglandins (e.g., PGE\textsubscript{2} and PGE\textsubscript{3} series), thromboxanes, leukotrienes, and lipoxins, which have a variety of important biologic properties. Polynsaturated fatty acids are found in most lipid classes, but they are especially important as acyl chains of the phos- pholipids where they contribute to the particular physical and biologic functional properties of the membranes (1, 2).

Bis- and polymethylene-interrupted acids

Fatty acids with bis- or polymethylene-interrupted double bonds, or a mixture of methylene- and polymethylene-separated unsaturation, are found in some plants and marine organisms. From sponges and other marine invertebrates, a wide range of such fatty acids, the demospongic acids, have been isolated. They have bis-methylene-interrupted cis-double bonds and chain lengths (both odd and even) that range from C16 to C34. Double bonds in positions 5 and 9 are usual, although additional bonds may be present in position 7 and/or 9 together with methyl branching.

Bis-methylene-interrupted acids with a ∆5 cis-double bond only are common in gymnosperms (conifers) with a typical example being pinolenic acid (5-cis, 9-cis, 12-cis-octade- catrienonic acid). This fatty acid has been found in several pine and larch species at the levels of 25–30% of total fatty acids. In angiosperms, 5-cis, 13-cis-docosadieioidenic acid (Fig. 6) was reported in the seed oil of meadowfoam (Limnanthes alba) (10%) of the total fatty acids). An analog of pinolenic acid with a trans double bond in position 5 (5t,9c,12c-18:3) was found to be the main fatty acid of the seed oil of Aquilegia vulgaris (columbine) (1, 2).
Lipids, Chemical Diversity of

Table 4 Some naturally occurring polyunsaturated fatty acids

<table>
<thead>
<tr>
<th>Number of C atoms</th>
<th>Systematic name</th>
<th>Trivial name</th>
<th>Melting point (°C)</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dienoic acids</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>cis,cis-6,9-octadecadienic</td>
<td>Petroselinic</td>
<td>−11</td>
<td>Minor acid in animal tissues.</td>
</tr>
<tr>
<td>18</td>
<td>cis,cis-9,12-octadecadienic</td>
<td>Linoleic</td>
<td>−5</td>
<td>Major acid in algae and plant tissues. Essential component for mammals.</td>
</tr>
<tr>
<td><strong>Trinonic acids (methylene interrupted)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>All-cis-7,10,13-hexadecatrienic</td>
<td>γ-Linolenic</td>
<td></td>
<td>A algae and higher plants.</td>
</tr>
<tr>
<td>18</td>
<td>All-cis-6,9,12-octadecatrienic</td>
<td>α-Linolenic</td>
<td>−11</td>
<td>Minor component of animal tissues, found in a few seed oils (e.g., evening primrose oil).</td>
</tr>
<tr>
<td>18</td>
<td>All-cis-9,12,15-octadecatrienic</td>
<td></td>
<td></td>
<td>Higher plants and algae, component of galatosyldiacylglycerols.</td>
</tr>
<tr>
<td><strong>Trinonic acids (conjugated)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>cis-9, trans-11, trans-13-octadecatrienic</td>
<td>Eeleostearic</td>
<td>49</td>
<td>Occasional in some seed oils, especially tung oil.</td>
</tr>
<tr>
<td><strong>Tetraenoic acids</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>All-cis-4,7,10,13-hexadecatetraenic</td>
<td>Arachidonic</td>
<td>−49.5</td>
<td>Most abundant PUFA of animal phospholipids. Common is some algae and bryophytes.</td>
</tr>
<tr>
<td>20</td>
<td>All-cis-5,8,11,14-eicosatetraenic</td>
<td></td>
<td></td>
<td>In some algae, e.g., Chlorella spp.</td>
</tr>
<tr>
<td><strong>Pentaenoic acids</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>All-cis-5,8,11,14,17-eicosapentaenic</td>
<td>Timnodonic</td>
<td></td>
<td>In marine algae and fish oils as a major component, ubiquitous component of animal phospholipids.</td>
</tr>
<tr>
<td>22</td>
<td>All-cis-7,10,12,16,19-docosapentaenic</td>
<td></td>
<td></td>
<td>In marine organisms.</td>
</tr>
<tr>
<td><strong>Hexaenoic acids</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>All-cis-4,7,10,13,16,19-docosahexaenic</td>
<td>Clupanodonic</td>
<td></td>
<td>In animals as phospholipid component and in fish.</td>
</tr>
</tbody>
</table>

Conjugated fatty acids

Fatty acids with two or more conjugated double bonds are found in some plants and animals. In tissues of ruminant animals (and, hence, in meat and dairy products), fatty acids with conjugated diene system were detected as intermediates or by-products in the biodegradation of linoleic acid by microorganisms in the rumen. The main isomer, 9-cis, 11-trans-octadecadienic acid, may account for up to 1% of the total fatty acids of milk fat. 9-cis, 11-trans-15-cis-octadecatrienic acid, derived from α-linolenic acid, is present in ruminant tissues only in trace levels. This fatty acid has been shown to have several medical properties, especially anti-cancer and anti-atherosclerosis effects.

The only conjugated dienoic fatty acid isolated from plant sources (the seed oil of Chilopsis linearis) is reported to be trans-10, trans-12-octadecadienic acid. In contrast, fatty acids with conjugated triene systems have been detected in various plant species. Tung oil is the commercial source of the most widespread fatty acid from this group, 9-cis, 11-trans, 13-trans-octadecatrienic (ω-eleostearic) acid (1, 2).

As an example of conjugated tetraenes, a naturally occurring cis- (or ω) parinaric acid (C18:4 9c,11t,13t,15c), should be mentioned. This acid was isolated from the seeds of the M akita tree (Parinarium laurinum), which is a tropical rainforest tree indigenous to Fiji. It is also isolated from Impatiens spp. cis-Parinaric acid has been used to measure phospholipase activity, lipase activity, and as an indicator of lipid peroxidation.

Cyclic fatty acids

Cyclic fatty acids with a carbon ring alone or at the end of the alkyl chain occur naturally in plants, especially in certain seed oils and in microorganisms. Cyclopropane fatty acids are reported occasionally from marine animals and may be synthesized by symbiotic bacteria. In addition, a variety of carbocyclic structures are formed from methylene-interrupted polyenes during food processing.

The cyclopropane fatty acid, cis-11,12-methylene-octadecanoic acid, was reported first for Lactobacillus arabinosus and was given the trivial name lactobacillic acid. By now, fatty acids with a mid-chain cyclopropane group have been found in many bacterial species, including aerobic, anaerobic, Gram-negative,
and Gram-positive species. They comprise up to 35% of the membrane lipids depending on the culture conditions.

This type of fatty acid is usually found at low levels (around 1%) in plant oils that contain cyclopropene acids. Cyclopropene acids are characteristic for oils of the Malvaceae, Sterculiaceae, Bombacaceae, Tiliaceae, and Sapiodicaceae families with sterulic acid (9,10-methylenesuccinic-9-enoic acid) and malvacic acid (8,9-methyleneoctadec-8-enoic acid) as two major compounds. Sterulic acid is more abundant (about 50% of the total fatty acids in Sterculia foetida oil). 2-Hydroxysterculic acid may also be present in this oil as a possible intermediate in the biosynthesis of malvacic acid (through α-oxidation of sterulic acid).

A fatty acid with a terminal ring structure, 11-cyclohexylundecenoic acid, was first isolated as a minor component of butter fat. It is produced by bacteria in the rumen and may be later digested and accumulated by the host animals. Together with homologous fatty acids, it has been also reported for many bacterial species, especially for those isolated from the extreme environments (e.g., hot springs).

For example, in strains of the acidophilic and thermophilic species in significant amounts. Other epoxy acids include yeasts, algae, marine bacteria, and plants (seeds, leaves, and fruits). A particular role of these acids as scavengers of hydroxy and hydroperoxyl radicals has been suggested (1, 2).

Fatty acids with oxygen-containing functional groups

Several fatty acids and their metabolites have oxygen-containing functional groups, which are most commonly a hydroxyl, epoxidic, or a furan ring.

Ricinoleic acid (R-12-hydroxy-9 cis-octadecenoic acid) (Fig. 6) accounts for 80–90% of fatty acids in castor oil from Ricinus communis. It is found in other plant species and in the sclerotic of the endospore (Claviceps purpurea). Lesquerolic acid (R-14-hydroxy-11-cis-eicosenoic acid), which is a C20 homolog of ricinoleic acid, occurs in Lesquerella species (up to 70% of total fatty acids). Isoricinoleic acid (R-9-hydroxy-12 cis-octadecenoic acid, or 9-OH 18:2 12c) is a major acid in the Wrightia species. In plants, several C16 and C18 mono-, di-, and trihydroxy fatty acids are structural components of cutin (a polyester constituent of plant cuticle).

2-Hydroxy or ω-hydroxy acids have been shown to occur in sphingolipids, skin lipids, wool wax, bacterial cell wall lipids, and in some seed oils. 3-Hydroxy or ω-hydroxy acids have been shown to occur in plant oils. 3-Hydroxy acids are present in bacterial lipids. Vernolic acid (or cis-12,13-epoxy-octadec-9-enoic acid) was the first naturally occurring epoxy fatty acid isolated from the seed oil of Vernonia anthelmintica. It is also found in several Compositae species. In significant amounts. Other epoxy acids include coronoric acid (9,10-epoxy-12 cis-octadecenoic acid), which is found in some Compositae species, sunflower oil, and other oils.

A C20 homolog of vernolic acid is reported for talks of 

References

27. Broad TE, Dawson RMC. Formation of ceramide phospho-
26. Chalfant CE, Spiegel S. Sphingosine-1-phosphate and ceramide-
25. Spiegel S, Milstien S. Sphingosine-1-phosphate: an enigmatic
24. Zeyda M, Stulnig TM. Lipid Rafts & Co.: an integrated model
22. Suzuki T, Hasegawa K. Diol lipids in the phospholipid frac-
21. Batrakov SG, Panosyan AG, Konova IV, Bergelson LD. Identi-
20. Ratledge C, Wilkinson SG. An overview of microbial lipids. In:
19. Eichenberger W, Gribi C. Lipids of
Lipids, Organization and Aggregation of

Ilpo Vattulainen, Institute of Physics, Tampere University of Technology, Finland; Laboratory of Physics and Nelimirki Institute of Physics, Helsinki University of Technology, Finland; MEMPHYS—Center for Biomembrane Physics, University of Southern Denmark, Odense, Denmark

doi: 10.1002/9780470048672.wecb664

Lipids constitute one of the main classes of molecules in biological systems. They are involved in numerous cellular functions either as individual molecules or as lipid aggregates with varying sizes and morphologies. For example, lipids are a crucial component of cellular membranes that surround and protect cells. Lipids comprise membrane domains that provide membrane proteins with a well-defined environment to carry out their functions. Lipids also interact specifically with some proteins, which render their functions possible. Lipids play the role of drugs and enzymes, and our skin and lung surfactant lining lung epithelial cells are composed largely of lipids. What is more, lipids are used in delivery vehicles to encompass drugs and other molecules. These functions and many others develop in part from lipids’ specific properties relevant on molecular scales and also from the assembly of lipids as fascinating structures observed over a multitude of scales beyond molecular size. Here, we discuss the functions of lipids and lipid structures together with their structural and dynamic properties, including examples and highlights of recent studies.

Lipids (see Fig. 1) constitute one of five classes of molecules that can be considered as crucial in biological systems. Together with proteins, nucleotides (DNA), carbohydrates, and water, lipids can be thought of as one of the basic building blocks of living systems.

For some reason, lipids cause a lot of emotions among the people. One often talks about “bad” and “good” fats, and in terms of health, this view is partly understandable. In a similar manner, it is common to talk about “bad” and “good” cholesterol; although in this case, most common people probably do not even know that only one type of cholesterol molecule exists, which is crucial for life. It is one of the most common and the most important lipids in eukaryotic cells. The “bad” and “good” cholesterols refer to the carrier particles (lipoproteins) that transport cholesterol molecules, and their effects on our health depend on the lipoprotein in question. Meanwhile, when a great deal of recent discussion has dealt with the genetic code and the proteomics of proteins, it seems that far less attention has been paid to the importance of lipids. In part, this attention is because no genes code lipids. Yet undoubtedly lipids are a crucial component of cells: They would not survive without lipids. It has been observed that polyunsaturated lipids that have several double bonds in their hydrocarbon chains are involved in the functioning of the eye, and DHA (perhaps the most important polyunsaturated fatty acid) is vital for normal brain development for infants and for the maintenance of normal brain function throughout life. Furthermore, we all appreciate the importance of lipids as a major source of energy, which may even be critical for survival under extreme conditions. In summary, lipids and fat are good. Those who love good food, such as sushi and fish in general, probably appreciate this view.

Appreciating the importance of lipids in a variety of biologically important cases, one is tempted to understand how they actually function and how that is affected by the structures of their assemblies. The answers to these questions are far from being resolved, but we do have some insight into the related issues. What is remarkable and deserves to be stressed here is that lipids are characterized by fascinating structures over a multitude of scales, which range from the size of individual molecules to the sizes of colloidal systems that comprise large amounts of lipid molecules. Lipids are the main component of lipoproteins that carry cholesterol, vesicles that transport molecules inside them, cell membranes that surround cells, lung surfactant that keeps us alive by allowing us to consume oxygen, and skin that protects us from the outside. Beautiful examples of the...
Lipids, Organization and Aggregation of
Figure 1 Examples of lipids found commonly and used in cells. (a) DPPC; (b) 1-palmitoyl-2-({1-pyrenedecanoyl}-sn-glycero-3-phosphocholine (PyrPC); (c) Palmitoyl-SM; (d) Cholesterol; (e) Dipalmitoylphosphatidylethanolamine; (f) 1,2-di-O-palmitoyl-3-β-D-galactosyl-sn-glycerol (DPGALA). Of these lipids, PyrPC is a pyrene-linked lipid probe and DPGALA is a glycolipid. The first four lipids are represented by a united-atom description, and the last two lipids are represented by a full-atom description.

Figure 1

Functions of lipids
No strict definition exists for the term "lipid" that is generally accepted. In a broad sense, lipids are compounds of low or intermediate molecular weight with a substantial proportion of hydrocarbons (1). Lipids have also been defined as "fatty acids and their derivatives, and substances related biosynthetically or functionally to these compounds" (4). Sometimes, it is also expected that molecules considered as lipids have some biological function. Many molecules based on fatty acids are lipids, but some vitamins and hormones are based on fatty acids as well. They serve many functions in living organisms, and the broad scope of these functions is truly fascinating.

The lipid membrane provides shelter for membrane proteins to do their functions. However, instead of working alone, membrane proteins such as ion channels work together with the membrane, such that the lipid composition around the protein actually affects the activation and the functioning of the protein. This idea is largely the essence of the lipid raft model (8), which highlights the importance of lipids in a variety of cellular functions. It has been observed, for example, that rhodopsin, which is the light sensitive membrane protein, favors interactions with polyunsaturated lipids (9).
Lipids are the main component in lipoproteins known as carriers of cholesterol (2). HDL (high-density lipoprotein, which is the "good" cholesterol) transports cholesterol and its esters from cells to liver for recirculation, whereas LDL (low-density lipoprotein, which is the "bad" cholesterol) carries cholesterol and cholesterol esters to the cells. The ratio of the two carrier particles partly determines one's risk for diseases such as atherosclerosis, although the understanding and the overall view of the related issues is still incomplete in many ways.

Some lipid species serve as second messengers that pass on signals and information in the cell, such as in programmed cell death in which a lipid known as ceramide has been proposed to be the messenger (10). The same lipid is often found in hair conditioners. Lipids also play the role of enzymes, receptors, and drugs, and our skin is largely composed of lipids. In a similar manner, a major amount of molecules that comprise the pulmonary surfactant, which is a thin liquid film that lines lung epithelial cells, are saturated lipids (11) characterized by a lack of double bonds in the chains. The lung surfactant stabilizes the alveoli during expiration, when lungs undergo compression, and reduces the re-expansion work during inhalation. Lipids are also widely used for storing energy in forms of triglycerides, and recently, disorder in the lipid spectrum of cells has been related to, for example, atherosclerosis and major psychiatric diseases. As for applications, lipids are used by nature as novel micro-encapsulation devices—which is an exciting application nowadays used for drug and gene delivery (12).

Scales of lipid systems

Temperature scales

The relevant temperature range for functions that deal with lipids is the physiologic temperature, which is about 330 K. The phase behavior of lipid systems should then be considered with respect to this value. Particular attention is usually paid to the main transition temperature \( T_{\text{m}} \) defined for a one-component lipid bilayer (see Fig 2). A bove \( T_{\text{m}} \), one finds a fluid (liquid-disordered) phase characterized by a lack of translational order in the bilayer plane and weak ordering of the lipid hydrocarbon chains. Below \( T_{\text{m}} \) in the gel phase, in turn, the lipids in the membrane plane position themselves to follow hexagonal packing, besides which the hydrocarbon chains are strongly ordered (all-trans) and tilted with respect to the membrane normal direction. These transitions in lipid conformational and translational order can be characterized readily by order parameters such as the second-order Legendre polynomial and the in-plane structure factor (13, 14). The second-order Legendre polynomial is used often because it can be exploited to provide insight into the orientation of lipid hydrocarbon chains, their tilt as well as the orientation of a lipid head group, and it can be accessed rather easily through simulations as well as experiments (nuclear magnetic resonance (NMR), see below). The main transition temperature varies from one lipid to another and is largest for lipids with saturated and long chains. For example, for a dipalmitoylphosphatidylcholine (DPPC) bilayer, it is 314.5 K. For comparison, for polyunsaturated lipids typical main transition temperatures are below 273 K. In a similar manner, other lipid systems such as droplets of triglycerides have distinctly different phases, usually crystalline at low temperatures and fluid above the transition point. For most of the physiologically important lipids \( T_{\text{m}} \) is below 310 K, which implies that the physiologically more important phase is the fluid phase.

However, natural membranes are not one-component bilayers but rather are comprised of a variety of different lipid components. This matter is discussed below in more detail. Consequently, the presence of many lipid types develops a formation of membrane domains with varying composition, and hence the main phase transition temperature is not well defined for a many-component membrane. Of particular interest are membranes with a large concentration of cholesterol, because cholesterol drives the formation of a new phase. The liquid-ordered phase (3), as it is commonly called, is characterized by a lack of translational order in the membrane plane, like in a fluid phase, and by a strong conformational order among the lipid hydrocarbon chains, which in turn is reminiscent to the gel phase. Currently, it is thought that the liquid-ordered phase and cholesterol in particular play a prominent role in the functions of lipid rafts, which are strongly ordered membrane domains involved in a variety of cellular functions such as signal transduction and protein sorting; see below for additional discussion. However, despite their significant ordering, rafts are essentially fluid-like membrane domains, which highlight the importance of fluidity for membrane functions under physiologic conditions.

Length scales

To understand cellular functions that deal with single-molecule properties, we must achieve deeper knowledge about the structure and the molecular organization of lipid membranes on the nanometer scale. This knowledge is particularly important in the context of membrane proteins, because studies indicate that some membrane proteins favor certain specific lipids in their vicinity (9). Recent crystallographic studies of membrane protein-tein structures have also revealed how lipids may be an integral part of the protein structure (15). However, as biological functions take place over a multitude of scales, one must also understand how structures and functions at larger scales emerge from corresponding ones at smaller scales. For example, membrane protein functions can be regulated by the (large-scale) elastic properties of membranes (16). Also, lipid aggregates such as micelles and vesicles are involved in intracellular transport and range in size from a few nanometers to micrometers; the size of a cell is typically about 30 micrometers. Therefore, no length scale would be specific to lipid structures.

Time scales

Time scales of dynamic processes in lipid systems are wide and range from picoseconds to hours or even months. Examples are discussed below. Here, we just refer to rotational motions as the fastest dynamical events and cell death as the slowest event. What is relevant to stress here is the lack of any specific time scale.

In summary, lipids play a prominent role in numerous cellular functions. Yet lipids and the structures composed of lipids are not characterized by any specific length or time scale. However, just like soft matter in general, there is one specific scale in
Lipids, Organization and Aggregation of

![Image of Lipid Bilayer](image)

**Figure 2** Lipid bilayer composed of DPPC lipids in two different phases. (a) Gel phase below the main transition temperature; (b) Fluid phase above the transition temperature. Water is not shown for clarity.

common: the well defined energy scale given by the thermal energy $kT$.

**Lipids in Various Forms**

The classification of lipids is largely arbitrary. It can be based on water solubility (hydration) or swelling of a lipid system at the air-water interface, for example. Here, we approach lipids in terms of increasing complexity and focus on those lipids that are found mostly in cells, see Fig. 1. A more thorough discussion of the topic is given by Hauser and Poupart (1) and Larson (4).

**From fats to fatty acids and lipids**

Fats are easy to recognize on the basis of our every-day experience. Usually, they are considered as frozen oils and used in cooking (17). The main component of a fat is a hydrocarbon moiety that is typically a long hydrocarbon chain, with a varying number of carbon atoms attached to each other through single (saturated) or double (unsaturated) bonds. Depending on the number of double bonds, the chains are called monounsaturated (one double bond), diunsaturated (two), or polyunsaturated (more than two). Fatty acids are obtained by attaching a carboxyl group at the end of a chain. Although hydrocarbons do not dissolve readily in water, fatty acids dissolve more easily because of the hydrophilic –COOH group. Yet fatty acids are not found free in the cell often. The rather few exceptions include the intercellular transport of fatty acids inside lipoproteins and the chemical reactions such as hydrolysis because of enzymes acting on lipids, in which a lipid is broken down into smaller pieces (one of them is a hydrocarbon chain). Fatty acids are linked more commonly to a chemical group that acts as the backbone of complex lipids. One of the most general groups is glycerol, which can form ester bonds in up to three positions. By doing so, one can form nonpolar lipids such as tri-acylglycerols (frequently also called triglycerides)
with three ester bonds and di-acyl glycerols with two ester linkages. Triglycerides are the major components of dietary fats and a common hydrophobic storage means of fat. They are also transported between cells via lipoproteins. Di-acyl glycerols function as second messengers in signal transduction.

Another common backbone for fatty acids is sphingosine, which is a long-chain amine. Through an amide linkage, it is bound to a fatty acid, which forms the ceramide found often in skin and hair conditioners. As mentioned above, ceramide is also one of the signaling molecules involved in processes such as apoptosis, which is the programmed cell death.

**Polar lipids**

The above molecules are nonpolar and therefore hydrophobic. The polarity of lipids can be promoted by binding polar groups to glycerol- or sphingosine-based molecules. For example, the vacant –OH group in the glycerol moiety of di-acyl glycerols can be linked to a polar group such as phosphatidylcholine (PC), phosphatidylethanolamine (PE), phosphatidylserine (PS), phosphatidylglycerol (PG), or phosphatidylinositol (PI). PC and PE groups are neutral and zwitterionic, whereas PS, PG, and PI are anionic. Using glycerol as a backbone, one obtains glycerophospholipids, in which one of the above polar moieties usually is acting as the head group. The best known exception is cardiolipin, in which two glycerophospholipids, which start from the phosphate group, are linked to one another through a glycerol group. Such a dimeric anionic lipid has usually four hydrocarbon chains. In a manner similar to glycerophospholipids, the free –OH group in ceramide (as the end group of sphingosine) can be linked to one of these head groups and one obtains sphingophospholipids. A common lipid of this group is sphingomyelin with a PC moiety as the head group.

One of the most complex classes of lipids in terms of their head group is glycolipids. Glycosylglycerolipids are di-acylglycerols linked to a monosaccharide or a disaccharide such as galactose or lactose, respectively. In turn, glycosphingolipids are sphingolipids with a carbohydrate head group; the most common types are cerebrosides and gangliosides based on ceramide as the fundamental structure. Another subclass complementing the above ones is lipopolysaccharides. The complexity of glycolipids is caused not only by their structural diversity, but also by the stereo-chemistry that plays a major role in their properties. Carbohydrates in general are characterized by the lack of a well-defined structure-function relationship, which means that the details do matter and often turn out to be very important. For example, recent studies have shown that glycolipid bilayers with either a galactose or glucose head group yield substantially different membrane properties. Yet galactose and glucose differ only in chirality. Nonetheless, the most fascinating and complex lipid type is sterols. They have a rigid steroid ring structure, which is a simple hydroxyl group as their polar head and a short flexible tail at the other end of the molecule. Of the many sterols found, two sterols exist whose significance in cells is above the others. Cholesterol is abundant in eukaryotic cells, and ergosterol is abundant in fungi. The reasons why nature has chosen these two specific sterols are not well understood. One is tempted to think that the properties of all sterols should be similar, because the structural differences between them are seemingly negligible. However, it has turned out that the details do matter. For example, adding just one double bond to the structure of cholesterol changes its membrane properties substantially (19). Consequently, cholesterol is the sterol in eukaryotic cells, and cholesterol esters are the form of cholesterol by which it is transported in lipoproteins.

**Lipid composition of biological membranes and lipoproteins**

The composition of lipids varies markedly in a cell, and biomembranes and lipoproteins provide crisp examples about the diversity of different lipids. In the plasma membrane of animals (1), the amount of cholesterol is usually around 20-30 mol%. The rest of the lipids are mainly PC, PE, and sphingomyelin (SM) lipids, with smaller amounts of PS, PI, and glycolipids. These lipids are distributed asymmetrically across the membrane, because most cholesterol, PC, and glycolipids are located in the extracellular (outer) leaflet, whereas PS and PE lipids are located mainly in the intracellular (inner) monolayer. The lipid composition can be highly different in other organelles, however, as is the case in mitochondria (1), in which the mitochondrial membrane is composed of two (inner and outer) membranes. There, the amounts of cholesterol, SM, and PS are negligible; most lipids are PC and PE. The major difference compared with plasma membrane is the concentration of cardiolipins. They are actually found only in bacterial and in mitochondrial membranes, where their numbers are significant; even in mitochondria they are located mainly on the inner membrane.

In lipoproteins, one finds essentially the same types of lipids but with somewhat different concentrations. In order of decreasing numbers, low density lipoproteins contain (2) mainly cholesterol esters, (unesterified) cholesterol, phosphatidylcho- lines, SM, triglycerides, and lyso-PC, and smaller numbers of PE, PI, and ceramide. In HDLs, one finds the same lipid types but in a different order, as HDLs are abundant in phospholipids and are complemented by cholesterol esters, cholesterol, and triglycerides. Besides their head group, also other features differentiate lipids from one another. Among phospholipids, which is the main group of lipids in cells, about 10% are charged. Except for sphingomyelin, which is cationic, all other charged lipids are anionic. What is even more striking is the diversity of unsaturation and chain length. The number of double bonds per hydrocarbon chain ranges from zero to six, and the chain length ranges typically from 14 to 22 carbons per chain. Usually the sn-2 chain in a glycerophospholipid is unsaturated, whereas the sn-1 chain is saturated, and these two chains may be asymmetric in terms of chain length. Furthermore, in addition to ester lipids, ether lipids such as plasmalogens exist, in which the hydrocarbon chain is linked to glycerol by an ether bond instead of an ester bond. For example, ether lipids act in cell signaling.

The above examples highlight the extraordinary diversity of lipids. On a cellular level, any single membrane may contain more than 100 different lipid species, each assumed to have some particular function. Because nature always has a reason,
Methods Commonly Applied to Lipid Systems

An extensive toolbox is often applied to lipid systems. Here, let us discuss some of the most important techniques, considering both experimental as well as computational ones. Although the description here is inevitably brief, it hopefully provides some flavor to the techniques in question.

Experimental techniques

One of the most versatile methods for studies of lipid systems is NMR (13, 20). In NMR, an applied magnetic field is used together with an alternating electromagnetic field to probe changes in molecular alignment. NMR can provide detailed information on the topology, the dynamics, and the structure of molecules. In the context of lipids, it allows determination of the orientation of a lipid or a group of atoms in a lipid, such as the head group region or a single C-H bond in a hydrocarbon chain. The latter in particular is often used to determine the ordering of lipid hydrocarbon chains by selective deuteration of acyl chains, as it provides one with a solid understanding of fluidity inside a membrane. As for dynamics, NMR is commonly employed to explore the rotational motion of atom groups in lipids and the diffusion of lipids in the plane of a membrane.

Small angle X-ray scattering (SAXS) (13, 21) is used increasingly in biological sciences to determine dynamic structures of various molecular systems. The X-ray sources with high intensity allow the observation of weak scattering features that are associated with the internal structures of molecules studied. In the context of lipids, SAXS is often used to measure lipid system structures and their phase behavior.

Differential scanning calorimetry (DSC) (22) is based on measuring changes in heat capacity because the temperature of a system is varied monotonically. As heat capacity is a thermodynamic response function, it is expected to exhibit critical behavior close to a phase transition boundary, which in turn can be detected by DSC. Consequently, DSC is a common way to detect phase behavior.

Considering techniques that allow the imaging of lipid surfaces, scanning probe microscopes such as the atomic force microscope (AFM) (13, 23) have become very appealing. The AFM allows measurements of native lipid samples under physiologic-like conditions and while biological processes are underway. It is hence often used to determine lipid membrane structures, structural defects in membranes, domain formation, and even the behavior of lipid rafts with high nanometer-scale lateral resolution.

Another appealing technique used often to image lipid structures is fluorescence microscopy and imaging (24). By using fluorescent probes attached covalently to lipid molecules, the fluorescence microscopy provides a wealth of spatially resolved information of individual lipid aggregates instead of averaging over a large number of them, see Fig. 3. Hence, fluorescent probes provide detailed information of structure as well as dynamics, and this information is often exploited to investigate phenomena such as trafficking and membrane dynamics in the spirit of single-particle tracking and single-molecule detection (14, 25). The downside is that probes inevitably perturb the system (26, 27), thus their use and the interpretation of the results warrant particular care.

Computational modeling

Unlike simulations of proteins and protein complexes, modeling and simulations of lipid systems is relatively “easy” in the sense that lipid molecules and (smallest) lipid aggregates are reasonably small, and the time scales related to many processes that take place in lipid systems are of the order of nanoseconds. Consequently, even atomistic modeling of lipid aggregates is feasible for reasonably complex systems. Here, we discuss briefly the three main levels of modeling associated with lipids.

Quantum-mechanical (QM) modeling

QM methods (28) are crucial in studies of processes in which one must account for electronic degrees of freedom, such as the action of sphingomyelinase acting on sphingomyelin. Such enzymes hydrolyze lipids, which cuts them into pieces. To describe the hydrolysis process fully, one should treat the action center in a QM manner. The problem is that QM techniques are feasible only to small scales, typical system sizes are a few hundred atoms and time scales range up to tens of picoseconds. Obviously, these scales are short compared with most biologically relevant processes. Consequently, QM techniques are often bridged to classical simulations in a manner in which the region of interest (such as the action center) is described in a QM fashion, whereas the rest of the system follows classical equations of motion using molecular mechanics (MM) force fields. Such QM/MM methods have gained increasing popularity in the field, but their wider use is limited to difficulties of treating the interface accurately between QM and classical regions. However, recent studies have shown that the QM/MM techniques work well, and the prospects for broader applications are promising (29).

Classical atomistic simulations

The most popular technique to deal with lipid systems has by far been classical molecular dynamics (MD) (28, 30). In MD, all interactions are classic, and the time evolution of the system is described by integrating Newton’s equations of motion. The particles can represent atoms or clusters of atoms; the most typical choice is the full-atom description in which all atoms including hydrogens are described explicitly, and the united-atom description in which each methyl and methylene group is described by a single particle. The particle-particle interactions are usually determined from QM calculations and tuned even more in an iterative manner by fitting system properties to experiments until simulation results and experimental data match sufficiently well. Usually, the largest system sizes are hundreds...
of thousands of atoms over hundreds of nanoseconds, which renders possible studies of various relevant processes. Examples are presented below. However, considering a practical example for a lipid membrane, the largest systems considered to date have been about $20 \text{ nm} \times 20 \text{ nm}$ in size in the membrane plane, which is rather small compared with typical membrane domain sizes. This finding implies that atomistic MD simulations currently are not the method of choice for many-component membrane systems, in which time scales of mixing and domain formation are far beyond the limits of atomistic simulations. To reach such scales, efficient ways are needed to treat the dynamics or the models that are simpler than the atomistic ones. The first idea can be conducted by, for example, Monte Carlo simulations (28, 30, 31). Instead of integrating the equations of motion in a deterministic manner, one employs random noise to evolve the system from one configuration to another in a random fashion. Despite the fact that Monte Carlo simulations (usually conducted with the Metropolis scheme) do not yield natural dynamics, they gauge equilibrium properties that are not functions of time. What is more, the Monte Carlo approach allows the use of nonphysical moves, which usually provides a major speed-up, hence facilitating studies of very large systems. The second idea to deal with simplified (coarse grained) models is more common, however, because it allows one to consider dynamic quantities and nonequilibrium properties.

Coarse-grained models

The key idea of coarse-grained (CG) models (32, 33) is to get rid of all details that are not relevant for the properties one is interested in. Replacing CH$_2$ groups by a united-atom particle can be considered coarse graining. In a similar fashion, one can replace several (say, four) methyl groups in a lipid hydrocarbon chain by a coarse-grained particle. Because the number of interacting beads is reduced, the computational burden decreases as well, which allows one to consider larger systems and time scales. This strategy is used more and more often in simulations of lipid systems, and the results have been very encouraging. The main challenges associated with CG models are the choice of the coarse-grained molecular description, for which one can employ systematic techniques such as self-organizing maps (34) together with plain intuition, and the choice of the interactions used in the CG model. For the latter issue, several approaches have been proposed. Perhaps the most promising approach is presented by Marrink et al. (35, 36), who used thermodynamic quantities such as solvation free energies to determine interaction strengths for different molecular groups. Another means to improve coarse graining is to get rid of the solvent (37, 38). This solvent-free approach can provide a major speed-up for dynamics, although it also has obvious limitations because of the absence of full hydrodynamics.
Structure of Lipid Membranes and Importance of Cholesterol

One of the main functions of lipids is to serve as the main structural components of cell membranes (7). Membranes resemble thin elastic sheets with a total thickness of about 5 nm. The membrane is composed of two lipid monolayers. The lipids in a membrane typically include two nonpolar and hydrophobic (water hating) acyl chains and one polar head group next to the head group, which in turn is usually polar and hydrophilic (water loving) and therefore can form hydrogen bonds with neighboring water molecules. This "schizophrenic" nature of lipid molecules causes them to self-assemble as closed objects such as liposomes, such that the head groups face water molecules while the hydrophobic hydrocarbon chains are protected from the water phase.

The essential structure of membranes is captured by the single-component lipid bilayer shown in Fig. 2. Although this view is a highly simplified description of an actual biological membrane, it readily demonstrates the importance of lipids on membrane structure and dynamics. Native biological membranes found in living systems are composed of lipids in terms of a lipid bilayer; although in those cases, it is not a single-component but rather a many-component membrane that consists of hundreds of different types of lipids that differ from one another in several ways such as size, unsaturation level, chemical composition in the polar head group region, number of hydrocarbon chains, and charge. The bilayer in native membranes also acts as a soft fluid-like environment for integral and peripheral proteins embedded in or attached peripherally to the membrane. In addition, the membrane proteins are involved in a dynamic rubber-like network known as the cytoskeleton attached to the inner surface of the membrane. The cytoskeleton moves the cell, gives even more rigidity to the membrane, and also allows the membrane to adjust its shape to varying nonspherical shapes. Moving on, also the outer leaflet of the membrane is covered by a network, which in this case is made of carbohydrates. The glycocalyx network, as it is called, is involved in cell–cell recognition and adhesion to other cells, among other functions.

In eukaryotic cells, one of the most important, or even the most important, lipid is cholesterol. Cholesterol constitutes about 30–40 mol% of the plasma membrane, and in the ocular lens membrane, its amount can be as large as about 80 mol%. Therefore, the role of cholesterol deserves the particular attention discussed below.

Influence of cholesterol

Cholesterol affects a large variety of membrane properties in animal cells (39). It is involved in modifying dynamical membrane properties by reducing passive permeability, slowing down the lateral diffusion of molecules in fluid-like membranes, and speeding up diffusion in gel-like membranes. It also affects bilayer properties by condensing the bilayer, which changes its elastic properties and promotes the order of phospholipid acyl chains in the hydrophobic membrane core. In this manner, cholesterol develops the formation of the liquid-ordered phase (3), which is characterized by significant conformational order in the lipid hydrocarbon chain region and the absence of translational long-range order in the membrane plane. Through the formation of the liquid-ordered phase, cholesterol governs membrane fluidity and is associated with membrane domain formation, in particular the formation of lipid rafts (8). The role of cholesterol in rafts perhaps best underlines the biological importance of cholesterol, see below for discussion. What is more, cholesterol seems to have a unique structure-function relationship because many cells do not do well without cholesterol. For example, desmosterol, which differs from cholesterol only by one double bond in the short hydrocarbon tail, cannot substitute for cholesterol (19).

The ordering capability of cholesterol is illustrated by Fig. 4, which shows the effect of increasing cholesterol concentration on the ordering of lipid hydrocarbon chains in a fluid DPPC bilayer (39). In the absence of cholesterol, the NMR order parameter, $S_{zz}$, indicates reasonable conformational order close to the membrane-water interface, and monotonically decreasing order toward membrane center. For increasing cholesterol concentration, the conformational order increases significantly. For the largest cholesterol concentration of 50 mol%, the order parameter is close to its maximal value of 0.5, in which case the chain would be in a full-trans conformation standing along the membrane normal direction. The enhanced ordering of hydrocarbon chains because of increasing cholesterol concentration is coupled to stronger packing, which in turn reduces the amount of free volume and changes the shape and size distributions.
of free volume pockets inside the membrane (40). It is readily clear why cholesterol plays such a strong role in many dynamic membrane processes. Yet the biological relevance of cholesterol in cells is related largely to lipid rafts.

Lipid rafts

In 1997, Simons and Ikonen (8) proposed that strongly ordered membrane domains rich in cholesterol and sphingolipids would be involved in a variety of cellular processes such as signal transduction, protein sorting, and programmed cell death. Ever since, the research that focuses on lipid rafts (see Fig. 5, as they are commonly called, has been very intense (41). Although the studies have demonstrated the role of rafts in numerous processes, the definition and the structure of rafts are still under debate. The uncertainties regarding the structure of rafts is largely caused by the small length and time scales, because the sizes of rafts seem to range from a few nanometers to hundreds of nanometers, and the time scales are short because of...
the transient nature of rafts. These conditions pose a challenge to gauge raft systems through experiments, thus the precise understanding of the structure within rafts has remained limited. Recent atom-scale simulations have complemented experiments and have shed some light on the matter (42). As Fig. 6 shows, cholesterol strongly orders the lipids in its vicinity, which condenses the membrane. The interplay of cholesterol with sphingolipids develops domains that are distinctly highly packed, which in turn slows down the lateral diffusion rate substantially compared with other membranes with large amounts of cholesterol. Cholesterol plays a role in the lateral pressure profile that acts on proteins embedded in a membrane. Recent simulations have shown that lipid rafts have distinctly different pressure profiles compared with other membrane systems, and that the contribution of the pressure profile for the free energy barrier for membrane protein activation (using a model of MscL, as an example) can be considerable compared with the total free energy barrier (42). All together, the results highlight the distinct nature of rafts and the importance of cholesterol in membranes overall. However, whereas the understanding of rafts and their role in cellular functions has made considerable progress during the last decade, major gaps remain to be covered by a combination of novel experimental and theoretical efforts.

Dynamics of Lipid Systems

The dynamics of lipids and lipid aggregates is driven mainly by thermal fluctuations through $k_BT$. This finding is truly fascinating because it implies that nature uses random walks in essentially all dynamic processes: The diffusion of lipids in membranes is a random walk, the growth of microtubulin follows a random-walk line pattern, and the (passive) diffusion inside a cell overall is caused by the thermal forces that act on lipid systems (43). To complement thermal diffusion, active dynamic processes occur in which the chemical energy contained in ATP is converted into mechanical work done by motor proteins. These active processes are also involved partly in the dynamics of lipids, such as in the one-dimensional diffusion of motor proteins that drag cellular cargo inside a lipid droplet along a tubulin. However, we focus here on the passive dynamic processes driven by thermal energy.

Considering dynamic processes in fluid membranes as an example and starting from the fastest dynamical processes, one may first consider rotational diffusion of individual carbon-hydrogen bonds in CH$_2$ groups in lipid hydrocarbon chains. The time scale of these rotational motions is on the order of picoseconds. The rotational motion of whole lipids around their principal axes of rotation is a slower process and usually takes place over a scale of nanoseconds. Lateral diffusion, in turn, involves diffusion of matter and hence longer time scales. This time scale is characterized by the diffusion length $L = (2dDt)^{1/2}$, where $d$ is the dimensionality, $D$ is the diffusion coefficient (≈10$^{-10}$ cm$^2$/s in fluid membranes), and $t$ is the time scale considered for diffusion. On average, a lipid in a fluid membrane diffuses over a distance of its own size (≈0.8 nm) in about 15-20 ms. The mixing of different lipid components and domain formation requires longer times and depends on the length scale. A smearing a membrane domain with a radius of 100 nm, an individual lipid would cross this length on average in about 250 microseconds. In a similar manner, assuming a roughly constant diffusion coefficient in all regions in a membrane, the average time scale for diffusion from one point in a plasma membrane (surrounding the cell) to the other side of the cell would be about a minute. One minute is this a lot? No. It is vanishingly small compared with the time scales of many cellular functions or, say, protein folding. Diffusion over cellular scales along the membrane plane is an efficient means to transport molecules, and for a cell this comes for free. No ATP is wasted. Instead, thermal fluctuations drive the motion.

All dynamic events are not so rapid, however. Lipid flip-flop, in which a lipid translocates from one leaflet to another in a membrane, is usually a profoundly slow process and occurs in a time scale of minutes or even hours. As for the longest time scale associated with biological lipid systems, the typical life time of a cell ranges from a few days to several months, which depends on the tissue type.

The above estimates are for fluid-like systems. In gel-like systems with features of frozen order, the time scales are much longer. For example, the lateral diffusion coefficient in a gel-like one-component membrane is about 10$^{-12}$-10$^{-13}$ cm$^2$/s (43), whereas in a fluid membrane it is usually $\approx$10$^{-7}$ cm$^2$/s. In a similar manner, the diffusion of matter inside lipid droplets is a much slower process compared with lipid interfaces caused by entanglement effects, as the situation is largely similar to a polymer melt. This effect is the case inside LDL. It has been estimated that the diffusion coefficient for cholesterol esters inside LDL particles is roughly 10$^{-8}$ cm$^2$/s (44) and is intermediate to diffusion in fluid- and gel-like membranes. Overall, the mechanisms associated with dynamic processes in lipid systems are complex and are understood rather poorly, although the combination of experiments and computer simulations has improved the situation recently. As an example, let us consider a more concrete situation, the formation of pores in a cell membrane and its significance for cellular functions.

Pore formation in lipid membranes

Transient water pores in cellular membranes are involved in several relevant processes, such as maintenance of osmotic balance, drug and antibody delivery into cells, and ion transport across the membrane. Understanding ion transport across membranes is especially important, because membranes strive to maintain a cationic electrochemical gradient used for ATP synthesis. Yet, ions leak through lipid membranes, and understanding the mechanisms associated with ion leakage would allow one to control membrane properties better in related applications.

Figure 7 illustrates the complexity of transient pore formation (45) under conditions that closely resemble physiologic conditions. The initial ion concentration imbalance across the membrane develops a strong local electric field that induces the formation of a pore. The pore formation starts with the creation of a single water defect in terms of a chain of water molecules, which spans the entire membrane. The defect then expands within less than 1 ms through redistribution of lipid head groups.
Influence of Probes

Often, experimental studies of lipid systems are based on spectroscopic approaches, which in turn frequently employ probes for enhancement of sensitivity and resolution. For example, in NMR, hydrogen atoms of lipids are replaced with deuterium, and in fluorescence spectroscopy and imaging, native lipid molecules are replaced with lipids in which one or more hydrocarbon chains is linked covalently to a fluorescent marker such as pyrene or diphenylhexatriene. Fluorescent markers allow one to follow numerous cellular processes in real time, such as intracellular trafficking of molecules and formation of domains within a biomembrane, see Fig. 3. The downside is that the probes tend to perturb their environment and affect the thermodynamic state of the system. Experiments have shown, for example, that probes may change the main transition temperature of a lipid membrane, and that the dynamics of probes may deviate considerably from the dynamics of corresponding native molecules (see discussion in Reference 27). Therefore, we wish to pose several questions. What is the range of perturbations induced by the probe? How significant are these perturbations actually?

Figure B shows a plot of a fluid-like DPPC bilayer, in which a small fraction of the lipids are replaced by a pyrene-containing PyPC probe, see Fig. 1 (26). The study has demonstrated that the perturbations in the vicinity of the probe are substantial, as the conformational order parameter \( S_{CD} \) of lipid hydrocarbon chains close to the probe may change as much as about 100%. However, what is also found is the short range of perturbations, because the perturbations are negligible beyond a distance of about 1.5 nm. In practice, this finding implies that about 20–30 lipid molecules are affected by the marker, but the global properties averaged across the membrane are affected only little.

This brief example highlights the possible problems associated with using fluorescent probes and antibodies and with interpreting the results obtained through probes. However, fluorescent labels are one of the most appealing means to follow a variety of intriguing dynamic processes in biological matter. Although there is a reason to be cautious, there is even more reason to develop better probes that mimic the properties of native molecules as closely as possible.

Acknowledgments

We wish to thank our collaborators whose contribution for the results presented here has been crucial. Additionally, Jarmila Repakova, Empu Salonen, Tomasz Rog, Ole Mouritsen, and Luis Bagatolli are thanked for help with illustrations.

References


Figure 7 Snapshots of pore formation and the resulting ion leakage across a lipid membrane (45). The plots (a)–(f) demonstrate the process at different times after the initial electric field has been established: (a) 20 ps; (b) 450 ps; (c) 1000 ps; (d) 1070 ps; (e) 9180 ps; (f) 60 ns. The membrane is not shown for clarity, whereas water is shown in intermediate grey and Na\(^+\) ions as light spheres.
Lipids: Organization and Aggregation of

Figure 8. Snapshot of a DPPC bilayer with a small concentration of DPPCs replaced with PyrPC probes, see Fig. 1 (26). One finds PyrPC probes (shown in dark grey) to penetrate (interdigitate) significantly into the opposing leaflet, thus causing perturbations in both membrane monolayers. Water is not shown for clarity.


Further Reading

See Also
Passive Diffusion Across Membranes
Lipid Domains, Chemistry of Membranes, Fluidity of Lipids, Chemical Diversity of
Lipids constitute a diverse and important group of biomolecules. Most lipids can behave as lyotropic liquid crystals. In the presence of water, they self-assemble in a variety of phases with different structure and geometry. The lipid polymorphic and mesomorphic behavior, i.e., their ability to form various ordered, crystalline, gel, or liquid-crystalline phases as a function of water content, temperature, and composition, is one of the most intriguing features of lipid-water systems. The mutual transformations between these phases and their physiologic implications are the subject of this article.

**Biologic Background**

Lipids have varied biologic roles: in energy storage and fat digestion, as enzyme cofactors, electron carriers, light-absorbing pigments, intracellular messengers, hormones, constituents of the pulmonary surfactant and the skin stratum corneum, and so forth. However, the issue of their phase behavior in water is particularly relevant and most often has been discussed in relation to their major structural function—as building blocks of the biologic membranes. The major discovery in the field of biologic membranes is undoubtedly the finding that the biomembrane is a liquid-crystalline lipid bilayer with embedded proteins (1–3). This so-called fluid-mosaic model (1) has been the central paradigm in membrane biology for more than three decades and has been very successful in rationalizing a large body of experimental observations. The model includes two references to the lipid phase state—liquid crystalline and bilayer—both of which are of vital importance for the proper membrane functioning. The liquid-crystalline phases can form in stacks with interbilayer aqueous spaces typically ~1–2 nm and thus build up the best-known lipid phase—the lamellar liquid-crystalline phase Lα. However, in addition to the Lα phase, the membrane lipids also can form a large variety of other phases (subgel, gel, cubic, hexagonal, and micellar). All these phases are interrelated and transform into each other via different kinds of phase transitions. On the temperature scale, the existence range of the Lα phase typically is limited from above by lamellar–nonlamellar transitions into cubic, hexagonal, and micellar phases and limited from below by fluid-solid transitions into gel and subgel phases. A remarkable property of the lipid dispersions from a biologic viewpoint is that the transition temperatures that limit the stability ranges of the different phases can be altered by tens of degrees by varying the composition of the lipid-water system. The possibility of modulating the lipid phase behavior in very broad limits by varying the lipid composition seems to represent the basis of important regulatory mechanisms involved in the biomembrane responses to external stimuli, such as changes in the environmental conditions, and the basis for the regulation of various membrane-associated processes (see the sections that follow).
Nonlamellar lipid phases and biologic membranes

Several classes of lipids common for the biomembranes can form inverted nonlamellar phases under physiologic conditions (4). The principle ones are phosphatidylethanolamines and monogalactosyldiglycerides. Also, cardiolipins and phosphatidic acids can form inverted phases in the presence of divalent cations, and phosphatidylserines and phosphatidic acids both form inverted phases at low pH. Moreover, biomembrane lipid extracts and membrane-mimicking lipid compositions form nonlamellar phases if heated above physiologic temperatures, dehydrated, or treated with divalent cations (5–7).

Recent developments show that the ability of lipids to form nonlamellar structures seems to be a prerequisite for important membrane-associated cell processes (5). It has been demonstrated, for example, that prokaryotic organisms maintain a delicately adjusted balance between lamellar-forming and nonlamellar-forming lipids (6). Growing evidence suggests that nonlamellar-forming membrane lipids play essential roles in many aspects of membrane functioning. Short-lived nonbilyayer structures are supposed to mediate the processes of fusion and fission, and long-lived bilyayer structures with a small radius of curvature occur in some types of biologic membranes (e.g., endoplasmic reticulum, inner mitochondrial membrane, and prolamellar bodies). Membrane phase transitions take place in the course of some cellular processes (8–11). For example, the action of anesthetic agents is believed to correlate with a lamellar-cubic transition in membranes (12). The prolamellar bodies in the etioplasts of dark-grown seedlings are organized into a cubic lipid phase; they undergo a light-induced phase transition to a lamellar phase—the thylakoid membranes of chloroplasts. Cubic patterns have been inferred from the electron micrographs of many cytomembranes (13, 14). Thus, the study of the roles played by membrane lipids, the functional lipidomics (see Lipidomics), is becoming increasingly important in membrane biology.

Modulation of membrane protein activity

A biologic reason for the abundance of nonlamellar lipids in membranes is that they possess the ability to modulate the activities of membrane proteins (15, 16). It has been recognized that membranes exist in a state of curvature frustration, which may be sufficiently large to have significant effect on certain membrane-associated cell processes (5). It shows the close relation between fusion and the transition from lamellar into bilayer cubic and hexagonal nonlamellar phases if heated above physiologic temperatures, dehydrated, or treated with divalent cations (5–7).

Lipid rafts

Nonuniform lipid distribution in membranes: lipid rafts

Growing awareness regarding the nonuniform lateral and transmembrane organization of lipids in membranes (9, 27, 28) has given rise to new developments of the fluid-mosaic biomembrane model (5). A remarkable recent advance is the widely discussed formation of membrane rafts, which presents a specific case of a lipid demixing transition (29) (see Lipid rafts). Rafts are microdomains, enriched in certain kinds of lipids such as cholesterol, sphingomyelin, saturated glycerophospholipids, and glycosphingolipids, that reside in the so-called liquid-ordered phase, which is immiscible with the regular liquid-crystalline (disordered) phase. Whereas the in vivo evidence of lipid rafts...
still remains a controversial topic, the liquid-ordered mi-
crodomain and the cases of liquid-liquid immiscibility have attracted much attention both as novel features in the membrane models and for their potential importance in phe-
nomena such as membrane signaling and differential trafficking of various membrane components.

Membrane adaptation and protection in extreme external conditions

External temperature variations, for example, seasonal variations, pose a severe challenge for the maintenance of normal physio-
logic activity in poikilothermic organisms, which cannot regu-
late body temperature and in plants and bacteria. The dominant hypothesis that accounts for acquired protection to low tempera-
ture damage relates to membrane lipid composition. Cold condi-
tioning generally leads to an increased proportion of unsaturated fatty acids in the major phospholipid classes. A consequence of increased chain saturation strongly reduces the gel-liquid-
crystalline phase transition temperature of lipids (see the section entitled "Gel-liquid-crystalline phase transition"). This phenomenon, in the context of unsaturated fatty acids serving as an adaptation mechanism that allows the membrane to maintain its physiologic behavior at low temperatures, high salinity, and dehydration, is attained by means of synthesis of water-soluble compounds. Low-molecular compounds with various chemical structures—saccharides, polyols, and amino acids—are known to be natural protectors and osmoregulators. It is important to note that all protectors modulate the membrane phase behav-
ior in a virtually identical way. Despite their different chemical structures, all natural cryo- and hypotonic protectors are kosmotropic agents (water-structure makers) that have been shown to stabi-
"Lung surfactant"

Lung surfactant forms a layer at the surface of lung alveoli, thereby preventing alveolar collapse. The alveolar surfactant consists of about 90% lipids, mainly phospholipids and cholesterol. After secretion from the epithelial cells, the surfactant lipids form lamellar bodies with spherically concen-
tric bilayers. During their release into the alveolar space, lamellar bodies swell, reorganize, and transform into a charac-
teristic cross-hatched structure termed tubular myelin. It exhibits complex topology, fitting an infinite periodic surface with a tetragonal structure (39). This structure supplies lipids to the surface, thus regulating the surface tension of lung alveoli.

Phase Transitions in Lipid-Water Systems

Lipids are amphiphilic molecules, which self-assemble in a va-

- 439 -
Lipids, Phase Transitions of

I. Lamellar phases: (A) subgel, Lc; (B) gel, untitled chains, Lβ; (C) gel, tilted chains, Lβ'; (D) rippled gel, Lβ'; (E) fully interdigitated gel, Lβ int; (F) partially interdigitated gel; (G) mixed interdigitated gel; (H) liquid crystalline, Lα.

II. Mesomorphic (liquid crystalline) aggregates of various topology: (A) spherical micelles; (B) cylindrical micelles (tubules); (C) disks; (D) inverted micelles; (E) a fragment of a rhombohedral phase; (F) lamellar; (G) inverted hexagonal phase; (H) inverted micellar cubic phase; (I) bilayer cubic; Lm3m phase; (J) bilayer cubic; Pn3m phase; (K) bilayer cubic; Ia3d phase.

Figure 1 Examples of various lipid phases: I. Lamellar phases: (A) subgel, Lc; (B) gel, untitled chains, Lβ; (C) gel, tilted chains, Lβ'; (D) rippled gel, Lβ'; (E) fully interdigitated gel, Lβ int; (F) partially interdigitated gel; (G) mixed interdigitated gel; (H) liquid crystalline, Lα. II. Mesomorphic (liquid crystalline) aggregates of various topology: (A) spherical micelles; (B) cylindrical micelles (tubules); (C) disks; (D) inverted micelles; (E) a fragment of a rhombohedral phase; (F) lamellar; (G) inverted hexagonal phase; (H) inverted micellar cubic phase; (I) bilayer cubic; Lm3m phase; (J) bilayer cubic; Pn3m phase; (K) bilayer cubic; Ia3d phase.

Lipid phase nomenclature

Lipid polymorphic and mesomorphic phases generally are characterized by their 1) symmetry in one, two, or three dimensions, 2) hydrocarbon chain ordering and specific chain arrangements in the ordered gel and crystalline phases, and 3) type (normal or inverted) for the curved mesomorphic phases. For four decades, the nomenclature introduced by Luzzati (40) has been used to designate lipid phases. Because of the rapid growth of the number of new lipid phases identified in the 1980s and 1990s, the introduction of new simplified notation has been attempted as part of the Lipid Data Bank project (41), but still the Luzzati scheme is used mostly in the literature. In the latter scheme, lattice periodicity is characterized by uppercase Latin letters: L for a one-dimensional lamellar lattice; H for a two-dimensional hexagonal lattice; P for a two-dimensional oblique or rectangular lattice; T, R, and Q for the three-dimensional rectangular, rhombohedral, and cubic lattices, with space groups specified according to the International Tables (42). A lowercase Greek or Latin subscript is used as a descriptor for the chain conformation: α for disordered (liquid crystalline), β for ordered (gel), γ for ordered tilted, and C for crystalline (subgel). Roman numerals are used to designate the aggregate type: I for the oil-in-water (normal) type, and II for the water-in-oil (inverted) type.

Phase transition types

Temperature and water content are primary variables in the lipid-water systems, responsible for their thermotropic and lyotropic phase behavior, respectively.
Lamellar crystalline (subgel) L₃ phase transforms into a lamellar gel L₂ phase at a higher temperature, the latter phase undergoes a melting transition into the lamellar liquid-crystalline L₁ phase. With an additional increase of temperature, a series of mesomorphic phase transitions follow the sequence: bilayer cubic Q₃II ↔ inverse hexagonal H₁II ↔ micellar cubic Q₃MI ↔ micellar M₁II. Some lipids can form two or more modifications of a given phase, for example, gel phases of different structures (interdigitated, noninterdigitated, tilted, rippled, etc., see Fig. 1-II j, k, l, etc., see Fig. 3a) and mesomorphic cubic phases of different topology (in P₃m, P₃m, etc., see Fig. 3b I, j, k, and l). Intermediate lipid phases have been reported as well, for example, the liquid-ordered phase has attracted much attention in recent years because of its relevance to the functional lipid rafts in membranes (29) (see lipid rafts).

From a biologic viewpoint, of greatest interest are the transitions that involve the physiologically important lamellar liquid-crystalline phase, namely, the gel—liquid-crystalline (melting) transition, and the lamellar—nonlamellar mesomorphic transitions.

**Gel—liquid-crystalline phase transition**

The lamellar gel—lamellar liquid-crystalline (L₃ → L₁) phase transition, frequently also referred to as (chain-)melting, order—disorder, solid—fluid, or main transition, is the major energetic event in the lipid bilayers and takes place with a large enthalpy change. It is associated with rotameric disordering of the hydrocarbon chains, increased headgroup hydration, and increased intermolecular entropy (45). The energy required to expand the hydrocarbon chain region against attractive van der Waals interactions (volume expansion) and to increase the bilayer area (increased hydrophobic exposure at the polar—apolar interface) contributes to the large transition enthalpy change.

The melting gel—liquid-crystalline transitions in fully hydrated lipids are accompanied by large increases in lipid surface area (~25%) and specific volume (~4%). In calorimetric measurements they manifest as sharp, narrow, high-heat capacity peaks with an enthalpy of ~20–40 kJ/mol (46, 47). A too, large volume fluctuations give rise to a strong increase of the isothermal bilayer compressibility at the melting transition temperature (48).

Because of a dramatic increase of the bending elasticity, large bilayer undulations (anomalous swelling) have been observed at the melting transition (49). The temperature of the chain-melting transition is determined largely by the hydrocarbon chains. The longer and more saturated they are, the higher the transition temperature (Fig. 3c). A summary of the phase transition temperatures of the major membrane lipid classes with different chain lengths is given in Table 1 (33, 46, 52, 55, 56). Most membrane lipids have two different hydrocarbon chains, usually one saturated and one unsaturated; most common are the glycerophospholipids with a saturated sn-1 chain typically 16–18 carbon atoms long and an unsaturated sn-2 chain typically 18–20 carbon atoms long.

The gel—liquid-crystalline (L₃ → L₁) transition temperatures of mixed-chain phosphatidylcholines are summarized in Table 2.
Figure 3  Dependence of the phase transitions temperature on lipid chemical structure and aqueous phase composition: (a) hydrocarbon chain length dependence of the $L_{\beta} - L_{\alpha}$ (black squares) and $L_{\alpha} - H_{\text{II}}$ (open circles) phase transition temperatures in saturated diacyl phosphatidylethanolamines (53, 73); (b) dependence of the $L_{\beta} - L_{\alpha}$ (black squares) and $L_{\alpha} - H_{\text{II}}$ (open circles) phase transition temperatures on the degree of N-methylation of the polar headgroup of ditetradecylphosphatidylethanolamine (reproduced with permission from (73); copyright (1983) American Chemical Society); (c) dependence of the $L_{\beta} - L_{\alpha}$ phase transition temperature on the double-bond position for dioctadecenoyl phosphatidylcholine bilayers (58); (d) dependence of the phase transition temperatures of dihexadecyl phosphatidylethanolamine on the concentration of sucrose (open squares) and NaSCN (full circles) (reproduced from (33) with kind permission of Springer Science and Business Media); (e) pH dependence of the chain-melting transition temperature of dimyristoylglycerophospholipid bilayers (superscripts give the lipid charge; abbreviations as in Table 1; MPA, methylphosphatidic acid) (reproduced with permission from (67)); (f) pH dependence of the gel-fluid (circles) and lamellar-hexagonal (squares) phase transition temperatures of didodecylphosphatidylethanolamine; the dashed line indicates the appearance of additional lines in the region of the lamellar-hexagonal transition (reproduced with permission from (68); copyright (1983) American Chemical Society).

It is evident from these data that altering the lipid chain length and unsaturation modulates the lipid phase state in very broad limits and therefore provides the basis of a mechanism for membrane adaptation to large fluctuations in the environmental temperature (see the section entitled "Membrane adaptation and protection in extreme external conditions").

**Formation of nonlamellar phases in membrane lipids**

Dispersions of double-chain nonlamellar membrane lipids most frequently display a lamellar-inverted hexagonal, $L_{\alpha} - H_{\text{II}}$, phase transition. In some instances, they also can form inverted phases of cubic symmetry. The membrane elastic energy plays an important role in lamellar–nonlamellar transformations (15, 23).

The $L_{\alpha} - H_{\text{II}}$ transition may be considered a result of competition between the spontaneous tendency of the lipid layers to bend and the resulting hydrocarbon chain packing strain; thus, membranes exist in a state of frustrated curvature stress (17). Respectively, the $L_{\alpha} - H_{\text{II}}$ transition is believed to be driven by the relaxation of the curvature of the lipid monolayers toward their spontaneous curvature. Conversely to the $L_{\beta} - L_{\alpha}$ transition, the $L_{\alpha} - H_{\text{II}}$ transition temperature decreases with the hydrocarbon chain length increase (Fig. 3a). At sufficiently long chains,
Lipids, Phase Transitions of

Table 1 Gel — liquid-crystalline and lamellar — nonlamellar phase transition temperatures [°C] of fully hydrated lipids as a function of the lipid polar head group and hydrocarbon chain length (33, 46, 52, 55, 56).

<table>
<thead>
<tr>
<th>Headgroup</th>
<th>PC</th>
<th>PE</th>
<th>PG</th>
<th>PS</th>
<th>PA</th>
<th>PI</th>
<th>CL</th>
<th>Glc</th>
<th>Gal</th>
<th>Mal</th>
<th>N-Sph</th>
<th>N-Sph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chain sn-1/sn-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gel — liquid — crystalline transition (Lβ → Lα)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:0/10:0</td>
<td>−5.7</td>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:0/11:0</td>
<td>−13.9</td>
<td>16.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:0/12:0</td>
<td>−1.9</td>
<td>31.3</td>
<td>−2</td>
<td>14.2</td>
<td>32</td>
<td>25.4</td>
<td>19.5</td>
<td>26.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:0/13:0</td>
<td>13.5</td>
<td>42.1</td>
<td>10.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:0/14:0</td>
<td>23.4</td>
<td>50.4</td>
<td>24</td>
<td>35.4</td>
<td>54</td>
<td>20</td>
<td>40</td>
<td>40.5</td>
<td>48.7</td>
<td>40.9</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>15:0/15:0</td>
<td>33.7</td>
<td>58.4</td>
<td>33</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:0/16:0</td>
<td>41.7</td>
<td>64.4</td>
<td>41.5</td>
<td>51.4</td>
<td>65</td>
<td>58</td>
<td>57.2</td>
<td>61.6</td>
<td>56.6</td>
<td>42</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>17:0/17:0</td>
<td>48.6</td>
<td>70.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:0/18:0</td>
<td>54.5</td>
<td>74.2</td>
<td>54</td>
<td>63.7</td>
<td>75.4</td>
<td>70</td>
<td>68.4</td>
<td>73.5</td>
<td>66.7</td>
<td>44.5</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>19:0/19:0</td>
<td>60.2</td>
<td>79.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20:0/20:0</td>
<td>65.3</td>
<td>83.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21:0/21:0</td>
<td>70.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22:0/22:0</td>
<td>73.6</td>
<td>90.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23:0/23:0</td>
<td>77.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24:0/24:0</td>
<td>80.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:1ω7</td>
<td>−4.0</td>
<td></td>
<td>−33.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:1ω9</td>
<td></td>
<td></td>
<td>−18.0</td>
<td>−7.3</td>
<td>−18.3</td>
<td>−11.0</td>
<td>−11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1ω7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1ω9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:2ω9,12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lamellar — nonlamellar transition (Lα → HII, unless otherwise indicated)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:0/14:0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:0/15:0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:0/16:0</td>
<td>118.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:0/17:0</td>
<td>107.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:0/18:0</td>
<td>101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:0/19:0</td>
<td>97.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20:0/20:0</td>
<td>94.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22:0/22:0</td>
<td></td>
<td>90.0°</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:0/18:1ω9</td>
<td>70.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:1ω9/16:1ω9</td>
<td>43.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1ω9/18:1ω9</td>
<td>8.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:0/18:1ω9</td>
<td>62.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1ω9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20:4ω5,8,11,14</td>
<td>32.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Lβ → HII;
1Lα → QII;
2QII → HII;
3Lα → HII; transition PC, diacylphosphatidylcholines; PE, diacylphosphatidylethanolamines; PG, diacylphosphatidylglycerols; PS, diacylphosphatidylserines; PA, diacylphosphatidic acids; PI, diacylphosphatidylinositols; CL, cardiolipins; Glc, diacylglycosylglycerols; Gal, diacylgalactosylglycerols; Mal, dialkylmaltosylglycerols; N-Sph PC, sphingomyelins; N-Sph Gal, galactocerebrosides (for the sphingolipids, chain length refers to the single fatty acid chain).
Table 2 Decrease of the gel–liquid crystalline (L_β→L_α) transition temperatures of fully hydrated acyl-chain phosphatidylcholines with increasing sn-2 chain unsaturation (33)

<table>
<thead>
<tr>
<th>Chains, sn-1/sn-2</th>
<th>Temperature [°C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:0/16:0</td>
<td>41.7</td>
</tr>
<tr>
<td>16:0/16:1c9</td>
<td>30.0</td>
</tr>
<tr>
<td>16:0/18:0</td>
<td>49.0</td>
</tr>
<tr>
<td>16:0/18:1c9</td>
<td>2.5</td>
</tr>
<tr>
<td>16:0/18:2c9,12</td>
<td>-19.6</td>
</tr>
<tr>
<td>16:0/20:0</td>
<td>51.3</td>
</tr>
<tr>
<td>16:0/20:4c5,8,11,14</td>
<td>-22.5</td>
</tr>
<tr>
<td>16:0/22:0</td>
<td>52.8</td>
</tr>
<tr>
<td>16:0/22:1c13</td>
<td>11.5</td>
</tr>
<tr>
<td>16:0/22:4c7,10,13,16,19</td>
<td>-3.0</td>
</tr>
<tr>
<td>18:0/18:0</td>
<td>54.5</td>
</tr>
<tr>
<td>18:0/18:1c9</td>
<td>6.9</td>
</tr>
<tr>
<td>18:0/18:2c9,12</td>
<td>-14.4</td>
</tr>
<tr>
<td>18:0/18:3c9,12,15</td>
<td>-12.3</td>
</tr>
<tr>
<td>18:0/20:0</td>
<td>60.4</td>
</tr>
<tr>
<td>18:0/20:1c11</td>
<td>13.2</td>
</tr>
<tr>
<td>18:0/20:2c11,14</td>
<td>-5.4</td>
</tr>
<tr>
<td>18:0/20:3c8,11,14</td>
<td>-9.3</td>
</tr>
<tr>
<td>18:0/20:4c5,8,11,14</td>
<td>-12.9</td>
</tr>
<tr>
<td>18:0/20:5c5,8,11,14,17</td>
<td>-10.4</td>
</tr>
<tr>
<td>18:0/22:0</td>
<td>61.9</td>
</tr>
<tr>
<td>18:0/22:1c13</td>
<td>19.6</td>
</tr>
<tr>
<td>18:0/22:4c7,10,13,16</td>
<td>-8.5</td>
</tr>
<tr>
<td>18:0/22:5c4,7,10,13,16</td>
<td>-6.4</td>
</tr>
<tr>
<td>18:0/22:6c4,7,10,13,16,19</td>
<td>-3.8</td>
</tr>
<tr>
<td>18:0/24:0</td>
<td>62.7</td>
</tr>
<tr>
<td>18:0/24:1c15</td>
<td>31.8</td>
</tr>
<tr>
<td>20:0/18:0</td>
<td>57.5</td>
</tr>
<tr>
<td>20:0/18:1c9</td>
<td>11.5</td>
</tr>
<tr>
<td>20:0/20:0</td>
<td>65.3</td>
</tr>
<tr>
<td>20:0/20:1c11</td>
<td>20.5</td>
</tr>
<tr>
<td>20:0/20:2c11,14</td>
<td>5.4</td>
</tr>
<tr>
<td>20:0/20:3c11,14,17</td>
<td>1.8</td>
</tr>
<tr>
<td>20:0/20:4c5,8,11,14</td>
<td>-7.5</td>
</tr>
<tr>
<td>20:0/22:0</td>
<td>69.6</td>
</tr>
<tr>
<td>20:0/22:1c13</td>
<td>29.2</td>
</tr>
<tr>
<td>20:0/24:0</td>
<td>70.6</td>
</tr>
<tr>
<td>20:0/24:1c15</td>
<td>36.6</td>
</tr>
<tr>
<td>22:0/18:0</td>
<td>58.6</td>
</tr>
<tr>
<td>22:0/18:1c9</td>
<td>15.1</td>
</tr>
<tr>
<td>22:0/20:0</td>
<td>67.7</td>
</tr>
<tr>
<td>22:0/20:1c11</td>
<td>22.9</td>
</tr>
<tr>
<td>22:0/22:0</td>
<td>73.6</td>
</tr>
<tr>
<td>22:0/22:1c13</td>
<td>32.8</td>
</tr>
<tr>
<td>22:0/24:0</td>
<td>77.1</td>
</tr>
<tr>
<td>22:0/24:1c15</td>
<td>41.7</td>
</tr>
<tr>
<td>24:0/18:0</td>
<td>58.9</td>
</tr>
<tr>
<td>24:0/18:1c9</td>
<td>20.7</td>
</tr>
<tr>
<td>26:0/18:0</td>
<td>68.4</td>
</tr>
<tr>
<td>24:0/20:0</td>
<td>24.5</td>
</tr>
</tbody>
</table>

The cis-double bond causes the biggest transition temperature drop to occur, whereas additional increases of chain unsaturation have much smaller effects.
long-chain glycolipids, a direct Lc bic phases with space groups greatest interest are the inverted bicontinuous or bilayer cu-
ning on heating may intervene in the cooling phase sequence the phase sequence. Interesti ngly, intermediate phases miss-
glycerols (52, 55). In many cases, a Q exotherm.
small, if any, latent heats. Alth ough energetically inexpensive,
the α1 transition even is ob-

Table 3 Examples of heating, cooling, and isothermal phase sequences in lipid dispersions (44)

<table>
<thead>
<tr>
<th>Lipid</th>
<th>Scan direction</th>
<th>Phase sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPPC</td>
<td>heating</td>
<td>Lc13.5°C → Lp 30°C → Pcl 41.5°C → Lc1</td>
</tr>
<tr>
<td></td>
<td>cooling</td>
<td>Lc12 → Lc1</td>
</tr>
<tr>
<td></td>
<td>isothermal equilibration</td>
<td>-8°C</td>
</tr>
<tr>
<td>DLPE</td>
<td>1st heating</td>
<td>Lc145°C → Lp 30°C → Lc12</td>
</tr>
<tr>
<td></td>
<td>cooling</td>
<td>Lc12 → Lc1</td>
</tr>
<tr>
<td></td>
<td>isothermal equilibration</td>
<td>2°C, 9 days</td>
</tr>
<tr>
<td></td>
<td>2nd heating</td>
<td>Lc150°C → Lc1</td>
</tr>
<tr>
<td>DMPE</td>
<td>1st heating</td>
<td>Lc156.6°C → Lp 36°C → Lc1</td>
</tr>
<tr>
<td></td>
<td>cooling</td>
<td>Lc11 → Lc1</td>
</tr>
<tr>
<td></td>
<td>isothermal equilibration</td>
<td>2°C, 9 days</td>
</tr>
<tr>
<td>DOPE</td>
<td>heating</td>
<td>Lc155°C → Lp 68°C → HII</td>
</tr>
<tr>
<td></td>
<td>cycling (n &gt; 100)</td>
<td>Q11 → Q12 → Q13</td>
</tr>
<tr>
<td></td>
<td>deep cooling (&lt; 1°C)</td>
<td>Lc165°C → Lc1</td>
</tr>
<tr>
<td>DOPE-Me</td>
<td>heating (&gt; 1°C)</td>
<td>Lc133°C → Lp 52°C → HII</td>
</tr>
<tr>
<td></td>
<td>isothermal equilibration</td>
<td>55°C, 20h</td>
</tr>
<tr>
<td></td>
<td>deep cooling (&lt; 20°C)</td>
<td>Lc172°C → Lc1</td>
</tr>
<tr>
<td>14-Gal</td>
<td>1st heating</td>
<td>Lc165°C → Lp 68°C → Lc12</td>
</tr>
<tr>
<td></td>
<td>cooling</td>
<td>Lc130°C → Lp 47°C → Lc1</td>
</tr>
<tr>
<td></td>
<td>2nd heating</td>
<td>Lc165°C → Lp 38°C → Lc12</td>
</tr>
<tr>
<td>18-Gal</td>
<td>1st heating</td>
<td>Lc165°C → Lp 68°C → Lc1</td>
</tr>
<tr>
<td></td>
<td>cooling</td>
<td>Lc130°C → Lp 47°C → Lc1</td>
</tr>
<tr>
<td></td>
<td>2nd heating</td>
<td>Lc165°C → Lp 38°C → Lc12</td>
</tr>
<tr>
<td></td>
<td>isothermal equilibration</td>
<td>60–70°C, 1 h</td>
</tr>
</tbody>
</table>

DOPE-Me, disteryl-methylthanolamine; 14-Gal, ditetradecylgalactosylglycerol; 18-Gal, dioctadecylgalactosylglycerol.

the Lc phase is eliminated completely, and direct Lc-H II tran-
sitions take place on heating. Such direct transitions have been
observed for diacyl PEs of 22-carbon chains and monogly-
cosyldiacylglycerols of 19-20 carbon chains (Table 5). With
long-chain glycolipids, a direct Lc-H II transition even is ob-
served in which both the Lc and Lp phases are eliminated from
the phase sequence. Interestingly, intermediate phases miss-
ing on heating may intervene in the cooling phase sequence (see the section entitled “Reversibility of the phase transitions”;

Table 3.

Among the seven cubic phases so far identified in lipids, of

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
these transitions typically are rather slow. The slow formation, hysteretic behavior, and extended metastability ranges of the cubic phases create significant difficulties in their study and applications.

Inverted micellar cubic phases have been observed mainly in mixtures of double-chain polar lipids with fatty acids or diacylglycerols but also in some single-component dispersions of glycolipids (61). The most frequently observed inverted micellar cubic phase in lipids is of space group $Q^{227}(Fd	ilde{3}m)$. For medium-chain lipids ($\geq 16$ C atoms), it typically forms via an $H-II-Q-II$ transition; however, the $L_\beta$ gel phase of dC19-xylopyranosyl has been found to melt directly into the $Fd	ilde{3}m$ cubic phase (62).

Polymorphic transitions between solid lipid phases

At temperatures below the main transition, a basic equilibrium structure is the subgel (crystalline) $L_c$ phase. Its formation usually requires prolonged low-temperature incubation. In addition to the $L_c$ phase, many intermediate stable, metastable, and transient lamellar gel structures are adopted by different lipid classes—with perpendicular or tilted chains with respect to the bilayer plane, with fully interdigitated, partially interdigitated, or noninterdigitated chains, rippled bilayers with various ripple periods, and so forth. (Fig. 1). Several polymorphic phase transitions between these structures have been reported. Well-known examples of polymorphic transitions are the subtransition ($L_c-L_\beta$) and the pretransition ($L_\beta-L_\beta'$) in phosphatidylcholines (33). Recently, a polymorphic transition that included rapid, reversible transformation of the usual gel phase into a metastable, more ordered gel phase with orthorhombic hydrocarbon chain-packing (so-called Y-transition) was reported to represent a common pathway of the bilayer transformation into a subgel (crystalline) $L_c$ phase (62).

Reversibility of the phase transitions: formation of metastable phases

Because of long relaxation times, especially in the transition vicinity, lipid phase transitions often are not reversible and end up with the formation of metastable phases, which replace the equilibrium phases in cooling scans. The metastable phases can be very long-lived and display no spontaneous conversion to the ground state in sensible time scales. Many rate-limiting factors have been suggested as physical reasons that lead to the formation of metastable phases: long hydration/dehydration times, slow reformation of hydrogen-bond networks, restricted molecular motion in the low-temperature solid–solid transformations, relative stability of the interfaces between solid and fluid domains, large spatial rearrangements in lamellar–nonlamellar transitions, low rate of appearance of critical-size nuclei of the nascent phase, and arrestment in local free-energy minima. A comparison between heating and cooling phase sequences observed in aqueous dispersions of lipids shows the frequent occurrence of additional, metastable phases, which only form in a cooling direction (Table 3).
Phase transitions in lipid mixtures: phase diagrams

Composition is another important variable that strongly modulates the lipid phase behavior. The phase properties of a lipid mixture are best presented by means of a temperature-composition phase diagram. Such diagrams may be constructed by using various experimental techniques (63); a most appropriate one is the differential scanning calorimetry. In addition to being non-perturbing, it also has the advantage of recording not only the temperature but also the enthalpy of the phase transitions.

Various types of lipid phase diagrams reported in the literature are shown in Fig. 5 (64–73). The lens-like diagram in Fig. 5a is characteristic for lipid mixtures that are completely miscible in both gel and liquid-crystalline phases. To display such complete miscibility, the two components must be very similar structurally. This kind of diagram is typical for lipid species with the same headgroup, differing by not more than two methylene groups in their hydrocarbon chains, such as the DMPC/DPPC binary (Fig. 5a). A usual complication of the lens-type diagram is the frequently occurring solid-state miscibility gap, where the mixture separates into two solid phases of different composition. In mixtures of lipids with sufficiently different structures, the miscibility gap may overlap with the region of the solid–liquid-crystalline phase coexistence and give rise to eutectic (Fig. 5b) or peritectic phase diagrams (Fig. 5d), in which single three-phase points exist. Horizontal solidus lines, reporting for such kind of behavior, have been observed for numerous lipid mixtures. Miscibility gaps also may occur in the liquid-crystalline phase of certain lipid mixtures. A phase diagram with a liquid-liquid immiscibility region is the
**Lipids, Phase Transitions of**

Monotropic phase diagram shown in Fig. 5e with a monotropic triple point of coexistence of one solid and two liquid phases. Deviations from ideal mixing may occur not only with a tendency for clustering of the like molecules and eventually leading to phase separation but also when contacts between unlike molecules are preferred - when the nearest-neighbor pairs tend to be made up of unlike molecules (a "chessboard" arrangement). Such mixtures often display phase diagrams with an upper isocencentration (azeotropic) point, such as the DPPC/palmitic acid diagram shown in Fig. 5g. Except for the phosphatidylycholine/cholesterol acid mixtures, phase diagrams with an upper isocencentration point are typical for mixtures that contain a charged lipid. An example of a phase diagram with a lower isocencentration point is shown in Fig. 5f.

The DPPC/cholesterol phase diagram in Fig. 5h contains a critical point. It is related to the existence of a peculiar, liquid ordered (lo) phase in the mixtures, which is believed to be the prototype of the lipid rafts (see Lipid rafts).

**Figs. 5I–L** illustrate the phase behavior in lipids/water mixtures. A quenched phase diagrams of monoglycerols of various chain length and saturation show the effect of molecular geometry on the lipid phase behavior. Increasing the chain length from C16 to C20 introduces cubic and hexagonal phases between the lamellar liquid-crystalline and the liquid L2 phase (72). In the unsaturated C18:1 glyceryl monostearate, two cubic phases, Ia3d (I) and Pn3m (D), and the inverted hexagonal HII phase form (73) (Fig. 5I).

Ternary phase diagrams are another important tool for the characterization of the phase properties of complex lipid mixtures. This kind of diagram has proven especially useful recently in the analysis of domains in model systems. Ternary mixtures of one phospholipid that has a relatively high melting temperature and another phospholipid that has a relatively low melting temperature together with cholesterol are viewed as useful models for the outer leaflet of animal-cell plasma membranes. An example of a ternary phase diagram is shown in Fig. 6 (74). It illustrates the rich phase behavior displayed by ternary lipid mixtures represented in this particular case by four regions of two-phase coexistence and one region of three-phase coexistence.

**Role of the aqueous phase composition**

**The Hofmeister Effect**

The interactions of the lipid polar groups with water have an important contribution to the energy balance of a given phase. The relatively high hydration, typical for the membrane lipids, is responsible in particular for their ability to form liquid-crystalline bilayers separated by aqueous spaces. Many lipid phase transitions take place with large changes in the lipid surface area and consequently in the amount of bound water. The lipid hydration is determined by the chemical structure of the polar groups, but it is essential to note that in fully hydrated systems with water in excess, the extent of the polar group hydration depends also on the state of the bulk water. On the other side, various low-molecular solutes are known for their ability to modulate strongly the bulk water structure: "water-structure makers" (kosmotropes) and "water-structure breakers" (chaotropes). It thus turns possible that, even without direct interaction with the lipid polar heads, solutes largely can modulate the properties of the lipid-water interface and hence the lipid phase behavior. Changes in the aqueous phase composition can shift substantially the temperature regions of stability of the different lipid phases and induce or suppress the formation of certain phases. Indirect solute effects of such kind on the interfacial properties, generally termed the Hofmeister effect, have been found in many lipid-water phases (34). Any studies on the nature of the Hofmeister effect indicate that it results from an interplay of electrostatics, dispersion forces, thermal motion, fluctuations, hydration, polarizability, ion size effects, and the impact of interfacial water (75, 76).

According to their effect on the lipid phase transitions, the Hofmeister solutes fall into two categories: 1) chaotropic solutes that favor the formation of the lamellar liquid-crystalline phase Lα, at the expense of the neighboring HII and Lβ phases and 2) kosmotropic solutes that favor the formation of the HII and Lβ phase at the expense of the Lα phase. Their effects are described correctly by an equation of the Clapeyron–Clausius type between phase transition temperature and solute concentration (34). The sign and magnitude of the transition shifts induced by the different solutes depend on the solute ability to distribute unevenly between interlamellar and free water. Kosmotropic solutes tend to minimize the area of the lipid-water contact. They suppress the Lα phase, as it has the largest surface area in contact with water. At a high enough concentration of kosmotropic solutes, the latter phase may disappear completely from the phase diagram. This disappearance is precisely what is seen with sucrose, trehalose, proline, and some salts and is consistent with the opposite effect caused by chaotropic solutes (Fig. 3d). The addition of chaotropic solutes also can induce the appearance...
Effect of pH
Changes in pH modulate the lipid phase behavior as a con-
sequence of protonation/deprotonation of the head group,
which results in a change of the surface charge of the mem-
brane (77). They also modify the surface polarity and hydration. Typ-
ical protonation decreases lipid hydration and increases the
main transition temperature (53). The effects of pH titration on
the chain-melting transition temperature $T_m$ of dimyristoyl
phospholipids is illustrated in Fig. 3e which shows that sin-
gle protonation increases the melting transition temperature
about 5–15°C.

The shifts of the lamellar-hexagonal transition during lita-
rations are greater and in the opposite direction relative to changes
of $T_m$. Thus, for didodecyl PE, the lamellar-hexagonal transi-
tion decreases by 43°C during phosphate protonation (pK$_a$ = 1.9)
and by 50°C during amine protonation (pK$_a$ = 3.9), whereas for
$T_m$ these shifts are 6°C and 15°C, respectively, in the opposite
direction (Fig. 3f) (54).

Tools and Techniques
From the analysis of the data in the LIPIDAT database (41),
more than 150 different methods and method modifications have
been used to collect data related to the lipid phase transitions.
Almost 90% of the data is accounted for by less than 10 meth-
ods. Differential scanning calorimetry strongly dominates the
field with two thirds of all phase transition records. From the
other experimental techniques, various fluorescent methods ac-
count for ~10% of the information records. X-ray diffraction,
nuclear magnetic resonance (NMR), Raman spectroscopy, elec-
tron spin resonance (ESR), infrared (IR) spectroscopy, and po-
nuclear magnetic resonance (NMR), Raman spectroscopy, elec-
tron spin resonance (ESR), infrared (IR) spectroscopy, and po-
larizing microscopy each contribute to about or less than 2–3% of
the chain-melting transition temperature

References
1. Singer SJ, Nicolson GL. Fluid mosaic model of structure of
2. Simhrick EJ, McConnell HM. Lateral phase separation in phos-
3. Traballo R, Sackmann E. Studies of crystalline-lipid crystalline
phase-transition of lipid model membranes.3. Structure of a
sterol-lectin system below and above lipid-phase transition. J.
4. Lewis RNH, Mannock DA, R-Delany RN. Membrane lipid
molecular structure and polymorphism. In: Current Topics in
Membranes, Volume 44. Lipid Polymorphism and Membrane
p. 25–102
5. Mouritsen OG, Andersen OS, eds. in Search of a New Biomem-
6. Riitonen L, Lindblom G. Regulation of lipid composition in biologi-
cal membranes—biophysical studies of lipids and lipid synthesizing
7. Konyava R, MacDonald KC. Natural lipid extracts and biomem-
brane-mimicking lipid compositions are disposed to form non-
lamellar phases, and they release DNA from lipoplexes most
2382.
8. Bilstien RI. A statistical-thermodynamic view of cooperative
structural changes in phospholipid bilayer-component of cell-membranes-A perspective.
9. Havel Jr. Thermal adaptation in biological-membranes—Is homo-

genic adaptation the explanation. Annu. Rev. Physiol. 1995; 57:
19–42.
10. Heimburg T, Jackson AD. The thermodynamics of general anes-
11. Larson K. Anesthetic effect and a lipid bilayer transition involv-
12. Landh T. From entangled membranes to ectodytic membranes—
Cubic membranes as subcellular space organizers. FEBS Lett.
13. Aalmeperi ZA, Kohlwein SD, Dong Y. Cubic membranes: a
legend beyond the flatland of cell membrane organization. J. Cell
14. Gruner SM. Coupling between bilayer curvature elasticity and
membrane protein activity. In: Biomembrane Electrochemistry
15. McIntosh TJ, Simon SA. Roles of bilayer material properties
in function and distribution of membrane proteins. Annu. Rev.
16. Gruner SM. Intrinsic curvature hypothesis for biomembrane
17. Lee AG. How lipids affect the activities of integral membrane
19. Kiinunen PKJ. On the molecular-level mechanisms of peripheral
protein-protein interactions induced by lipids forming inverted
20. Botelho AV, Gibson NJ, Thurmond RL, Wang Y, Bunn HF. Con-
formational energetics of rhodopsin modulated by nonlamellar-
21. Lee AG. Lipid-protein interactions in biological membranes: a
1612: 1–40.
22. Siegel DP. The relationship between bicontinuous inverted cubic
phases and membrane fusion. In: Bicontinuous Lipid Cubic
Lynch ML, Spicer PT, eds. 2005. Taylor & Francis Group, CRC
23. Kozlov MM, Markin VS. Possible mechanism of membrane-fus-
24. Siegel DP. The modified stalk mechanism of lamellar/inverted
phase transitions and its implications for membrane fusion. Bio-
25. Yang L, Huang HW. A rhombohedral phase of lipid containing
a membrane fusion intermediate structure. Biochim. J. 2003;
1808–1817.
Lipids, Phase Transitions of


Further Reading


See Also

Drug Delivery
Lipidomics
Lipid Rafts
Membrane Fusion, Mechanisms of
Alkaloids are highly diverse groups of natural products related only by the presence of a basic nitrogen atom located at some position in the molecule. Even among biosynthetically related classes of alkaloids, the chemical structures are often highly divergent. Although some classes of natural products have a recognizable biochemical paradigm that is centrally applied throughout the pathway, for example, the "assembly line" logic of polyketide biosynthesis (2), the biosynthetic pathways of alkaloids are as diverse as the structures. It is difficult to predict the biochemistry of a given alkaloid based solely on precedent, which makes alkaloid biosynthesis a challenging, but rewarding, area of study.

Biologic Background

Hundreds of alkaloid biosynthetic pathways have been studied by chemical strategies, such as isotopic labeling experiments (2, 3). However, modern molecular biology and genetic methodologies have facilitated the identification of alkaloid biosynthetic enzymes. This article focuses on pathways for which a significant amount of genetic and enzymatic information has been obtained. Although plant alkaloid natural products are produced by insects, plants, fungi, and bacteria, this article focuses on four major classes of plant alkaloids: the isoquinoline alkaloids, the terpenoid indole alkaloids, the tropane alkaloids, and the purine alkaloids.

In general, plant biosynthetic pathways are understood poorly when compared with prokaryotic and fungal metabolic pathways. A major reason for this poor understanding is that genes that express complete plant pathways typically are not clustered together on the genome. Therefore, each plant enzyme often is isolated individually and cloned independently. However, several enzymes involved in plant alkaloid biosynthesis have been cloned successfully, and many more enzymes have been purified from alkaloid-producing plants or cell lines (4–6). Identification and study of the biosynthetic enzymes has a significant impact on the understanding of the biochemistry of the pathway. Furthermore, genetic information also can be used to understand the complicated localization patterns and regulation of plant pathways. This article focuses on the biochemistry responsible for the construction of plant alkaloids and summarizes the biosynthetic genes that have been identified to date. Some of these pathways have been the subject of metabolic engineering studies; the results of these studies are mentioned here also. An excellent, more detailed review that covers the biochemistry and genetics of plant alkaloid biosynthesis up until the late 1990s is available also (7).

Isoquinoline Alkaloids

The isoquinoline alkaloids include the analgesics morphine and codeine as well as the antibiotic berberine (Fig. 1a). Morphine and codeine are two of the most important analgesics used in medicine, and plants remain the main commercial source of the alkaloids (8). Development of plant cell cultures of *Eschscholzia californica*, *Papaver somniferum*, and *Coptis japonica* has aided in the isolation and cloning of many enzymes involved in the biosynthesis of isoquinoline alkaloids (9).

Early steps of isoquinoline biosynthesis

Isoquinoline biosynthesis begins with the substrates dopamine and p-hydroxyphenylacetaldehyde (Fig. 1b). Dopamine is made from tyrosine by hydroxylation and decarboxylation. Enzymes that catalyze the hydroxylation and decarboxylation steps in either order exist in the plant, and the predominant pathway...
Alkaloid Biosynthesis

Figure 1  (a) Representative isoquinoline alkaloids. (b) Early biosynthetic steps of the isoquinoline pathway yield the biosynthetic intermediate (S)-reticuline, the central biosynthetic intermediate for all isoquinoline alkaloids. (c) Berberine and sanguinarine biosynthesis pathways. (d) Morphine biosynthesis. NCS, norcoclaurine synthase; 6-OMT, norcoclaurine 6-O-methyltransferase; CNMT, coclaurine N-methyltransferase (Cyp80B); NMTC, N-methylcoclaurine 3′-hydroxylase; 4′-OMT, 3′-hydroxy-N-methylcoclaurine 4′-O-methyltransferase; BBE, berberine bridge enzyme; SOMT, scoulerine 9-O-methyltransferase; CS, canadine synthase; TBO, tetrahydroprotoberberine oxidase; CHS, cheilanthifoline synthase; SYS, stylopine synthase; NMT, N-methyltransferase; NMSH, N-methylstylopine hydroxylase; P6H, protopine 6-hydroxylase; DHPO, dihydrobenzophenanthridine oxidase; RO, reticuline oxidase; DHR, dihydroreticulium ion reductase; STS, salutaridine synthase; SaR, salutaridine reductase; SaAT, salutaridinol acetyltransferase; COR, codeinone reductase.

for formation of dopamine from tyrosine is not clear. The second substrate, p-hydroxyphenylacetaldehyde, is generated by transamination and decarboxylation of tyrosine (10, 11). Condensation of dopamine and p-hydroxyphenylacetaldehyde is catalyzed by norcoclaurine synthase to form (S)-norcoclaurine (Fig. 1b). Two norcoclaurine synthases with completely unrelated sequences were cloned (Thalictrum flavum and C. japonica) and heterologously expressed in E. coli (12–14). One is homologous to iron-dependent dioxygenases, whereas the other is homologous to a pathogenesis-related protein. Undoubtedly, future experiments will shed light on the mechanism of these enzymes and on how two such widely divergent sequences can catalyze the same reaction.

One of the hydroxyl groups of (S)-norcoclaurine is methylated by a S-adenosyl methionine-(SAM)-dependent O-methyltransferase to yield (S)-coclaurine. This enzyme has been cloned, and the heterologously expressed enzyme exhibited the expected activity (15–17). The resulting intermediate is
then N-methylated to yield N-methylcoclaurine, an enzyme that has been cloned recently (18, 19). N-methylcoclaurine, in turn, is hydroxylated by a P450-dependent enzyme (CYP80B), N-methylcoclaurine (S)-hydroxylase, that has been cloned (20, 21). The 4′ hydroxyl group then is methylated by the enzyme 3’-hydroxy-N-methylcoclaurine 4′-O-methyltransferase (4′-OMT) to yield (S)-reticuline, the common biosynthetic intermediate for the berberine, benzocistanthidinone, and morphinan alkaloids (Fig. 1c). The gene for this methyl transferase also has been identified (15, 22). These gene sequences also were used to identify the corresponding T. flavum genes that encode the biosynthetic enzymes for reticuline from a cDNA library (23). At this point, the biosynthetic pathway then branches to yield the different structural classes of isoquinoline alkaloids.

Berberine biosynthesis

(S)-reticuline is converted to (S)-scoulerine by the action of a well-characterized flavin-dependent enzyme, berberine bridge enzyme (Fig. 1c). This enzyme has been cloned from several plant species, and the mechanism of this enzyme has been studied extensively (24–28). (S)-scoulerine is then O-methylated by scoulerine 9-O-methyltransferase to yield (S)-tetrahydrocolumbamine. Heterologous expression of this gene in E. coli yielded an enzyme that had the expected substrate specificity (29). A variety of O-methyl transferases also have been cloned from Thalictrum tuberosum (30). The substrate-specific cytochrome P450 oxidase canadine synthase (31) that generates the methylene dioxy bridge of (S)-canadine has been cloned recently (32). The final step of berberine biosynthesis is catalyzed by a substrate-specific oxidase, tetrahydrocolumbamine dehydrogenase, the sequence of which has not been identified yet (33).

Overproduction of berberine in C. japonica cell suspension culture with a high producing cell line (34) with reported productivity of berberine reaching 7 g/L (35). This overproduction is one of the first demonstrations of production of a benzylisoquinoline alkaloid in cell culture at levels necessary for economic production. This cell line has facilitated greatly the identification of the biosynthetic enzymes.

Sanguinarine biosynthesis

The biosynthesis of the highly oxidized benzocistanthidinone alkaloid sanguinarine is produced in a variety of plants and competes with morphine production in opium poppy. The pathway to sanguinarine has been elucidated at the enzymatic level (Fig. 1c) (36). Sanguinarine biosynthesis starts from (S)-scoulerine, as in berberine biosynthesis. Methyleneoxy bridge formation then is catalyzed by the P450 chelanthidine synthase (37). A second P450 enzyme, stylopine synthase, catalyzes the formation of the second methylenedioxy bridge of stylopine (37). Stylopine synthase from E. californica has been cloned recently (38). Stylopine then is N-methylated by (S)-tetrahydroprotoberberine cis-N-methyltransferase to yield (S)-cis-N-methylstylopine, an enzyme that has been cloned recently from opium poppy (39). A third P450 enzyme, (S)-cis-N-methylstylopine hydroxylase, then forms protopine. Protopine is hydroxylated by a fourth P450 enzyme, protopine 6-hydroxylase, to yield an intermediate that rearranges to dihydrosanguinarine (40). This intermediate also serves as the precursor to the benzocistanthidinone alkaloid macraine (Fig. 1a). The copper-dependent oxidase dihydrobenzocistanthidinone oxidase, which has been purified (41, 42), then catalyzes the formation of sanguinarine from dihydrosanguinarine.

Additional enzymes from other benzocistanthidinone alkaloids have been cloned. For example, an O-methyl transferase implicated in palmistine biosynthesis has been cloned recently (43).

Morphine biosynthesis

The later steps of morphine biosynthesis have been investigated in P. somniferum cells and tissue. Notably, in morphine biosynthesis, (S)-reticuline is converted to (R)-reticuline, thereby epimerizing the stereocenter generated by norcoclaurine synthase at the start of the pathway (Fig. 1d). (S)-reticuline is converted to (R)-reticuline through a 1,2-dehydroreticuline intermediate. Dehydroreticuline synthase catalyzes the oxidation of (S)-reticuline to 1,2-dehydroreticuline ion (44). This enzyme has not been cloned yet but has been purified partially and shown to be membrane-associated. This intermediate then is reduced by dehydroreticuline reductase, an NADPH-dependent enzyme that stereoselectively transfers a hydride to dehydroreticuline ion to yield (R)-reticuline. This enzyme has not been characterized, but it has been cloned yet (45).

Next, the key carbon–carbon bond of the morphinan alkaloid is formed by the cytochrome P450 enzyme salutaridinol synthase. Astonishingly, this enzyme has been detected in different palms and plants (46), but the sequence has not been identified (46). The keto moiety of the resulting product, salutaridinol, then is stereoselectively reduced by the NADPH-dependent salutaridinol reductase to form salutaridinol. The enzyme has been purified (47), and a recent transcript analysis profile of P. somniferum has resulted in the identification of the clone (48). Salutaridinol acetyltransferase, also cloned, then transfers an acetyl group from acetyl-CoA to the newly formed hydroxyl group, which results in the formation of salutaridinol-N-7-acetate (49). This modification sets up the molecule to undergo a spontaneous reaction in which the acetate can act as a leaving group. The resulting product, thebaine, then is demethylated by an as yet uncharacterized enzyme to yield neopine, which exists in equilibrium with its tautomer codeinone. The NADPH-dependent codeinone reductase reduces codeinone to codeine and has been cloned (50, 51). Finally, codeine is demethylated by an uncharacterized enzyme to yield morphine.

The localization of isoquinoline biosynthesis has been investigated at the cellular level in intact poppy plants by using in situ RNA hybridization and immunofluorescence microscopy. The localization of 4′-OMT (reticuline biosynthesis), berberine bridge enzyme (sanguinarine biosynthesis), salutaridinol acetyltransferase (morphine biosynthesis), and codeinone reductase (morphine biosynthesis) has been probed. 4′-OMT and salutaridinol acetyltransferase are localized to parenchyma cells, whereas codeinone reductase is localized to laticifer cells in...
sections of capsule (fruit) and stem from poppy plants. Berberine bridge enzyme is found in parenchyma cells in roots. Therefore, this study suggests that two cell types are involved in isoquinoline biosynthesis in poppy and that intercellular transport is required for isoquinoline alkaloid biosynthesis (52). A further study, however, implicates a single cell type (sieve elements and their companion cells) in isoquinoline alkaloid biosynthesis (53, 54). Therefore, it is not clear whether transport of pathway intermediates is required for alkaloid biosynthesis or whether the entire pathway can be performed in one cell type. Localization of enzymes in alkaloid biosynthesis is difficult, and, undoubtedly, future studies will provide more insight into the trafficking involved in plant secondary metabolism.

Metabolic engineering of morphine biosynthesis

In attempts to accumulate thebaine and decrease production of morphine (a precursor to the recreational drug heroin), codeinone reductase in opium poppy plant was downregulated by using RNAi (8). Silencing of codeinone reductase results in the accumulation of (S)-reticuline but not the substrate codeinone or other compounds on the pathway from (S)-reticuline to codeine. However, overexpression of codeinone reductase in opium poppy plants did result, in fact, in an increase in morphine and other morphinan alkaloids, such as morphine, codeine, and thebaine, compared with control plants (55). Gene expression levels in low morphine-producing poppy plants have been analyzed also (56). Silencing of berberine bridge enzyme in opium poppy plants also resulted in a change in alkaloid profile in the plant latex (57).

The cytochrome P450 responsible for the oxidation of (S)-N-methylcoclaurine to (S)-3′-hydroxy-N-methylcoclaurine has been overexpressed in opium poppy plants, and morphinan alkaloid production in the latex is increased subsequently to 4.5 times the level in wild-type plants (58). A diastereomeric, suppression of this enzyme resulted in a decrease in morphinan alkaloids to 16% of the wild-type level. Notably, analysis of a variety of biosynthetic gene transcript levels in these experiments supports the hypothesis that this P450 enzyme plays a regulatory role in the biosynthesis of benzylisoquinoline alkaloids. Collectively, these studies highlight that the complex metabolic networks found in plants are not redirected easily or predictably in all cases.

Terpenoid Indole Alkaloids

The terpenoid indole alkaloids have a variety of chemical structures and a wealth of biologic activities (Fig. 2b) (59, 60). Terpenoid indole alkaloids are used as anticancer, antimalarial, and antiarrhythmic agents. Although many biosynthetic genes from this pathway remain unidentified, recent studies have correlated terpenoid indole alkaloid production with the transcript profiles of Catharanthus roseus cell cultures (61).

Early steps of terpenoid indole alkaloid biosynthesis

All terpenoid indole alkaloids are derived from tryptophan and the indole terepne secologanin (Fig. 2a). Tryptophan decarboxylase, a pyridoxal-dependent enzyme, converts tryptophan to tryptamine (62, 63). The enzyme strictosidine synthase catalyzes a stereospecific Pictet–Spengler condensation between tryptamine and secologanin to yield strictosidine. Strictosidine synthase (64) has been cloned from the plants C. roseus (65), Rauwolfia serpentina (66), and, recently, Ophiorrhiza pumila (67). A crystal structure of strictosidine synthase from R. serpentina has been reported (68, 69), and the substrate specificity of the enzyme can be modulated (70). Strictosidine then is deglycosylated by a dedicated β-glucosidase, which converts it to a reactive hemiacetal intermediate (71–73). This hemiacetal opens to form a dialdehyde intermediate, which then forms dehydrogeniisschizine. The end form of dehydrogeniisschizine undergoes a conjugate addition to produce the heterocyclic amine catharanthine (74–76). A variety of rearrangements subsequently act on dehydrogeniisschizine to yield a diversity of indole alkaloid products (77).

Ajmaline biosynthesis

The biosynthetic pathway for ajmaline in R. serpentina is one of the best-characterized terpenoid indole alkaloid pathways. Much of this progress has been detailed in a recent extensive review (78). Like all other terpenoid indole alkaloids, ajmaline, an antiarrhythmic drug with potent sodium channel-blocking properties (79), is derived from dehydrogeniisschizine (Fig. 2c).

A membrane-protein fraction of an R. serpentina extract transforms labeled strictosidine (80, 81) into saragpane-type alkaloids. The enzyme activity is dependent on NADPH and molecular oxygen, which suggests that saragpane bridge enzyme may be a cytochrome P450 enzyme. Polyneuridine aldehyde esterase hydrolyzes the polyneuridine aldehyde methyl ester, which generates an acid that decarboxylates to yield epi-vellosamine. This enzyme has been cloned from a Rauwolfia cDNA library, heterologously expressed in E. coli, and subjected to detailed mechanistic studies (82, 83).

In the next step of the ajmaline pathway, vinorine synthase transforms the saragpane alkaloid epi-vellosamine to the ajmaline alkaloid vinorine (84). Vinorine synthase also has been purified from Rauwolfia cell culture, subjected to protein sequencing, and cloned from a cDNA library (85, 86). The enzyme, which seems to be an acetyl transferase homolog, has been expressed heterologously in E. coli. Crystallization and site-directed mutagenesis studies of this protein have led to a proposed mechanism (87). Vinorine hydroxylase hydroxylates vinorine to form vomicline (88). Vinorine hydroxylase seems to be a P450 enzyme that requires an NADPH-dependent reductase. This enzyme is labile and has not been cloned yet. Next, the indolenine bond is reduced by an NADPH-dependent reductase to yield 1,2-dihydropomilene. A second enzyme, 1,2-dihydropomilene reductase, then reduces this product to...
Alkaloid Biosynthesis

Figure 2  (a) Representative terpenoid indole alkaloids. (b) Early biosynthetic steps of the terpenoid indole alkaloid pathway yield the strictosidine, the central biosynthetic intermediate for all terpenoid indole alkaloids. (c) Ajmaline biosynthesis. (d) Ajmalicine and tetrahydroalstonine biosynthesis. (e) Vindoline biosynthesis from tabersonine. TDC, tryptophan decarboxylase; STR, strictosidine synthase; SGS, strictosidine glucosidase; SB, sarpagan bridge enzyme; PNAE, polyneuridine aldehyde reductase; VS, vinorine synthase; VH, vinorine hydroxylase; VR, vomilenine reductase; DHVR, dihydrovomilenine reductase; AAE, 17-O-acetyl-ajmalanesterase; NMT, norajmaline-N-methyltransferase; T16H, tabersonine-16-hydroxylase; HTOM, 16-hydroxytabersonine-16-O-methyltransferase; NMt, N-methyltransferase; DH, dihydrovomilenine-O-methyltransferase.

Acetylnorajmaline. Partial protein sequences have been obtained for both of the purified reductases. Although several putative clones that encode these proteins have been isolated, the activity of these clones has not been verified yet (89, 90). An acetylersterase then hydrolyzes the acetyl link of acetyl norajmaline to yield norajmaline. This esterase has been purified from R. serpentina cell suspension cultures, and a full-length clone has been isolated from a cDNA library. Expression of the gene in tobacco leaves successfully yielded protein with the expected enzymatic activity (91). In the final step of ajmaline biosynthesis, an N-methyl transferase introduces a methyl group at the indole nitrogen of norajmaline. Although this enzymatic activity has been detected in crude cell extracts, the enzyme has not been characterized additionally (92).
Figure 2 (continued)

**Ajmalicine and tetrahydroalstonine**

Ajmalicine (raubasine) affects smooth muscle function and is used to help prevent strokes (93), and tetrahydroalstonine exhibits antipsychotic properties (Fig. 2d) (94). These compounds are found in a variety of plants, including *C. roseus* and *R. serpentina*. A partially purified NADPH-dependent reductase isolated from a tetrahydroalstonine that produces a *C. roseus* cell line was shown to catalyze the conversion of cathenamine, a spontaneous reaction product that results after strictosidine deglycosylation, to tetrahydroalstonine in vitro (95). A second *C. roseus* cell line contains an additional reductase that produces ajmalicine. Labeling studies performed with crude *C. roseus* cell extracts in the presence of D2O or NADPD support a mechanism in which the reductase acts on the iminium form of cathenamine (96).

**Vindoline**

Vindoline, an aspidosperma-type alkaloid produced by *C. roseus*, is a key precursor for vinblastine, an anticancer drug that is the most important pharmaceutical product of *C. roseus*. Vindoline, like ajmalicine and ajmaline, is produced from deglycosylated strictosidine. Deglycosylated strictosidine is converted to tabersonine through a series of biochemical steps for which no enzymatic information exists. More details are known about...
the six steps that catalyze the elaboration of tabersonine to vindoline (Fig. 2e) (97). Tabersonine-16-hydroxylase, a cytochrome P450, hydroxylates tabersonine to 16-hydroxy-tabersonine in the first step of this sequence and has been cloned (98, 99). The newly formed hydroxyl group is methylated by a SAM-dependent O-methyl transferase to yield 16-methoxy-tabersonine; this enzyme (16-hydroxytabersonine-16-O-methyltransferase) has been purified but not cloned (100). In the next step, hydration of a double bond by an uncharacterized enzyme produces 16-methoxy-2,3-dihydro-3-hydroxytabersonine. Transfer of a methyl group to the indole nitrogen by an N-methyl transferase yields desacetoxyvindoline. This methyl transferase activity has been detected only in differentiated plants, not in plant cell cultures (101). The resulting intermediate, desacetylvindoline, is produced by the oxoglutarate-dependent dioxygenase enzyme desacetylvindoline 4-hydroxylase. This enzyme has been cloned and is absent from plant cell cultures (102). In the last step, desacetylvindoline is acetylated by desacetylvindoline O-acetyl transferase. This enzyme, also absent from nondifferentiated plant material, has been cloned successfully (103).

As in morphine biosynthesis, the knowledge of the enzyme sequences allows a more detailed understanding of the localization of the enzymes (104). Strictosidine synthase (Fig. 2b) seems to be localized to the vacuole (105), and strictosidine glucosidase is believed to be associated with the membrane of the endoplasmic reticulum (73, 106). Tabersonine-16-hydroxylase is associated with the endoplasmic reticulum membrane (98); N-methyl transferase activity is believed to be associated with the thylakoid, a structure located within the chloroplast (101, 107); and vindoline-4-hydroxylase and desacetylvindoline O-acetyltransferase are believed to be localized to the cytosol (Fig. 2e) (107, 108). Overall, extensive subcellular trafficking of biosynthetic intermediates is required for vindoline biosynthesis. Aside from subcellular compartmentalization, specific cell types are required for the biosynthesis of some terpenoid alkaloids. Several enzymes involved in the early stages of secologanin biosynthesis seem to be localized to the phloem parenchyma, as evidenced by immunocytocchemistry and in situ RNA hybridization studies (109). However, additional studies have suggested that these genes also are observed in the epidermis and laticifers (110). Studies of the localization of vindoline biosynthetic enzymes by using immunocytocchemistry and in situ RNA hybridization strongly suggest that the mid-part of the vindoline pathway (tryptophan decarboxylase, strictosidine synthase, and tabersonine-16-hydroxylase) takes place in epidermal cells of leaves and stems. However, the later steps catalyzed by desacetylvindoline 4-hydroxylase and desacetylvindoline O-acetyltransferase take place in specialized cells, the laticifers, and idioblasts (109–112). As with isoquinoline alkaloid biosynthesis, deconvolution of the enzyme localization patterns remains a challenging endeavor.

Vinblastine

Vinblastine is a highly effective antitumor agent currently used clinically against leukemia, Hodgkin’s lymphoma, and other cancers. (113, 114). Vinblastine is derived from dimerization of vindoline and another terpenoid indole alkaloid, catharanthine.
Arabidopsis

Interestingly, expression of a transcription factor from Arabidopsis thaliana in C. roseus leaves has been demonstrated to convert vindoline and catharanthine to anhydrovinblastine in vitro (122, 123). Because the dimerization of these C. roseus alkaloids also can be catalyzed by peroxidase from horseradish in reasonable yields (124), it is interesting to speculate that anhydrovinblastine may be a by-product of isolation; after lysis of the plant material, nonspecific peroxidases are released from the vacuole and may act on vindoline and catharanthine.

Metabolic engineering of terpenoid indole alkaloids

Strictosidine synthase and tryptophan decarboxylase have been overexpressed in C. roseus cell cultures (125, 126). Generally, overexpression of tryptophan decarboxylase does not seem to have a significant impact on alkaloid production, although overexpression of strictosidine synthase does seem to improve alkaloid yields. Overexpression of tryptophan and secolagenin biosynthetic enzymes in C. roseus hairy root cultures resulted in modest increases in terpenoid indole alkaloid production (127, 128). Secologanin biosynthesis seems to be the rate-limiting factor in alkaloid production (129). Precursor-directed biosynthesis experiments with a variety of tryptamine analogs suggest that the biosynthetic pathway can be used to produce alkaloid derivatives (130). Strictosidine synthase and strictosidine glucosidase enzymes also have been expressed successfully heterologously in yeast (131); however, efforts to express heterologously terpenoid indole alkaloids currently are limited because the majority of the biosynthetic genes remain uncloned.

Transcription factors that upregulate strictosidine synthase (132), as well as a transcription factor that coordinate upregulates expression of several terpenoid indole alkaloid biosynthetic genes, have been found (133). Several zinc finger proteins that act as transcriptional repressors to tryptophan decarboxylase and strictosidine synthase also have been identified (134). Manipulation of these transcription factors may allow tight control of the regulation of terpenoid indole alkaloid production. Interestingly, expression of a transcription factor from Arabidopsis thaliana in C. roseus cell cultures results in an increase in alkaloid production (135).

Tropane Alkaloids

The tropane alkaloids hyoscyamine and scopolamine (Fig. 3a) function as acetylcholine receptor antagonists and are used clinically as parasympatholytics. The illegal drug cocaine also is a tropane alkaloid. The tropane alkaloids are biosynthesized primarily in plants of the family Solanaceae, which includes Hyoscyamus, Duboisia, Atropa, and Scopolia (136, 137). Nicotine, although perhaps not apparent immediately from its structure, is related biosynthetically to the tropane alkaloids.

Tropane alkaloid biosynthesis has been studied at the biochemical level, and several enzymes from the biosynthetic pathway have been isolated and cloned, although the pathway has not been elucidated completely at the genetic level (Fig. 3b). L-Arginine is converted to the nonproteogenic amino acid L-ornithine by the urease enzyme arginase. Ornithine decarboxylase then decarboxylates ornithine to yield the diamine putrescine. In Hyoscyamus, Duboisia, and Atropa, putrescine serves as the common precursor for the tropane alkaloids. Putrescine is N-methylated by a SAM-dependent methyltransferase that has been cloned to yield N-methylputrescine (139, 140). Putrescine N-methyltransferase now has been cloned from a variety of plant species (141–143), and site-directed mutagenesis and homology modeling models have led to insights into the structure function relationships of this enzyme (144). N-methylputrescine then is oxidized by a diamine oxidase to form 4-methylaminobutanal, which then spontaneously cyclizes to form the N-methyl-D-putrescine ion (144, 145). This enzyme, which recently has been cloned, seems to be a cooper-dependent amine oxidase (146, 147). Immunoprecipitation experiments suggest that this enzyme associates with the enzyme S-adenosylhomocysteine hydrolase (148). The pyrrolinium ion then is converted to the tropane skeleton by yet uncharacterized enzymes (Fig. 3b). Although no enzymatic information is available, chemical labeling studies have indicated that an acetate-derived moiety condenses with the pyrrolinium ion; one possible mechanism is shown in Fig. 3b (136). Tropamine then is reduced via an NADPH-dependent reductase to tropine that has been cloned from Hyoscyamus niger (149, 150). All tropane-producing plants seem to contain two tropine reductases, which create a branch point in the pathway. Tropine reductase I yields the tropane skeleton (Fig. 3b), whereas tropine reductase II yields the opposite stereocenter, pseudotropine (151). Tropamine is converted to scopalamine or hyoscymamine, whereas the TRI I product pseudotropine leads to callyclavine (152). These two tropine reductases have been crystallized, and site-directed mutagenesis studies indicate that the stereoselectivity of the enzymes can be switched (153, 154). The biosynthesis of scopalamine is the best characterized of the tropane alkaloids. After action by tropine reductase I, tropine is condensed with phenyllactate through the action of a P450 enzyme to form littorine (155). The phenyllactate moiety is believed to derive from an intermediate involved in phenylalanine metabolism (156). Littorine then undergoes rearrangement to form hyoscyamine. The enzyme that catalyzes this rearrangement, which has been purified partially, seems to proceed via a radical mechanism using S-adenosylmethionine as the source of an adenosyl radical (157). L-Labeling studies have been used to examine the mechanism of rearrangement (136, 157–159). Hyoscymamine is converted to the hydroxylation of hyoscymamine to 6β-hydroxyhyoscymamine as well as the epoxidation to scopolamine (Fig. 3b) (160, 161).
Alkaloid Biosynthesis

Figure 3. (a) Representative tropane and nicotine alkaloids. (b) Tropane biosynthesis. ODC, ornithine decarboxylase; PMT, putrescine N-methyltransferase; MPO, diamine oxidase; TR1, tropinone reductase 1; H6H, hyoscyamine 6b-hydroxylase.

H6H, which has been cloned and expressed heterologously (162), is a nonheme, iron-dependent, oxoglutarate-dependent protein. It seems that the epoxidation reaction occurs much more slowly than the hydroxylation reaction. The tropane alkaloids seem to be formed in the roots and then transported to the aerial parts of the plant (163).

Metabolic engineering of tropane alkaloids

Atropa belladonna plants have been transformed with an H6H clone from H. niger. A. belladonna normally produces high levels of hyoscyamine, the precursor for the more pharmaceutically valuable alkaloid scopolamine. However, after transformation with the H6H gene, transgenic A. belladonna plants were shown to accumulate scopolamine almost exclusively (164). Additionally, the levels of tropane alkaloid production in a variety of hairy root cultures were altered by overexpression of methyltransferase putrescine-N-methyltransferase and H6H. Overexpression of both of these enzymes in a hairy root cell culture resulted in significant increases in scopolamine production (164, 165). Fluorinated phenyllactic acid substrates could be incorporated into the pathway (166), and several substrates derived from putrescine analogs were turned over by the enzymes of several Solonaceae species (167).

Purine Alkaloids

Caffeine biosynthesis

Caffeine, a purine alkaloid, is one of the most widely known natural products. Caffeine is ingested as a natural component of coffee, tea, and cocoa, and the impact of caffeine on human health has been studied extensively. The biosynthetic pathway of caffeine has been elucidated recently on the genetic level. Caffeine biosynthesis has been studied most widely in the plant species Coffea (coffee) and Camellia (tea) (168, 169).

Xanthosine, which is derived from purine metabolites, is the first committed intermediate in caffeine biosynthesis (Fig. 4). Xanthosine can be formed from de novo purine biosynthesis, S-adenosylmethionine (SAM) cofactor, the adenylate pool, and the guanylate pool (169). De novo purine biosynthesis and the
The biosynthesis of caffeine begins with the methylation of xanthosine to yield N-methylxanthine by the enzyme xanthosine N-methyltransferase (XMT) (also called 7-methylxanthosine synthase) (171-173). N-methylxanthine is converted to N-methylxanthine by methylxanthine nucleosidase, an enzyme that has not been cloned yet (174). N-methylxanthine is converted to theobromine by 7-methylxanthine-N-methyltransferase (MXMT) (also called theobromine synthase), a second N-methyltransferase (171, 175). Theobromine is converted to caffeine by 7-methylxanthine-N-methyltransferase (2XMT) (also called caffeine synthase) (171).

Coffee and tea plants seem to contain a variety of N-methyltransferase enzymes that have varying substrate specificity (168, 169). For example, a caffeine synthase enzyme isolated from tea leaves catalyzes both the N-methylation of N-methylxanthine and theobromine (176). The substrate specificity of the methyltransferases can be changed by site-directed mutagenesis (177), and the crystal structure of two of the methyltransferases has been reported recently (178).

Metabolic engineering of caffeine biosynthesis

Caffeine may act as a natural insecticide in plants. When the three N-methyltransferase genes were overexpressed in tobacco, the resulting increase in caffeine production improved the tolerance of the plants to certain pests (179). Conversely, coffee beans with low caffeine levels would be valuable commercially, given the demand for decaffeinated coffee. Because of the discovery of these N-methyltransferase genes, genetically engineered coffee plants with reduced caffeine content now can be constructed (180, 181). For example, a 70% reduction in caffeine content in engineered coffee plants with reduced caffeine content now can be constructed (180, 181). For example, a 70% reduction in caffeine content in engineered coffee plants with reduced caffeine content now can be constructed (180, 181). For example, a 70% reduction in caffeine content in engineered coffee plants with reduced caffeine content now can be constructed (180, 181).

References


Further Reading


See Also

Chemical Diversity of Natural Products in Plants

Enzymatic Synthesis of Biomolecules

Metabolic Engineering

Metabolic Profiling

Natural Products: An Overview
Chlorophylls and Carotenoids, Chemistry of

Hugo Scheer,
Department Biologie I, Botanik, Universität München, München, Germany

doi: 10.1002/9780470048672.wecb083

Chlorophylls and carotenoids are essential pigments of photosynthesis, in which they have complementary functions for capturing light and transducing it to biochemical energy and for protecting against its deleterious effects. Carotenoids are also present in most other organisms where they have, besides photoprotection, a variety of other functions. Typical structures of both types of pigments are reviewed in this article as well as their functions, chemical properties, spectroscopy, biosynthesis, and applications.

Photons contain, by biologic standards, relatively high energies (120–300 kJ/mol). In photosynthesis, this energy is captured collectively by hundreds of pigment molecules (chlorophyll, carotenoids, and biliproteins) that are organized in light-harvesting complexes (LHCs). The excitation energy is then funneled with high quantum efficiencies to reaction centers (RCs), where few specialized chlorophylls initiate the charge separation across the photosynthetic membrane. The resulting electrochemical potential is used by a vectorial electron/proton transport chain to generate energy-rich compounds (adenosine triphosphate, reduced nicotinamide adenine dinucleotide) that are used eventually to reduce atmospheric carbon dioxide to carbohydrates (see the section entitled "Photosynthesis").

Chlorophylls (Chls) and carotenoids (Cars) are the two indispensable pigments of photosynthesis, where they have complementary functions. Chls are key pigments for the productive functions of photosynthesis, which includes light harvesting and primary energy transduction of light into an electrochemical gradient. The combination of intense light absorption and long excited state lifetimes renders them, however, highly phototoxic and potentially damaging. This finding may be the reason that only few examples exist where Chl (derivatives) have been found in nonphotosynthetic tissue and organisms and, generally, in low concentrations. Chls and their precursors are also involved in intracellular communication and regulation. Cars have essential light-protective functions in all photosynthetic organisms and in many nonphotosynthetic organisms. Cars are functionally more diverse, however. In photosynthetics, they can function in a productive way by light harvesting. They contribute to the coloration of flowers, their derivatives are photosensory pigments (rhodopsin), they are involved in the alternative photosynthesis of halobacteria that is based on \( \text{H}^+ \)-pumping, they act as hormones in plants (abscisic acid) and animal (retinoic acid), and they act as volatile defensive or attractive principles in plants. This short review is focused on the biochemistry and biophysics of Chls and Cars in photosynthesis, and it will provide a short outline of applications.

In the dynamic natural light environment, the combination of Chls and Cars allows photosynthetic organisms to maintain the balance between efficiently competing for light and damage by an overdose of light. Photosynthesis based on these two pigments has conquered nearly all habitats on earth where light is available, even at very low levels. It has produced the atmospheric oxygen we breathe, and it fixes \( 5 \times 10^{11} \) tons of carbon annually. The greening and de-greening of the vegetation in temperate zones, because of the biosynthesis of Chls in spring and their degradation in fall, is probably the most obvious life process on earth, which is visible clearly from outer space (1).

Chls and Cars both absorb light very strongly. In the LHC, these absorptions cover, in combination, most of the visible and the near-infrared spectral range (350–1050 nm). The two groups of pigments differ, however, by more than three orders of magnitudes in their excited state lifetimes: Those of the Chls live for several nanoseconds (\( 10^{-9} \) s), and those of most Cars live for only few picoseconds (\( 10^{-12} \) s). This finding has major consequences for their potential functions. The short lifetimes of the Cars suffice only for an efficient transfer in the LHC over very short distances, whereas the much longer lifetimes of Chls allow for transfer over tens of nanometers. Only the latter are also involved in charge separation. However, the short lifetimes of Cars with relatively excited state energies are favorable for light protection, where Cars are indispensable. Whenever, under varying light (cloud cover, leaf movements) and physiologic conditions (water, temperature), the energy cannot be transformed rapidly in a productive way; high-energy side-products can be formed that can lead to severe damage sunburn. Cars are involved in minimizing this damage at several levels. Excess energy is drained from the Chls and converted into heat, triplet states of Chls are quenched, and reactive oxygen species (ROS) are quenched as well, in particular singlet
Chlorophylls and Carotenoids, Chemistry of oxygen. The widespread occurrence of photosynthesis in many habitats and its evolution over more than $3 \times 10^9$ years led to considerable variations of the photosynthetic apparatus, which includes the pigments. The variety within the two pigment classes and the principally similar properties of the different pigments within each of the two groups will be reviewed.

Chlorophylls

Structures

The basic structure of Chls is the tetrapyrrolic porphyrin macrocycle with an attached isocyclic ring (Fig. 1). This structure has been termed "phytoporphyrin" by the IUPAC commission (2). The four nitrogens bind a magnesium ion ($\text{Mg}^{2+}$) as central metal. In rare cases, it is replaced by $\text{Zn}^{2+}$, and it is missing in the pheophytins that are present in type II reaction centers. Three spectrally distinct types of Chls can be distinguished by the degree of unsaturation of the macrocycle; they are shown in Fig. 1. It is fully unsaturated in the phytoporphyrin type Chls, hydrogenated in ring D in the phytochlorins, and hydrogenated in ring B in the phytobacteriochlorins. The red-most absorption band ($Q_x$) becomes increasingly intense and red-shifted with increasing saturation (Fig. 2). Green plants, green algae, cyanobacteria, and green photosynthetic bacteria contain mainly chlorin-type Chls. Bacteriochlorin-type Chls abound in the purple photosynthetic bacteria and heliobacteria, and the porphyrin-type Chls is found in brown algae and dinoflagellates (Table 1). The Chls vary even more in their peripheral substituents (Fig. 1). In most Chls, the C-17 propionic acid side chain is esterified by a long-chain alcohol, which is mainly the C20 terpenoid alcohol.
phytochlorin-type Chls that absorb at wavelengths ing energetics of water splitting, their reaction centers contain totrophs, viz. organisms that generate oxygen from water, can such as bacteria contain Chls at higher energies. Plants, red and green algae, and cyanobacteria Therefore, their light-harvesting systems also need to absorb bands in the blue spectral region where light in clear waters prevails. Plants, red and green algae, and cyanobacteria contain Chls a and b, few cyanobacteria contain Chl d. Brown algae and dinoflagellates, contain, in addition, Chls of the c-type. They absorb in the “green gap” where all other Chls absorb only weakly (Fig. 2) and they have particularly intense absorption bands in the blue spectral region where light in clear waters prevails.

RCs contain few, more specialized Chls (Fig. 1 and Table 1, see References 3 and 4). They form a chain of pigments across the membrane, over which in a stepwise fashion electrons are transferred from the primary donor, which is a pair of Chls located on the periplasmic side of the membrane, to the acceptors on the cytoplasmic side. Two types of reaction centers can be distinguished. In type I, the primary donor is a heterodimer composed of one molecule of Chl a or d (oxygentic photosynthesis) or BCHl g (an oxygentic Photosynthesis), and one molecule of the respective C-13P-epimer (Chl′ a, d, BCHl ′ g). The electron acceptors of the type I RC are Chls a (or d) in oxygentic photosynthesis and Chls like Bacteriochlorin a in an oxogentic Photosynthesis. In type II reaction centers, the primary donor is Chl a or d (oxygentic photosynthesis), or BCHl a or b (an oxygentic Photosynthesis): The same pigments also act as primary acceptors, whereas the Mg-free phaeophytin a and bacteriopheophytin a (or b), respectively, act as secondary acceptors.

Because of their long excited-state lifetimes, chlorophyll derivatives are rare outside phototropic organisms. However, some chlorophyll-like pigments perform other functions in nature: Certain deep-sea fish use Chls as visual pigments; in the marine worm, Bonellia viridis, a chlorophyll-like derivative acts as a sex determinant; and in certain tunicates, the tunichlorins may be involved in nonphotosynthetic electron transport (5).

Spectroscopy

The spectral properties of Chls are described by the four-orbital model (6–8). It predicts four major absorptions termed Q_x, Q_y, Q_b, and Q_a, in the near ultraviolet (NIR), visible (Vis), and NIR spectral regions that are often accompanied by vibrational side bands at higher energies of the 0–0 transition. The band intensities vary among the Chls, however, and they partly overlap so that the four bands are observed only in the bacteriochlorophyll structures, viz. BCHls a, b, g. The type of spectrum is determined mainly by the degree of unsaturation of the tetrapyrrole macrocycle (Fig. 2). In the fully unsaturated Chl c that contains a phytophorphyin macroyclic system, the B-bands around 400 nm are very intense and overlap, the Q_a-band around 620 nm is weak, and Q_b-band is even weaker and discernible only with special techniques. In phytochlorin-type Chls (Chl a, b, d, BCHl c, d, e), the B-bands are reduced in intensity and still overlap, the Q_a-band is increased to nearly equal intensity and red-shifted to ~660 nm, and the Q_b-band is only weak. In the phytobacteriochlorin-type Chls (BChl a, b, g), the B-bands are blue-shifted to ~400 nm and are well separated, the Q_a-band is even more increased in intensity and red-shifted to ~750 nm, and the Q_b-band has gained intensity and becomes clearly visible around 570 nm. By these spectral characteristics, the chlorophyll type of any given pigment can be determined readily from the spectra and other details can be detected, such as substitution pattern, ligation to the central metal (e.g., by the protein), or aggregation (see below).

Chls dissolved in organic solvents have long-lived excited states (10−5 s). Therefore, they are highly fluorescent and have significant intersystem crossing to the triplet state, which generates phosphorescence and, under aerobic conditions can generate highly toxic ROS like singlet oxygen. Chlorophyll solutions bleach rapidly by attack of the pigment by these ROS, and photooxidize cosolutes. Excess Chls or precursor porphyrins are phototoxic, their biosynthesis is controlled tightly, and deregu- lation is a way of herbicidal action (9–12). Chlporphyrin injected into animals lead to severe “sunburn” and destruction of tissue.

Figure 2: Type spectra of phytophorphyin, phytol, and phytobacteriochlorin-type Chls. See the internet version for color coding.
Chlorophylls and Carotenoids, Chemistry of

Table 1. Occurrence and functions of chlorophylls

<table>
<thead>
<tr>
<th>Organism</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
<th>Chlorophyta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorin</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Chlorophyll a</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
</tr>
<tr>
<td>Chlorophyll b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorophyll c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorophyll d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorophyll e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophyll a</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
<td>AR</td>
</tr>
<tr>
<td>Bacteriochlorophyll b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophyll c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophyll d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophyll e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriochlorophyll f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Ala, antenna pigment; R, reaction center pigment. Major pigments bold caps, minor pigments in lowercase, and pigments present in few species in italics.
Chlorophyll and Carotenoids, Chemistry of

This effect is used in photodynamic therapy of cancer and other diseases (PDT) (12, 13).

Chls aggregate readily both in nonpolar solvents and in water. Aggregates have generally red-shifted absorption of the \( Q_y \)-band, and the excited state lifetimes (and thereby fluorescence and phosphorescence) are reduced drastically because of rapid conversion of the excitation energy into heat by intermolecular conversion (IC). Formation of such aggregates is probably responsible for “concentration quenching.” In both cases, one mechanism of preventing quenching seems to be the presence of “traps,” such as radicals. In large aggregates in which the excitation is highly delocalized over many pigment molecules, this mechanism reads generally to deexcitation that becomes more efficient with aggregate size. Exceptions are the aggregates of BChls \( c, d \), and \( e \) present in chlorosomes of green and brown bacteria, which are fluorescent.

Aggregation is also observed in photosynthetic pigment-protein complexes, where it may be considered a major organizing force. In addition to the red-shifted absorption, an additional red shift exists of all bands and is brought about by the protein environment. Both bands are most pronounced in the phytobacteriochlorin-type BChls. In photosynthetic systems, these aggregates show much less IC. In the antenna systems, an important function of the protein is to block (and control) deexcitation channels of Chl-aggregates in a manner still unresolved mechanistically. Isolated LHCs, which contain most (>99%) Chls of the photosynthetic apparatus therefore show high fluorescence. Light-harvesting Chls are coupled to those in the RC, such that now the excitation energy of the former is transferred efficiently to the latter. The RCs show only little fluorescence at moderate light intensities, here “photochemical quenching” by the uppermost indicator antenna system (BChls). To study the absorption bands of the RC, one needs to demetallate the BChls. To the best of our knowledge, the complete RC of green algae has not yet been isolated. It is not clear whether the RCs are a direct continuation of the antenna systems, or whether there is an additional reaction center. It is clear from a comparison of the BChls of different organisms that the arrangement of the pigments in the chlorosomes of purple bacteria is significantly different from the arrangement in the LHC of green bacteria, the chlorosomes, coordinating properties of the central Mg\(^{2+}\), and the ligation of special peripheral substituents combine to form large aggregates of BChls \( c, d \), and \( e \) that are nearly devoid of protein. Free valences of the central Mg\(^{2+}\) are also critical for interactions with the proteins in photosynthetic complexes. By coordinating to suitable amino acid side chains (e.g., histidine, methionine, and glutamate) or backbone C=O-groups, they are positioned optimally for efficient energy or electron transfer. Most (B)Chls have a single extra ligand; in this case, two isomers can be formed in which binding occurs to the (inequivalent) upper or lower face of the molecule (Fig. 1).

The esterifying alcohol, in most cases phyto (Fig. 1), comprises about 1/3 of the mass of Chls, yet its influence on the chemistry (and function) is still poorly understood. With the exception of the nonesterified Chls \( c, d \), it renders Chls amphiphilic, and is important both in aggregation in polar environments, and in the positioning of Chls in photosynthetic complexes. Variations of the alcohol are frequent in phototrophic bacteria, in particular in the chlorosomal membrane, where they contribute to formation of the fluorescent BChl \( c, d \), and \( e \) aggregates that are unique for these pigments (16).

**Metabolism**

The first dedicated intermediate in CN biosynthesis (17) is 5-aminolevulinic acid (ALA), which is the common precursor of all tetrapyrroles. It can be formed either in a single step from succinyl-CoA and glycine (C\(_5\)-pathway) or from glutamic acid (C\(_4\)-pathway) via an intermediate (Glu-IRNA\(^{\text{7p}}\)) that is generally involved in protein synthesis. Some photosynthetic bacteria (purple bacteria) use the C\(_4\)-pathway for BCHN formation; most other organisms use the C\(_5\)-pathway. In the second stage, which is ubiquitous for all natural tetrapyrroles, four ALA react to yield a pyrrole (porphobilinogen). Four of those then condense to a linear tetrapyrrole and then cyclize in a remarkable reaction in which one of the pyrroles is flipped, to the cyclic tetrapyrrole skeleton (uroporphyrinogen III). The next step involves a series of decarboxylations and oxidations to yield protoporphyrin. The latter is the first of a series of increasingly colored and phototoxic products, therefore, the organisms regulate the levels of ALA tightly to avoid any over production of protoporphyrin. Any deprotonation of this process, as well as mutations of the enzymes, may result in severe damage and often death (9–11).
The next stage, which is dedicated to Chls, is the insertion of Mg, the esterification at the C-13 propionic acid side chain, and its cyclization to yield protochlorophyllide. The biosynthesis of the phytoporphyrin-type Chls c is practically complete at this stage and requires only peripheral modifications. For the phycobobilin-type Chls, ring D is reduced either by the light-dependent protochlorophyllide-reductase (DPOR), or by the light-dependent LPOR. In angiosperms, only the latter is present; therefore, they require light for Chl biosynthesis. Subsequent reactions at the periphery, including esterification with an activated long-chain alcohol derived from the isopenoid pathway, complete the reaction sequence to the phycobobilin-type Chls. In the phycobilliprotein-type c BChls, ring B is reduced, too, by an enzyme that is homologous to DPOR, which is followed again by peripheral modifications.

Because of the high phototoxicity of Chls, their degradation is tightly controlled. It has been studied in detail only for Chls a and b, where it involves early on the (light-independent) opening of the macrocycle to yield the much less phototoxic open chain tetrapyrroles (bilins), which are modified over several stages and eventually are degraded to monocyclic compounds (14, 15). Very little is known on the degradation of the other Chls.

Applications

The best source for Chl a is the cyanobacterium Spirulina platensis, which is available commercially. Chl a/b mixtures can be obtained from all green plants. All other Chls are less readily accessible, which limits their applications. Several Chl derivatives are used as dyes for food colorants (Cu-chlorophyllin) and cosmetics. The “chlorophyll” used for the “chlorophyll” ink is a complex mixture of degradation products. More recently, (B)Chl derivatives have gained increasing interest as photosensitizers in photodynamic therapy of cancer, these compounds include pigments in which the isocyclic ring is opened and/or the central metal has been removed or replaced (e.g., by Pd(II)) to increase phototoxicity (12, 13).

Carotenoids

Structures

Most Cars are tetraterpenes that contain 6–15 conjugated double bonds. Two C20–units (originally geranyl-geraniol) are joined tail-to-tail to a chain of 32 carbon atoms bearing 8 methyl side-chains (see Reference 18 for nomenclature, and lycopene tail-to-tail to a chain of 32 carbon atoms bearing 8 methyl a similar fashion but starting from two C15-units (farnesol). In addition, a family of triterpenic Cars exists that are generated in acylation, and shortening or extension of the carbon skeleton. In oxygen-containing functional groups and their glycosylation or ble bonds, cyclization at one or both ends, the introduction of (18), including isomerization and rearrangement of the double bonds, cyclization at one or both ends, the introduction of oxygen-containing functional groups and their glycosylation or acylation, and shortening or extension of the carbon skeleton. In addition, a family of triterpenic Cars exists that are generated in a similar fashion but starting from two C30–units (farnesol). In combination, these modifications account for the more than 800 isomers. The number and the structural variations of Cars reflect a variety of functions. The modifications seem to be particularly far-reaching in Cars dedicated to light harvesting, as exemplified by peridinin; a highly modified C35 pigment from algae (Fig. 3).

Furthermore, many carotenoid metabolites exist that have distinct functions. One such example shown in Fig. 3 is retinal, the chromophore of visual pigments (rhodopsins) and the light-driven proton pump, bacteriorhodopsin. Other examples are the plant hormone, abscisic acid, or volatile compounds that contribute to the fragrance of roses, for example.

Occurrence and Functions

Car functions are as diverse as their structures. In contrast to the Chls, they are not defined to photosynthetic organisms but are rather ubiquitous in living organisms. However, animals (including humans) cannot synthesize them but rely on dietary supplies, for example of vitamin A.

The essential and indispensable function of Cars in photosynthesis is the protection of the photosynthetic apparatus. Variations of light quality (color) and intensity in the natural environment often lead to overload of the photosynthetic machinery. Whenever energy transfer in the light-harvesting systems is disturbed, or the RC cannot cope with the input from the LHC, the Cars are the prime protectants from toxic long-lived excited states of the Chls and the subsequently formed ROS (see above). This task is achieved by three major reaction mechanisms. The first is direct photoprotection: Cars have several “forbidden” states that can accept excess singlet excitations from Chls; this energy is subsequently converted rapidly into heat by IC (19). The second mechanism is quenching of Chl triplets by triplet energy transfer to the low-lying Car triplets that can no longer generate ROS. Even if ROS are formed by Chl photosensitization, Cars can detoxify them for example by energy transfer from singlet oxygen or by addition of ROS to the double bond system. All these processes require very close distances between the donor (excited Chls, ROS) and the energy-accepting Car. Therefore, all photosynthetic Chl proteins contain Cars, and they are always in contact with the Chls, as is revealed by several X-ray structures of photosynthetic complexes (20). Specialized Cars are found in the RC and in all Chl-containing LHCP. In the latter, the effective Cars in protection seem to be positioned strategically at critical sites where the energy is funneled to the RC.

Because ROS attack is an important defense mechanism against infections in animals, Cars have functions in bacteria that protect them against the immune system. Many bacteria synthesize Cars for their protection, and many virulent forms of Cars are colored by deep Colors.

Carotenoids also protect simply by their function as nonphototoxic light filters. They remove, by virtue of their high absorption, near ultraviolet and blue light region, and they degrade their excitation energy rapidly to heat by IC. Certain algae tolerant to extreme light stress contain droplets of pure Cars. Many nonphotosynthetic organisms synthesize Cars when subjected to increasing light intensities. As nonphototoxic pigments, Cars often function as “safe” colorants in nature. Examples are colors of flowers used for attracting and for communicating with
C. Chemistry of Chlorophylls and Carotenoids

**Precursors**
- Dimethylallyl-pyrophosphate
- Farnesyl-pyrophosphate
- Geranyl-pyrophosphate
- Geranylgeranyl-pyrophosphate
- Phytoene

**Phototrophic bacteria**
- Okonone
- 13-cis-Rhodopinal
- Spiroloxanthin
- Spheroidenone

**Algae**
- Peridinin
- Fucoxanthin
- Diadinoxanthin
- Diatoxanthin
- Prasinoxanthin

![Figure 3](https://example.com/figure3.png)

Figure 3: Selected carotenoid structures from bacteria, algae, plants, and animals, and of precursors and metabolic products with biologic function. The IUPAC numbering is given for lycopene (top right).
Chlorophylls and Carotenoids, Chemistry of

**Higher Plants**

- Lycopene
- α,β-Carotene
- Zeaxanthin
- Antheraxanthin
- Violaaxanthin
- 9-cis-Neoxanthin
- Abscisic acid

**Animals**

- Astaxanthin
- Diglucosyl C_{30}-carotenoid
- Halocynthiaxanthin
- Tunaxanthin
- 4-Oxomytiloxanthin
- Salmoxanthin
- Retinal (all-trans)

Figure 3 (continued)

...pollinating animals, mainly insects, or examples of crustaceans that blend into the marine background.
A rather function in photosynthetic organisms is light harvest-
ing in the “green gap” (470–600 nm) where Chls absorb only poorly. Because of rapid IC, most Cars transfer energy only with low efficiency to Chls. However, particularly in microal-
geae and phototrophic bacteria they can contribute prominently to photosynthesis despite their general protective function. Two factors are important to this function. One is the evolution of Cars in which the excited state lifetime is somewhat increased, and therefore IC is reduced. The two most abundant Cars, fucox-
anthin and peridinin, have lifetimes that reach 100 ps (22). The second factor is, again, a location of such Cars close enough to Chls that energy transfer becomes sufficiently effective within the short excited state lifetimes (electron exchange mechanism). Plants and some algae use, e.g., in the so-called violaxanthin cy-
cle) subtle structural modifications to manipulate Cars such that light-harvesting Cars are converted into protecting pigments, and vice versa, in response to the light supply and the status of the photosynthetic apparatus (22). The mechanism of this switch is still unclear.

Last are Car precursors for important metabolites. Only three examples shall be given. The first example is retinal (Fig. 3), which is the chromophore of the visual pigment rhodopsin (23) and is derived from β-carotene. Because the latter cannot be synthesized by mammals, they need it to be supplied as provitamin A. Retinal derivatives are also required for other regulatory functions. The second example is abscisic acid (Fig. 3), which is the plant hormone involved in the shedding of leaves in fall and in fruit ripening: it is derived from violaxanthin. Finally, certain fragrances of roses are not synthesized directly, but they are breakdown products of the flowers’ Cars.

Spectroscopy

Carotenoids are conjugated linear polyenes with 6–15 double bonds. The optical spectra of such pigments are characterized by some unusual features (Fig. 4, 19). First, the lowest energy S0 – S1 transition located in the red to NIR spectral range is “opti-
cally forbidden” for conventional one-photon excitation or emis-
sion; therefore, the related absorption and fluorescence are ex-
tremely weak. Theoretically, this is true only for C2-symmetric polyenes; but in practice, even highly asymmetric Cars like peridinin have negligible S0 absorption. Second, the most in-
tense absorption, which is responsible for the yellow-orange color of most Cars, is a series of closely spaced, sometimes overlapping bands in the 400–550 nm range. They belong to vibrational sub-bands of an S0 – S0 absorption that are gen-
erally well resolved, but they can be broadened to a degree that they appear only as a single, broad band. With an increas-
ing number of conjugated double bonds, these absorptions are shifted, in an asymptotic fashion, to the red. Third, at least two other transitions of intermediate energy exist, which are again forbidden and therefore are very weak. Fourth, cis-carotenoids show an additional band that is typically located 100–150 nm below to the blue of the main absorption. Last, the triplet states have comparably low energy. Cars with >7 double bonds it lies below 1250 nm, which is the energy required to generate sin-
glet oxygen. The “forbidden” bands and the triplet state(s) are fundamen-
tally important for the biologic functions of Cars (see abo-

above). Car triplets can quench singlet oxygen efficiently by a spin-allowed singlet-triplet exchange reaction. The “forbidden” states are important in singlet energy transfer with other Cars and, in particular, Chls. They can be reached only indirectly (e.g., after absorption into the energy-rich major absorption band and subsequent IC) or by energy transfer from neighboring pig-
ments like Chls, but thereby contribute to energy transfer and dissipation. The fluorescence of Cars is generally negligible even from the optically allowed S0 state because of rapid IC to the lower lying forbidden states. Generally, lifetimes are a few picoseconds, and even in extreme cases they reach only 100 ps, which corresponds to fluorescence yields of ~1%.

When incorporated into proteins, the spectral properties of Cars can be modified considerably. Generally, the bands are sharpened and red-shifted, and they can become strongly op-
tically highly even for achiral Cars because of twisting of the long-chain molecule. The red shifts are in the range of ~10 nm, but they can be much larger. Spectacular cases are the color change of astaxanthin from orange to green when it is bound in the crustacean protein, α-astaxanthin, and its reversion when the protein is denatured by boiling, or the red shift of retinals in rhodopsins.

Chemical Properties

The basic hydrocarbon skeleton of Cars is very hydrophobic, and almost all Cars are insoluble in water and soluble in nonpo-
lar solvents. Solubility is increased by polar functional groups and conjugation to sugars (glycosylation) (Fig. 3), but truly water-soluble Cars are very rare. The conjugated double-bond system is moderately stable chemically. It is subject to rear-
rangements, in particular in the light, and to additions (e.g., of oxygen). The chemical properties of the highly modified Cars are determined by their particular functional groups, which in some Cars are labile. A frequent example is the epoxy-group present e.g., in violaxanthin, diadinoxanthin, fucoxanthin or peridinin. With the latter two representing >75% of the total Cars. Because of the diversity of structures and substituents, the reader is referred to specialized treatments of carotenoid chemistry (see “Further Reading” section).
Metabolism

Carotenoids are terpenoids, which are derived from oligomerization of activated isoprene. It can be synthesized either by the mevalonate pathway from acetyl-Coenzyme A, or via the recently discovered deoxyxylulose pathway (24). The latter seems to be the pathway that leads to Cars in most photosynthetic organisms. Both pathways result in two isomers: isopentenyl pyrophosphate and dimethylallyl pyrophosphate. One molecule of the latter is condensed by prenyl transferases in a head-to-tail fashion sequentially with three molecules of isopentenyl pyrophosphate. The resulting geranylgeranyl pyrophosphate (GGPP) constitutes an important branching point in terpenoid metabolism that leads to Chls via esterification with the tetrapyrrolic chlorophyllides (see above), to longer isoprenoids via additional condensation with isopentenyl pyrophosphate, and in particular to Cars (25), beginning with a tail-to-tail condensation of two molecules of GGPP by the first dedicated enzyme, phytoene synthase. The resulting phytoene is still uncolored because it has only three conjugated double bonds (Fig. 3). Dehydrogenation by phytoene-desaturase still uncolored because it has only three conjugated double bonds (Fig. 3). Dehydrogenation by phytoene-desaturase via phytolpheine to lycopene is common to most known Cars. C30 molecules are derived, in an analogous fashion, from tail-to-tail condensation of two C15 units (farnesol-pyrophosphate) and subsequent dehydrogenations. A variety of enzymes is responsible for additional structural modifications, which includes cyclases to form the end-rings characteristic of many Cars, oxygenases to introduce OH groups which can be additionally modified, isomerases and (de)hydrogenases to modify the double-bond system, and many others, which provide the structural diversity of Cars.

Applications

The use of Cars as food colorants is of considerable economic importance. This task can be done directly, but is generally done indirectly by supplying Cars or carotenoid-rich algae as food additives to fish (salmon) or poultry (chicken eggs). In cosmetics, Cars are used as sunscreens. Provitamin A (β-carotene) is a dietary supplement that is given directly but more frequently is given indirectly, for example as milk supplement. Natural sources for Cars are algae like Hematococcus, plants like carrots, or genetically manipulated bacteria. Industrial scale synthetic methods have been developed.

References


Further Reading


See Also
Cancer, Topics in Chemical Biology
Photosynthesis, Electron Transfer Chemistry in Porphyrin and Corrin Biosynthesis
Spectroscopic Techniques: Overview of Applications in Chemical Biology
Terpenoid Biosynthesis
Vision, Topics in Chemical Biology
Whereas plants and certain microorganisms can generate all required coenzymes from CO₂ or simple organic precursors, animals must obtain precursors (designated as vitamins) for a major fraction of their coenzymes from nutritional sources. Still, most vitamins must be converted into the actual coenzymes by reactions catalyzed by animal enzymes. The structures and biosynthetic pathways of some coenzymes are characterized by extraordinary complexity. Enzymes for coenzyme biosynthesis have frequently low catalytic rates, and some of them catalyze reactions with highly unusual mechanisms.

Many coenzymes (cofactors) involved in human and animal metabolism were discovered in the first half of the twentieth century, and their isolation and structure elucidation were hailed as milestones as shown by the impressive number of Nobel prizes awarded for research in that area. Studies on coenzyme biosynthesis were typically initiated in the second half of the twentieth century and have generated a massive body of literature that continues to grow rapidly because the area still involves many incompletely resolved problems. In parallel, numerous novel coenzymes were discovered relatively recently by studies of microorganisms. In this article, the terms “cofactor” and “coenzyme” are used as synonyms.

Cofactor biosynthesis is a very broad and multifaceted topic. This article summarizes basic concepts of major cofactors. Detailed status reports on the biosynthesis of individual coenzymes can be found in excellent recent reviews that are quoted (1).
Vitamin B12 from nutritional sources must then be converted to coenzyme B12 by mammalian enzymes. Ultimately, the biosynthesis of vitamin B12 requires five enzymes, albeit of vital importance, which are involved in fatty acid and amino acid metabolism (6). Notably, because plants do not generate corrinoids, animals depend on bacteria for their supply of vitamin B12 (which may be obtained in recycled form via nutrients such as milk and meat) (7).

Some coenzymes develop by very complex pathways and/or reactions. Whereas tetrahydrobiopterin is biosynthesized from GTP via just three enzyme-catalyzed steps (2), some coenzyme biosynthetic pathways are characterized by enormous complexity. Thus, the biosynthesis of vitamin B12 requires a sequence of a deamination and reduction can proceed in a common theme, for example, in the biosynthesis of vitamin B12, a sequence of a deamination and reduction can proceed in different order. In other cases, entirely different reactions afford a given intermediate, for example in the case of nicotinic acid biosynthesis. A further example is the biosynthesis of isoprenoid building blocks and their downstream products, in which the existence of a second pathway besides the classic mevalonate pathway had been ignored until recently. Notably, higher plants use both isoprenoid pathways but for different final products.

### Specific Biosynthetic Pathways

**Biosynthesis of iron/sulfur clusters**

Iron/sulfur clusters are inorganic cofactors that are used in all cells (10). They comprise S\(^{2-}\)  ions and iron ions in the +2 or +3 state (Fig. 2). Iron/sulfur clusters are essential cofactors for numerous redox and nonredox enzymes, alone or in tandem with organic cofactors such as flavocoenzymes and/or pyridine nucleotides. The simplest structural type is the rhombic [2Fe-2S] cluster (3), [3Fe-4S] (4) and [4Fe-4S] (5) clusters are characterized by distorted cubic symmetry (10, 11). Clusters can form aggregates, and other metal ions can replace iron ions or can be present additionally.

Whereas many cognate apoenzymes can be reconstituted with iron/sulfur clusters by simple and essentially alchemistic procedures using Fe\(^{2+}\) and sulfide ions under anaerobic conditions, a highly complex enzymatic machinery is used in vivo for the synthesis of iron/sulfur clusters and their transfer to the target enzymes. Sulfide ions required for cluster synthesis are obtained from cysteine (1) via a persulfide of a protein-bound cysteine residue (2); pyridoxal phosphate is required for the formation of the persulfide intermediate (Fig. 1) (12).

In eukaryotes, the formation of iron/sulfur clusters proceeds inside mitochondria (13). The mitochondrial enzymes are orthologs of the subcellular isc proteins and are characterized
by very slow rates of evolution. Iron/sulfur clusters are initially assembled on Isc protein (prokaryotic) or Isu protein (eukaryotic) that serves as a scaffold. They can be exported to the cytoplasm in which they can become part of cytoplasmic enzymes by the assistance of proteins that serve as iron chaperones.

The persulfide intermediate 2 can also serve as a sulfur source for the biosynthesis of thiamine (6), lipoic acid (7), molybdopterin (8), and biotin (9) (Fig. 1).

Tetrapyrroles

A large and structurally complex family of coenzymes, including various hemes, chlorophylls, corrinoids (including coenzyme B12) and the archeal coenzyme F430, is characterized by their macrocyclic tetrapyrrole structure (14, 15). These coenzymes contain a metal ion (Fe, Mg, Co, or Ni) at the center of the tetrapyrrole macrocycle, which is specifically introduced by enzyme catalysis. These compounds are all derived from δ-aminolevulinic acid (12) that can be biosynthesized by two independent pathways, one that is, from glycine (11) and succinyl-CoA (10) in animals and some bacteria (e.g., Rhodobacter) or from glutamyl-tRNA (13) in plants, many eubacteria, and archaebacteria (Fig. 2) (16, 17). Two molecules of δ-aminolevulinic acid (12) are condensed under formation of porphobilinogen (14). Oligomerization of porphobilinogen affords hydroxymethylbilane (15), in which all pyrrole rings share the same orientation of their substituents. Ring D is then inverted by a rearrangement that affords uroporphyrinogen III (16) (3). Side-chain modification and the incorporation of iron by ferrochelatase (18) afford the various hemes, chlorophylls, corrinoids (including coenzyme B12) and the archaeal coenzyme F430 by a series of enzyme reactions that occur in the human host (7).

A family of coenzymes biosynthesized from GTP: tetrahydrofolate, tetrahydrobiopterin, flavocoenzymes, molybdopterin

Several coenzymes comprising a pyrimidine ring motif are derived from GTP (22) (Fig. 3). Specifically, this group comprises two members of the B vitamin group, riboflavin (vitamin B2) (24) and folic acid/tetrahydrofolate (33). Two other members of the group, tetrahydrobiopterin (32) and molybdopterin (8), are biosynthesized de novo in animals and do not have vitamin status (20, 21).

The first committed step in the biosyntheses of these compounds is the hydrolytic opening of the imidazole ring of GTP, which affords a diaminopyrimidine-type intermediate. In the biosynthetic pathways of folate, tetrahydrobiopterin, and methanopterin (34), the respective diaminopyrimidine intermediate undergoes ring closure by means of an intramolecular condensation that involves parts of the ribose side chains of GTP, which affords a 2-amino-4-pteridinone compound (29).

The 4-aminobenzoate moiety of tetrahydrofolic acid is obtained from the shikimate ring pathway of aromatic amino acid biosynthesis via chorismate. Interestingly, apicomplexan protozoa may have conserved the complex shikimate pathway for the single purpose to generate 4-aminobenzoate as a tetrahydrofolic acid precursor, whereas aromatic amino acids are obtained from external sources.

The formation of intermediate 27 (compound 2) in the biosynthesis of molybdopterin (8) proceeds a rearrangement that...
involves the ribose side chain as well as C-8 of GTP for the formation of the tetracyclic ring system.

The carbocyclic moiety of vitamin B2 is assembled from two molecules of a deoxytetrolose phosphate (22). The carbocyclic moiety of the deazaflavin-type coenzymes (36) is provided by the tyrosine precursor, 4-hydroxyphenylpyruvate.

Notably, the GTP cyclohydrolases that catalyze the first committed step in the pathways of tetrahydrofolate biosynthesis in plants and microorganisms and of tetrahydrobiopterin in animals are orthologs.

**Thiamine pyrophosphate**

The biosynthesis of thiamine pyrophosphate (46) in microorganisms and plants is characterized by extraordinary complexity (21, 23). Animals are dependent on nutritional sources but can convert unphosphorylated thiamine (that is not an intermediate of the bacterial biosynthetic pathway) into thiamine pyrophosphate in two steps.

In bacteria, the pyrimidine precursor (38) is derived from 5-aminoimidazole ribotide (39), an intermediate of the basic branch of the bacterial biosynthetic pathway into thiamine pyrophosphate in two steps.

In bacteria, the pyrimidine precursor (38) is derived from 5-aminoimidazole ribotide (39), an intermediate of the basic branch of the bacterial biosynthetic pathway into thiamine pyrophosphate in two steps.

In yeasts, a totally unrelated reaction sequence uses carbon atoms from vitamin B6 (39) that are indicated by roman letters in Fig. 4 for the assembly of the thiamine precursor (38), whereas the nitrogen atoms and one additional carbon atom are introduced from histidine (40).

In bacteria, the thiazole moiety (42) of thiamine is derived from 1-deoxy-D-xylulose 5-phosphate (43) that can also serve as a precursor for pyridoxal in many eubacteria (Fig. 5) and for isoprenoids via the nonmevalonate pathway (cf. isoprenoid cofactors). The sulfur atom is derived from the persulfide that also serves as precursor for iron-sulfur clusters and for biotin (6) and thiooctanoate (7). C2 and N3 of the thiazole moiety of thiamine have been reported to stem from tyrosine in Escherichia coli and from glycine in Bacillus subtilis, respectively. Yeasts use ADP-ribulose (44) derived from NAD as precursor (24).

In plants, little is known about the basic building blocks and the reactions involved in thiamine biosynthesis. An early study with chloroplasts of spinach indicated that 1-deoxy-D-xylulose 5-phosphate, tyrosine, and cysteine act as precursors for the thiazole moiety in analogy to the pathway in E. coli. More recently, it has been shown that a homolog of the THIC protein that converts 5-aminoimidazole ribotide into (38) is essential (25). These results suggest that the plant pathway is similar to the pathway in prokaryotes but not to that in yeast.

**Pyridoxal phosphate**

In many eubacteria, 1-deoxy-D-xylulose phosphate (43) serves as a common precursor for the biosynthesis of vitamins B6 and...
Figure 3  Coenzymes biosynthesized from GTP. 8, molybdopterin; 22, GTP; 23, 5-amino-6-ribitylamino-2,4(1H,3H)-pyrimidinedione; 24, riboflavin; 25, FMN; 26, 5,6-dimethylbenzimidazole; 27, precursor Z; 28, metal containing pterin; 29, dihydroneopterin triphosphate; 30, 6-pyruvoyl-tetrahydropterin; 31, 6,7,8-dideazafolate; 32, dihydroneopterin; 33, 6(R)-5,6,7,8-tetrahydrobiopterin; 34, dihydroneopterin; 35, 6(S)-5,6,7,8-tetrahydrofolate; 36, 5,6,7,8-tetrahydromethanopterin; 37, 5-deaza-7,8-dideazafolate; 38, coenzyme F420.

B3 and for the biosynthesis of isoprenoids via the nonmevalonate pathway. Condensation of 43 with 3-amino-1-hydroxyacetone phosphate (47) (biosynthesized from D-erythrose 4-phosphate) affords pyridoxine 5′-phosphate (48, Fig. 5A). A sequence of elimination, tautomerization, and water addition precedes cyclization via an aldol condensation (26, 27). Then, pyridoxine 5′-phosphate can be converted into pyridoxal 5′-phosphate (39) by oxidation with molecular oxygen.

A more recently discovered second pathway starts from ribulose 5-phosphate (49) that is condensed with ammonia and glyceraldehyde phosphate or its isomerization product dihydroxyacetone phosphate (50), which affords pyridoxal 5′-phosphate in a single enzyme-catalyzed reaction step (Fig. 5B) (28, 29). This pathway seems to be widely distributed: it is used in plants (30) and has also been shown to proceed in fungi, archaea, and most eubacteria.

In mammals, dietary vitamin forms, including pyridoxal, pyridoxol, and pyridoxamine, can all be converted to the respective coenzyme forms by phosphorylation.

Pyridine nucleotides

Animals and yeasts can synthesize nicotinamide from tryptophan via hydroxyanthranilic acid (52) and quinolinic acid (53, Fig. 6A) (31), but the biosynthetic capacity of humans is limited. On a diet that is low in tryptophan, the combined contributions of endogenous synthesis and nutritional supply of precursors, such as nicotinic acid, nicotinamide, and nicotinamide riboside, may be insufficient, which results in cutaneous manifestation of niacin deficiency under the clinical picture of pellagra. Exogenous supply of nicotinamide riboside was shown to promote NAD+−dependent Sir2 function and to extend life-span in yeast without calorie restriction (32).

Bacteria and plants use aspartate (54) and dihydroxyacetone phosphate (50) as precursors for the biosynthesis of nicotinamide via quinolinic acid (53, Fig. 6B) (33). The transformation of precursors into NAD (56) and NADP (57) follow the same pathway in all organisms. A ribosyl phosphate residue can be transferred to biosynthetic quinolinic acid.
Cofactor Biosynthesis

Figure 4  Biosynthesis of thiamine (vitamin B1). 37, aminoimidazole ribotide; 38, 2-methyl-4-amino-5-hydroxymethyl-pyrimidine phosphate; 39, pyridoxal 5′-phosphate; 40, histidine; 41, 2-methyl-4-amino-5-hydroxymethyl-pyrimidine pyrophosphate; 42, 4-methyl-5-hydroxyethylthiazole phosphate; 43, 1-deoxy-D-xylulose 5-phosphate; 44, 5-ADP-D-ribose; 45, thiamine phosphate; 46, thiamine pyrophosphate.

Figure 5  Formation of the pyridoxine ring in vitamin B6. (A) deoxyxylulose phosphate-dependent pathway; (B) deoxyxylulose phosphate-independent pathway; 43, 1-deoxy-D-xylulose 5-phosphate; 47, 3-amino-1-hydroxyacetone 1-phosphate; 48, pyridoxine 5′-phosphate; 49, ribulose 5-phosphate; 50, dihydroxyacetone phosphate; 39, pyridoxal 5′-phosphate.

or to preformed nicotinamide affording 55 (or its amide) that is converted to NAD by adenylation. Subsequent phosphorylation yields NADP.

Pantothenate

Pantothenate (also designated vitamin B5, 64, Fig. 7) is biosynthesized de novo in plants and many microorganisms but must be obtained from nutritional sources by animals (20, 34). The branched carboxylic acid 63 is obtained from α-ketoisovalerate (62), which is an intermediate of valine biosynthesis, via 62. β-Alanine (60) is obtained by decarboxylation of aspartate (54) in microorganisms. Plants and yeasts can biosynthesize β-alanine from spermine (58) (35). An additional pathway to 60 starting from uracil (59) has been reported in plants, where the downstream steps that lead to pantothenate are understood incompletely.

Organisms of all biological kingdoms convert 64 into the cysteamine derivative phosphopantetheine (65) using L-cysteine as substrate. 65 is converted to coenzyme A (66) by attachment of an adenosine moiety via a pyrophosphate linker and phosphorylation of the ribose moiety. Phosphopantetheine can be attached covalently to serine residues of acyl carrier proteins that are parts of fatty acid synthases and polyketide synthases.

Vitamin C

Whereas most mammals can synthesize ascorbic acid (vitamin C, 75) from D-glucose 1-phosphate (67) via the pathway shown in Fig. 8B, humans and guinea pigs lack the last enzyme of that
Cofactor Biosynthesis

Figure 6  Biosynthesis of pyridine nucleotides. A) in animals and yeasts; B) in plants and bacteria.

50, dihydroxyacetone phosphate; 51, tryptophan; 52, hydroxyanthranilic acid; 53, quinolinic acid; 54, aspartate; 55, nicotinic acid mononucleotide; 56, NAD; 57, NADP.

pathway and are therefore dependent on nutritional sources (36, 37). Plants use the pathway shown in Fig. 8A that has been elucidated relatively recently (38, 39). Yeasts produce and use a five-carbon analog, which is called erythroascorbic acid, instead of ascorbate. The biosynthetic pathway of erythroascorbic acid involves the oxidation of D-arabinose to D-arabino-1,4-lactone, which is then oxidized to erythroascorbic acid.

Pyroloquinoline quinone

Pyroloquinoline quinone (77, Fig. 9, PQQ) serves as cofactor of bacterial oxidoreductases (40). The heated debate whether PQQ has vitamin character for animals is still inconclusive (41-43). PQQ is derived from a peptide precursor that contains conserved glutamate (13) and tyrosine (76) residues. All carbon and nitrogen atoms of the precursor amino acids are incorporated into the product (44). Gene clusters involved in this pathway have been studied in considerable detail. The X-ray structure of the enzyme that catalyzes the final reaction step has been determined, and reaction mechanisms have been proposed on that basis (45). However, details of the biosynthetic pathway are still incompletely understood.

Biotin

Biotin (vitamin H, 6, Figs. 1 and 10) acts as cofactor of carboxylases. It can be produced in bacteria, plants, and some fungi (46). The biosynthetic pathway involves four steps that start from alanine (78) and pimeloyl-CoA (79). Carboxylation and cyclization of 81 affords diethiobiotin (82), which is then converted into biotin (6) by the iron/sulfur protein, biotin synthase, in an unusual radical mechanism (47).

Isoprenoid cofactors

Isoprenoids are one of the largest classes of natural products that comprise at least 35,000 reported members (48). Many of these compounds play crucial roles in human metabolism as hormones, vitamins (vitamins A, D, E and K), quinine-type cofactors of respiratory chain enzymes (ubiquinone), membrane constituents, and functionally important side chains of signal cascade proteins (Fig. 11). Chlorophyll (17, Fig. 2) and heme a (18) have isoprenoid side chains. Whereas vitamin E (99) and vitamin A or its biosynthetic precursor, β-carotene, must be obtained by animals from dietary sources, many other isoprenoids, including the quinone type coenzyme Q family (where individual representatives differ by the length of their side chains), can be synthesized de novo by vertebrates. 3-Hydroxy-3-methylglutaryl-CoA reductase, the enzyme that catalyzes the conversion of (S)-3-hydroxy-3-methylglutaryl-CoA (84) to mevalonate (85), is one of the most important drug targets for the prevention of cardiovascular disease (see below) (49, 50).

All isoprenoids are biosynthesized from two isomeric 5-carbon compounds, isopentenyl diphosphate (IPP, 86) and dimethylallyl diphosphate (DMAPP, 87) (Fig. 11). The mammalian pathway for the biosynthesis of these key biosynthetic precursors from three acetyl-CoA units (83) via mevalonate (85) had been elucidated in the 1950s (51). In the wake of that pioneering work, it became established dogma that all terpenoids are invariably of mevalonate origin, even in the face of significant aberrant findings.
The existence of a second pathway that affords IPP and DMAPP was discovered in the 1990s. The details of that nonmevalonate pathway were then established in rapid sequence by the combination of isotope studies, comparative genomics, and enzymology (52). The delayed discovery of the nonmevalonate pathway can serve as paradigm for pitfalls in the elucidation of biosynthetic pathways.

The mevalonate pathway starts with a sequence of two Claisen condensations that afford (S)-3-hydroxy-3-methylglutaryl-CoA (84) from three acetyl-CoA moieties. The pathway affords IPP that can be converted into DMAPP by isomerization. The first committed intermediate of the nonmevalonate pathway is 2C-methyl-D-erythritol 4-phosphate (90) obtained from 1-deoxy-D-xylulose 5-phosphate (43), which is a compound also involved in the biosynthesis of vitamins B1 (46, cf. Fig. 4) and B2 (39, cf. Fig. 3), by rearrangement and subsequent reduction. Three enzyme-catalyzed steps are required to convert the compound into the cognate cyclic diphosphate 91 that is then converted reductively into a mixture of IPP and DMAPP by the consecutive action of two iron/sulfur proteins.

It is now firmly established that green plants use both isoprenoid biosynthesis pathways (52). More specifically, sterols and triterpenes are generated in the cytoplasm via the mevalonate pathway, whereas isopentenyl diphosphate and isopentenyl pyrophosphates are generated in plastids via the nonmevalonate pathway. These assignments are not absolute because there is a level of crosstalk between the compartments.

Oligomerization of isoprenoids under elimination of pyrophosphate affords the precursors for the biosynthesis of monoterpenes, sesquiterpenes, diterpenes, triterpenes, and tetraterpenes (93). Long-chain oligomer pyrophosphates also supply the side chains of vitamin E (99, α-tocopherol, Fig. 11), heme a (138), chlorophyll (17, Fig. 2), and the quinone type co-factors, including vitamin K (menaquinone, 98) and coenzyme Q10 (ubiquinone, 97). The quinone moieties are derived from hydroxybenzoate that is synthesized from tyrosine in animals or from chorismate in microorganisms (53, 54).

The tetraterpene, β-carotene (95), is biosynthesized in microorganisms and in the chloroplasts of higher plants where it serves as an important component of the light-harvesting apparatus. In plants, the isoprenoid precursor units of carotenoids are predominantly of nonmevalonate origin. In vertebrates, β-carotene serves as provitamin A that can be converted into the vitamin retinol (96) by oxidative cleavage. Whereas vitamin A functions as a component of retinal photoreceptors and in signal transmission in vertebrates, it is also involved in bacterial photosynthesis, in which it serves as prosthetic group in a light-driven proton pump. The halobacterial proton pump and
Cofactor Biosynthesis

Figure 8: Biosynthesis of ascorbic acid (vitamin C). (A) in plants; (B) in mammals. 67, d-glucose 1-phosphate; 68, d-glucurionate; 69, l-gulonate; 70, l-gulono-1,4-lactone; 71, d-glucose-6-phosphate; 72, d-mannose 1-phosphate; 73, l-galactose; 74, l-galactono-1,4-lactone; 75, ascorbic acid. The enzyme converting 70 into 75 is missing in primates and guinea pigs.

Figure 9: Biosynthesis of PQQ. 76, glutamate (protein bound); 77, tyrosine (protein bound); 78, PQQ.

Figure 10: Biosynthesis of biotin. 78, alanine; 79, pimeloyl-CoA; 80, 7-keto-8-amino-pelargonic acid; 81, 7,8-diamino-pelargonic acid; 82, dethiobiotin; 83, biotin.

Humans can generate vitamin D₃ (cholecalciferol) by endogenous biosynthesis but require dietary sources under certain environmental conditions. More specifically, 94 that can be obtained via the endogenous mevalonate pathway can be photochemically converted into provitamin D in light-exposed skin areas. The transformation requires ultraviolet light. At higher geographical latitudes, light exposure of the skin can be a limiting factor. The fact that skin pigmentation retards the photochemical formation of vitamin D may have acted as a selective factor that favored pale skin when modern humans migrated to areas with higher geographic latitude.

Biosynthesis of archaeal coenzymes

Archaea bacteria that were discovered only during recent decades are now recognized as a third kingdom of life besides eubacteria and eukaryotes. An important subgroup of archaea can generate energy by conversion of carbon dioxide or low molecular weight organic acids into methane. The pathway of methanogenesis has been shown to implicate several unique coenzymes (55). Specifically, 5-deaza-8-hydroxy-10-ribosylisoalloxazine (factor
clusters (12). Use a pyrosulfide protein precursor that is also used for the coenzymes (thiamine (18) or pyridoxal phosphate). A tetrapyrrole-type coenzyme F430 (5) was named on basis of its absorption maximum at 430 nm. The nickel-chelating factor is biosynthesized via the porphyrin biosynthetic pathway (Fig. 2) (19). For the handling of one-carbon fragments that play a central role in their metabolism, methanogenic bacteria use methanopterin (34, Fig. 3). The tetrahydropterine system that serves as the business end of the methanopterin coenzyme family is structurally similar to tetrahydrofolate, and the biosynthetic pathway starting from GTP is similar to that of tetrahydrofolate (Fig. 3). The ribitylaniline moiety is derived from ribose and from the shikimate pathway via 4-amino-phenylpyruvate (56, 57). In contrast to the coenzymes limited to methanogenic bacteria and are also found in streptococci, those of the riboflavin biosynthetic pathway, which affords the benzenoid moiety, are derived from the shikimate pathway via 4-hydroxyphenylpyruvate (56, 57).

Some coenzymes serve as biosynthetic precursors that afford structural parts of other coenzymes. Thus, the benzenoid moiety of the flavo coenzyme FMN serves as a precursor for the lower ligand 26 of the central cobalt ion in vitamin B12 (20, Fig. 3). Pyridoxal and NAD are used as precursors for the biosynthesis of thiamine in yeast (Fig. 4) (23, 24).

Branching of pathways is relevant in several cases. Thus, intermediates of the porphyrin biosynthetic pathway serve as precursors for chlorophyll (12, Fig. 2) and for the central ring systems of vitamin B12 (20, Fig. 2) (17). 3-D-xylyl-5-phosphate (43) serves as an intermediate for the biosynthesis of pyridoxal phosphate (17, Fig. 3), for the terpenoid precursor IPP (86) via the nonmevalonate pathway (Fig. 11), and for the thialeole moiety of thiamine pyrophosphate (46, Fig. 4). 7,8-Dihydrobiopterin triphosphate (29, Fig. 3) serves as an intermediate in the biosynthetic pathways of tetrahydrofolate (33) and tetrahydrobiopterin (31). The closely related compound 7,8-dihydroneopterin triphosphate is the precursor of the archaean cofactor, tetrahydromethanopterin (34) (58). A common pyrimidine-type intermediate (23) serves as the precursor for flavin and deazaflavin coenzymes. Various sulfur-containing coenzymes (lipoic acid (7), biotin (6), Fig. 1) use a pyrosulfide protein precursor that is also used for the biosynthesis of inorganic sulfide as a precursor for iron/sulfur clusters (12).

**Coenzymes as biosynthetic precursors**

Several coenzymes are involved in the biosynthesis of their own precursors. Thus, thiamine is the cofactor of the enzyme that converts 1-deoxy-D-xylulose 5-phosphate (43) to tetrahydrofolate derivatives as coenzymes (cofactors) (Fig. 3). When a given coenzyme is involved in its own biosynthesis, we are faced with a “hen and egg” problem, namely how the biosynthesis could have evolved in the absence of the crucially required final product. The answers to this question must remain speculative. The final product may have been formed via an alternative biosynthetic pathway that has been abandoned in later phases of evolution or that may persist in certain organisms but remains to be discovered. Alternatively, the coenzyme under study may have been accessible by a prebiotic sequence of spontaneous reactions. An interesting example in this respect is the biosynthesis of flavin coenzymes, in which several reaction steps can proceed without enzyme catalysis despite their mechanistic complexity.

In terms of coenzyme evolution, it is also noteworthy that the biosynthesis of a given coenzyme frequently requires the cooperation of other coenzymes. For example, the biosynthesis of riboflavin (24) requires tetrahydrofolate (33) for the biosynthesis of GTP, which serves as a precursor (Fig. 3). Pyridoxal 5′-phosphate is required for the biosynthesis of the activated pyrosulfide type protein (2) that serves as the common precursor for iron/sulfur clusters and various sulfur-containing organic coenzymes (Fig. 1).

Covalent coenzymes

Whereas many coenzymes form noncovalent complexes with their respective apoenzymes, various flavoenzymes are characterized by covalently bound FMN (25) or FAD (Fig. 3). Covalent linkage involves the position 8 of the riboflavin or the benzenoid carbon atom of the flavin and a cysteine or histidine residue of the protein. The covalent CN or CS bond can be formed by autoxidation of the noncovalent apoenzyme/coenzyme precursor complex as shown in detail for nicotine oxidase (59). Biotin (6, Fig. 10) and lipoic acid (7, Fig. 1) are attached enzymatically to apoenzymes via carboxamide linkage to specific lysine residues (60, 61). The pantotheryl moiety (64, Fig. 7) can also be linked covalently to proteins via amide linkage (62). Covalently bound heme is involved in heme M (63) and heme L-catalyzed reactions (64, 65). Several covalently bound coenzymes, including pyruvyl, methylidene imidazolone, topaquinone, and tryptophan tryptophyl quinone-type prosthetic groups are generated by posttranslational modification (66).
Figure 11. Biosynthesis of isoprenoid type cofactors. 18, Homoe; 19, pyridoxal 5’-phosphate; 43, 1-deoxy-D-xylulose 5-phosphate; 46, thiamine pyrophosphate; 81, acetyl-CoA; 84, (S)-3-hydroxy-3-methylglutaryl-CoA; 95, 

β-carotene; 96, retinol; 97, ubiquinone; 98, menaquinone; 99, α-tocopherol.

Cellular topology of coenzyme biosynthesis in eukaryotes

In bacteria, coenzyme biosynthesis is located in the cytoplasm. In eukaryotic cells, organelles play important roles in the biosynthesis of certain cofactors. For example, certain steps of iron-sulfur cluster biosynthesis proceed in mitochondria (10, 13). In plants, some steps of the biosynthesis of tetrahydrofolate, biotin, and lipoate proceed in mitochondria (7, 15), whereas the biosynthesis of vitamin B2 is operative in plastids (67). In apicomplexan protozoa, enzymes in mitochondria, in the apicoplast (an organelle that is believed to have a common evolutionary origin with chloroplasts) and in the cytoplasmic
that results from mutations of phenylalanine hydroxylase has also been advocated for certain patients with classic phenylketonuria. In fact, the relatively large number of patients with classic phenylketonuria may provide an economic incentive for the development of a biotechnological process for the bulk production of the coenzyme.

Deficiencies of enzymes involved in the transformation of the vitamin pantothenic acid (64) into the cognate coenzyme forms (66, Fig. 7) result in severe developmental and neurological deficits that affect few human patients (34, 72). Therapy with megadoses of pantothenic acid has been advocated, but their efficiency has yet to be demonstrated by stringent clinical studies (73).

Genetic defects of molybdopterin biosynthesis (Fig. 3) also result in severe neurologic and developmental deficits (74). Genetic defects in the biosynthesis of the quinine-type coenzyme Q10 (97, Fig. 11) can result in encephalopathy, myopathy, and renal disease (53).

Inherited defects or porphyric biosynthetic enzymes can cause the accumulation of pathway intermediates that cannot be converted anymore with sufficient velocity. Various genetic forms of porphyria have been reported and result in liver toxicity, neurological damage and photosensitivity (75). Aquired forms of porphyria can be caused by a variety of toxic and pharmacologic agents.

The absorption of vitamin B12 (20) requires a glycoprotein-designated intrinsic factor that is secreted by the gastric mucosa. A acquired failure to produce the intrinsic factor results in a complex disease that can present with hematological (macrocytotic anemia and pernicious anemia), neurological, or psychiatric symptoms or a combination thereof. Prior to the discovery of vitamin B12, pernicious anemia was lethal. Initial treatment was based on the consumption of large amounts of uncooked liver. If diagnosed timely, the disease can now be cured easily by the parenteral administration of vitamin B12. Notably, the liver can store large amounts of the vitamin.

### Cofactor Biosynthesis

#### Three-dimensional structures of coenzyme biosynthesis enzymes

The rapid technological progress in X-ray crystallography has enabled the structural analysis of numerous enzymes involved in coenzyme biosynthesis. Complete sets of structures that cover all enzymes of a given pathway are available in certain cases such as riboflavin, tetrahydrobiopterin, and folic acid biosynthesis. Structures of orthologs from different taxonomic groups have been reported in certain cases. X-ray structures of enzymes in complex with substrates, products, and analogs of substrates, products, or intermediates have been essential for the elucidation of the reaction mechanisms. Structures of some coenzyme biosynthesis enzymes have been obtained by NMR-structure analysis.

Enzymes that are addressed by major drugs have been studied in particular detail. Thus, well above one hundred structures have been reported for dihydrofolate reductases from a variety of organisms, including major pathogens such as Mycobacterium tuberculosis, which is the causative agent of tuberculosis, and of Plasmodium falciparum, which is the most important of the Plasmodium spp. that causes malaria. The interaction of mammalian dihydrofolate reductases with inhibitors that are used as cytostatic agents and/or immunosuppressants is also documented extensively by X-ray structures.

The rapidly growing number of three-dimensional coenzyme biosynthesis enzyme structures in the public domain and the cognate publications are best addressed via the internet server of Mycobacterium tuberculosis (76). The factor binds the vitamin and enables its subsequent transport across the ileal mucosa. Acquired failure to produce the intrinsic factor results in a complex disease that can present with hematological (macrocystotic anemia and pernicious anemia), neurological, or psychiatric symptoms or a combination thereof. Prior to the discovery of vitamin B12, pernicious anemia was lethal. Initial treatment was based on the consumption of large amounts of uncooked liver. If diagnosed timely, the disease can now be cured easily by the parenteral administration of vitamin B12. Notably, the liver can store large amounts of the vitamin.

#### Medical Aspects

**Genetic deficiency of coenzyme biosynthesis**

Genetic defects have been reported for the biosynthesis of several coenzymes in humans. Typically, these rare anomalies cause severe neurological deficits that become apparent at birth or in early childhood.

Specifically, the deficiency of certain enzymes of tetrahydrobiopterin biosynthesis (GTP cyclohydrolase I, pyruvoyltetrahydrobiopterin synthase, [Fig. 3](#)) result in severe neurological and developmental deficits designated as atypical phenylketonuria. The condition can be treated with some success by the oral application of synthetic tetrahydrobiopterin in large amounts. More recently, tetrahydrobiopterin therapy has also been advocated for certain patients with classic phenylketonuria that results from mutations of phenylalanine hydroxylase (71). This therapeutic approach is based on the concept that the function of certain defective phenylalanine hydroxylases can be bolstered by increased amounts of the cognate coenzyme tetrahydrobiopterin. In fact, the relatively large number of patients with classic phenylketonuria may provide an economic incentive for the development of a biotechnological process for the bulk production of the coenzyme.

Genetic defects of molybdopterin biosynthesis (Fig. 3) also result in severe neurologic and developmental deficits (74). Genetic defects in the biosynthesis of the quinine-type coenzyme Q10 (97, Fig. 11) can result in encephalopathy, myopathy, and renal disease (53).

Inherited defects or porphyric biosynthetic enzymes can cause the accumulation of pathway intermediates that cannot be converted anymore with sufficient velocity. Various genetic forms of porphyria have been reported and result in liver toxicity, neurological damage and photosensitivity (75). Aquired forms of porphyria can be caused by a variety of toxic and pharmacologic agents.

The absorption of vitamin B12 (20) requires a glycoprotein-designated intrinsic factor that is secreted by the gastric mucosa. A acquired failure to produce the intrinsic factor results in a complex disease that can present with hematological (macrocystotic anemia and pernicious anemia), neurological, or psychiatric symptoms or a combination thereof. Prior to the discovery of vitamin B12, pernicious anemia was lethal. Initial treatment was based on the consumption of large amounts of uncooked liver. If diagnosed timely, the disease can now be cured easily by the parenteral administration of vitamin B12. Notably, the liver can store large amounts of the vitamin.

#### Coenzyme biosynthesis enzymes as anti-infective drug targets

Enzymes involved in coenzyme biosynthesis represent targets for anti-infective agents (77). The sulfonamides that were discovered in the 1930s were the first group of synthetic agents for anti-infective agents (77). The sulfonamides that were discovered in the 1930s were the first group of synthetic agents with a broad spectrum of activity against pathogenic bacteria and protozoa. Their mode of action, via inhibition of dihydrofolate synthetase in the biosynthetic pathway of tetrahydrofolate biosynthesis (Fig. 3), was elucidated only much later. Subsequent studies on compounds with antifolate activity afforded inhibitors of dihydrofolate reductase, which is the en-zyne in that pathway that catalyzes the formation of tetrahydrofolate from dihydrofolate in organisms that synthesize the coenzyme de novo and from folate in organisms that rely on dietary sources. Trimethoprim, which is an inhibitor of dihy-drofolate reductase that is required for the use of nutritional folate and dihydrofolate as well as for the metabolic recycling
of tetrahydrofolate coenzymes) became a widely used antimicrobial agent that is typically applied in combination with a sulfonamide.

Fosmidomycin, initially discovered as a product of Streptomyces lavendulae with antibacterial and herbicide activity, was recently shown to act via the inhibition of IspC protein that catalyzes the first committed step in the nonmevalonate pathway of isoprenoid biosynthesis that is absent in humans (43 → 90, Fig. 11). Based on these findings, the compound is now under clinical evaluation as an antimalarial drug (78).

In principle, other coenzyme biosynthetic pathways that occur in pathogenic bacteria but not in humans should qualify as anti-infective drug targets with a favorable toxicity profile. Novel anti-infective principles would be highly desirable in light of the rapid spread of resistant pathogens.

Coenzyme biosynthesis as a target for cytostatic agents

The development of the sulfonamides as antibacterial and antiparasitic agents had preceded the discovery of its metabolic target in the biosynthesis of tetrahydrofolate (that was per se unknown in the 1930s) (Fig. 3). The discovery of the vitamin in the 1940s triggered a wave of research directed at additional inhibitors of its biosynthesis. This work resulted in the discovery of methotrexate that is widely used as a cytostatic agent predominantly for hematological malignancy, and also an immunosuppressive agent used in the therapy of autoimmune disease such as Crohn’s disease (79).

Drug interaction with coenzyme biosynthesis pathways

Mevastatin is an inhibitor of 3-hydroxy-3-methylglutaryl-CoA reductase (Fig. 11, 84 → 85) that was isolated from Penicillium citrinum in 1971. A group of structurally related compounds designated as statins is now widely used for the prevention and treatment of cardiovascular disease (80). Statins exert their desired effects via the lowering of low-density lipoprotein and probably also via reduced prenylation of small G-proteins of the Ras protein family that are involved in proinflammatory signaling. The highly pleiotrophic mevalonate pathway is the source of numerous other highly important metabolites including coenzyme Q10 (97). However, a recent review concludes that the suppression of coenzyme Q10 biosynthesis, which can be expected as a side result of statin therapy, is not a significant cause of rhabdomyolysis, a dreaded side effect of statin therapy (81).

Modulation of vitamin D biosynthesis by environmental factors

Fair-skinned humans face a dilemma because ultraviolet light exposure carries the risks of carcinogenesis and skin aging, whereas insufficient ultraviolet light carries the risk of vitamin D deficiency (94, Fig. 13). Insufficiency. The case is unique in so far as an endogenous biosynthetic pathway is subject to regulation by external lifestyle factors such as ultraviolet exposure and the use of chemical and physical sunscreens. Recent studies indicate that vitamin D deficiency is widespread in the human population and is a risk factor for a wide variety of conditions, including cancer and autoimmune disease (82). The dilemma of ultraviolet protection and vitamin D sufficiency can clearly be addressed easily by vitamin D supplements.

Biotechnology

Harnessing biosynthetic pathways for vitamin production

Certain vitamins that serve as coenzymes (vitamin C) or as precursors of coenzymes (all B group vitamins and certain carotenoids) are commercially produced in bulk amounts, which are used for human nutrition and animal husbandry, as antioxidants (vitamins C and E) and food colorants (vitamin B12, carotenoids). Only a fraction of technically manufactured vitamins is used for inclusion in drugs.

Vitamin B12 is produced exclusively by bacterial fermentation technology. The chemical synthesis of vitamin B12 has been suspended during the past two decades by fermentation processes using bacteria and yeasts. Carotene, various other carotenoids and vitamin A are produced by chemical synthesis, but a variety of biotechnological processes have been also been explored for their production. Notably, vitamin A deficiency continues to be a major cause for acquired blindness in developmental countries, although vitamin A and β-carotene, which serves as a provitamin, can be produced at modest cost and in virtually unlimited quantity by existing technology. At least certain steps of the various technical vitamins production are also conducted by fermentation. For various other vitamins, biotechnological production may become competitive in the future.

References

Cofactor Biosynthesis


Cofactor Biosynthesis


Lantibiotics, Biosynthesis and Mode of Action of

L.A. Furgerson-Ihnken and Wilfred A. van der Donk,
University of Illinois at Urbana-Champaign, Illinois
doi: 10.1002/9780470048672.wecb277

Lantibiotics are a unique class of antimicrobial peptides produced by Gram-positive bacteria. They are synthesized ribosomally as precursor peptides and undergo extensive posttranslational modification to attain their biologically active structures. These modifications include dehydration of serine and threonine residues that produce dehydroalanines and dehydrobutyrines, followed by intramolecular addition of cysteine thiols to the unsaturated amino acids. The lantibiotic biosynthetic machinery displays low substrate specificity both in vivo and in vitro, which allows for the engineering of novel lantibiotics. The polycyclic peptides exert their biological activity through several modes of action including the sequestration of the cell wall precursor lipid II and pore formation in bacterial membranes. Because of their high potency, several commercial applications of lantibiotics are being developed.

Biological Background

The prototypical lantibiotic, nisin, was discovered in 1928 for its antibacterial properties and has been used as a preservative in dairy products since the 1950s (1). Nisin and other lantibiotics exhibit nanomolar efficacy against many Gram-positive strains of bacteria (2), which include methicillin resistant Staphylococcus aureus, vancomycin resistant enterococci, and oxacillin resistant bacteria. On the other hand, some lantibiotics function as morphogenetic peptides rather than antibiotics and are important for spore formation in streptomycetes (3). Since the structural elucidation of nisin in the early 1970s, extensive research efforts have been directed at understanding the biosynthesis and mode of action of various lantibiotics.

Molecular structures

Lantibiotics are synthesized ribosomally as precursor peptides with an amino-terminal leader sequence and a carboxy-terminal structural region. The structural region undergoes extensive posttranslational modification catalyzed by enzymes unique to lantibiotic biosynthesis. After structural region modification, the unaltered leader sequence is removed by a protease, which results in the biologically active species.

Lantibiotics contain several unusual amino acids, including the thioether lanthionine (Lan) linkage and its methyl substituted analog methyl-lanthionine (MeLan) (Fig. 1a) that unifies all members of the class and accounts for their family name. In addition to Lan, lantibiotics commonly contain 2,3-dehydroalanine (Dha) and (Z)-2,3-dehydrobutyryne (Dhb). In all, no less than 15 different posttranslational modifications have been documented in lantibiotics (for a selection see Fig. 1a), and up to 58% of their amino acids are modified. These extensive structural alterations overcome the constraints imposed by the use of 20 amino acids in ribosomally synthesized peptides. Some less common, posttranslationally crafted residues in lantibiotics are β-hydroxy aspartate, lysinoalanine, aminovinyl cysteine (AviCys), D-alanine, 2-oxobutyrate, 2-oxopropionate, and 2-hydroxypropionate. The presence of these unusual residues is thought to be important for the biological activity of lantibiotics.

Currently, the lantibiotic family contains more than 50 members with varying structures, properties, and biological activities (Fig. 1b). Depending on the ring topology of mature lantibiotics, they may have an elongated three-dimensional structure.
Lantibiotics, Biosynthesis and Mode of Action of

Figure 1. (a) Selection of the structural motifs found in lantibiotics that result from posttranslational modification. A shorthand notation for each structure is also illustrated. (b) Representative structures of mature lantibiotics using the shorthand notation.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
(e.g., nisin), a globular conformation (e.g., cinnamycin), or have both regions of linearity and of globular nature (e.g., lacticin 481). The thioether linkages are thought to confer stability and enhance activity by locking the mature lantibiotics in their biochemically active form and by decreasing the susceptibility to protease-mediated degradation. The mature species vary widely in size, shape, charge, and biological activity. They have been classified by groupings named after a representative member (e.g., the nisin, lacticin 481, and cinnamycin subgroups) (4). Members of each group share a high degree of sequence homology, biosynthetic machinery, and also have similar biological activities. A unique group of lantibiotics is the two-component systems such as lacticin 3147 (Fig. 2a) that are composed of two posttranslationally modified peptides that act in synergy to exert their bactericidal activity.

**Mode of action**

Nisin displays several biological activities including inhibition of cell wall biosynthesis, disruption of the cell membrane, and inhibition of spore outgrowth. Nisin is able to disrupt cell wall biosynthesis by binding to the essential peptidoglycan precursor lipid II (Fig. 2a) (2). Based on studies with fluorescently labeled nisin, the molecule not only binds lipid II, but also removes it from its functional location in the cell (7). Mutagenesis studies have demonstrated that the A and B rings of nisin are essential for binding to lipid II (5), and a nuclear magnetic resonance (NMR) structure of a 1:1 nisin-lipid II analog complex has revealed that the mechanism of molecular recognition involves hydrogen bonding interactions between the lipid II pyrophosphate moiety and the nisin amide linkages in the A and B-rings (Fig. 2b) (6). After binding to lipid II at the membrane surface, the C-terminus of nisin inserts perpendicularly into the membrane (Fig. 2c). Mutagenesis studies have also shown that the C and D rings of nisin as well as a hinge region are important for pore formation (8). This hinge region, which is composed of three residues that link the C and D rings (A200, M201, and Ly22), gives nisin the conformational flexibility necessary to traverse the lipid bilayer. The mechanism of bacterial spore outgrowth inhibition is less defined, but it is thought to involve the electrophilic nature of Dha5 (9).

The mode of action of the two-component lantibiotic lacticin 3147 also involves binding to lipid II and pore formation, but it requires two posttranslationally modified peptides to act in synergy to affect their bactericidal properties. In a recent model, lacticin 3147 A2 was proposed to first bind lipid II, which causes a change in its conformation such that lacticin 3147 A2 is recruited to the cell surface. Interaction of A2 then promotes deeper insertion into the membrane and pore formation (10). Other lantibiotics, like mersacidin, also bind lipid II but do not form pores. They are thought to prevent the transglycosylation step of cell wall biosynthesis by preventing action of the transglycosylation enzyme on lipid II. This inhibition is achieved by binding to the disaccharide pyrophosphate region of the peptidoglycan precursor (11). Vancomycin also prevents cell wall biosynthesis by binding to the D-Ala-D-Ala moiety of lipid II. Considering their different binding sites, it is not surprising that alteration of the D-Ala-D-Ala structure in vancomycin-resistant bacterial strains does not affect the efficacy of mersacidin (11). Thus, mersacidin represents a promising antibiotic candidate with the potential of treating multi-drug resistant bacterial infections.

**Lantibiotic Biosynthesis**

The genes that encode lantibiotics and their biosynthetic machinery are designated generically by the locus symbol lan, with a more specific annotation for each family member. The lanA gene encodes the precursor peptide LanA, the lanE gene for the nisin precursor, LctA for the lacticin 481 prepeptide, and so on. The lanBCDMPTXY genes encode for various modification enzymes, the lanKROX genes produce regulatory proteins, and the lanEG1 genes provide immunity proteins. Typically, a subset of these genes is present in a given biosynthetic gene cluster. The cluster for nisin is shown in Fig. 3a as a representative example.

**Enzymatic dehydration and cyclization**

The unsubstituted residues Dha and Dib are formed by dehydration of serine and threonine residues, respectively, and the thioether linkages Lys and MeLys are generated by intramolecular Michael-type addition of cysteine thiol(s) to the unsubstituted sites (e.g., Fig. 3b). These modifications can be performed by either two separate enzymes (LanB and LanC) in class I lantibiotics or a single bifunctional enzyme (LanM) in class II lantibiotics. Typically, proteolysis of the leader sequence is performed by a dedicated protease, either a LanP serine protease (class I) or the cysteine protease domain of a LanT protein (class II). The lanB genes encode large (~1000 residues) predominantly hydrophilic dehydratases that may be membrane associated. To date, the dehydratase activity of a LanB protein has not been reconstituted in vitro and little is known about the mechanism of catalysis of this group of enzymes.

The lanC genes encode smaller (~400 residues), zinc metalloenzymes that catalyze cysteine thiol addition to unsaturated amino acids to form the lanthionine and methyllanthionine residues in a regiospecific, stereoselective, and chemoselective manner (Fig. 3b). The first experimental evidence for the role of LanC proteins in lantibiotic biosynthesis was the accumulation of dehydrated NisA in strains that lack the NisC protein (14). In 2006, the crystal structure of NisC was solved, and its cyclase activity was reconstituted in vitro (15). Its active site contains a single zinc atom with one histidine and two cysteine...
Figure 2. Mode of action of the prototypical lantibiotic nisin. (a) The peptidoglycan precursor lipid II is composed of an N-acetylmuramic acid disaccharide (GlcNAc-MurNAc) that is attached to a membrane anchor of 11 isoprene units via a pyrophosphate moiety. A pentapeptide is linked to the muramic acid. Transglycosylase and transpeptidase enzymes polymerize multiple lipid II molecules and crosslink their pentapeptide groups, respectively, to generate the peptidoglycan. (b) The NMR solution structure of the 1:1 complex of nisin and a lipid II derivative in DMSO (6). (c) The amino-terminus of nisin binds the pyrophosphate of lipid II, whereas the carboxy-terminus inserts into the bacterial membrane. Four lipid II and eight nisin molecules compose a stable pore, although the arrangement of the molecules within each pore is unknown (5).

ligands as well as a coordinated water molecule. A cysteine thiol of the substrate is thought to displace this water molecule, followed by deprotonation to generate a thiolate that then attacks the beta-carbon of a Dha or Dhb residue, which results in an enolate that is prototated stereoselectively to provide the D-configuration at the alpha-carbon (Fig. 3c).

Class II lantibiotic gene clusters (e.g., lacticin 481) do not contain lanBC genes, but instead contain a lanM gene that encodes a large bifunctional LanM protein (∼115–120 kDa) that catalyzes both dehydration and cyclization (Fig. 4). The C-termini of LanM proteins share a low degree of sequence identity with LanC proteins, which include the zinc binding ligands, but have no homology to LanB proteins. Both the dehydratase and cyclase activities of lacticin 481 synthetase (LctM) and haloduracin synthetase (HalM) were reconstituted recently in vitro (16, 17).

Other modifications

Another common posttranslational modification found in mature lantibiotics is a C-terminal AviCys (Fig. 1a). This residue is formed by EpiD in epidermin (18). Currently, not much is known about the enzymes that catalyze formation of other unique lantibiotic residues. The cinnamycin gene cluster was cloned recently (19), which set the stage for studies to determine which gene products are responsible for lysinoalanine and beta-hydroxy aspartate formation.

Proteolysis and export

After posttranslational modification of the precursor peptide, the unmodified leader sequence is cleaved by a protease to reveal the biologically active compound. Depending on the lantibiotic, this cleavage occurs either just prior to or after transport outside the cell. For most class I lantibiotics, the cleavage reaction is...
Lantibiotics, Biosynthesis and Mode of Action of

Figure 3 (a) The nisin biosynthetic gene cluster. (b) Posttranslational modifications during the biosynthesis of nisin. Dehydration of serine and threonine residues in the structural region of the precursor peptide NisA is performed by the dehydratase NisB. Then, the (Me)LaN rings are installed by the cyclase NisC. After secretion, the unmodified leader sequence is removed by the serine protease NisP, which generates the biologically active species. (c) The proposed cyclization mechanism for NisC.

catalyzed by a LanP serine-type protease. Many LanP proteases contain a presequence, which indicates they themselves may be secreted and therefore act after the modified precursor peptide has been exported. NisP is required for nisin maturation because strains that lack the nisP gene but contain all other nisin biosynthetic machinery produce fully modified but biologically inactive nisin with the leader peptide still attached (20). For all lantibiotics investigated to date, removal of the leader sequence is required for biological activity.

Lantibiotics are secreted out of the cell by the LanT ATP-binding cassette transporters. These enzymes are transmembrane, homodimeric proteins that use the energy of ATP hydrolysis to secrete either the mature lantibiotic or a fully modified precursor peptide with the leader sequence still attached. A second type of LanT transporter is found in the biosynthetic machinery of class II lantibiotics, whose precursor peptides contain a "double-glycine" type cleavage site (21). These bifunctional LanT proteins contain an N-terminal proteolytic domain in addition to the membrane-spanning portion and C-terminal ATP-binding cassette. Cleavage occurs C-terminal to the G(−2)G/A(−1) motif at the junction between leader and structural region, and it is thought to take place concomitant with export.

Chemical Tools and Techniques

Because lantibiotics are gene-encoded antimicrobials, genetic, molecular biology, and protein chemistry techniques have been used traditionally to explore their structure and biosynthesis. These approaches include gene deletions and disruptions, and
Lantibiotics, Biosynthesis and Mode of Action of

in vivo and in vitro site-directed mutagenesis. The following sections will detail more recent tools used in various aspects of lantibiotic research with an emphasis on chemical biology techniques.

Discovery

Until recently, the discovery of new lantibiotics relied on the painstaking isolation and purification of compounds that exhibited bactericidal activity from the producer strains. Characterization of those isolates by amino acid analysis, and in some cases determination of their structure by NMR spectroscopy would classify them as a lantibiotic. More recently, the advent of bioinformatics has provided another avenue to discover new lantibiotics by scanning the ever-increasing number of fully sequenced genomes for genes homologous to known lantibiotic producing genes. The two-component lantibiotic haloduracin was discovered in this way (17, 22).

Structure determination

Over the years, several techniques have been developed to elucidate the structure of a new lantibiotic. The extensive post-translational modifications limit Edman degradation to a stretch of amino acids from the N-terminus to the first modification. Various chemical derivatization techniques have been used to aid in the resolution of the sequence and to reveal the position of the posttranslationally modified residues. Originally, these techniques relied on treatment with ethanethiol under highly basic conditions that resulted in the incorporation of dehydroamino acids and dehydrated residues already present in the parent structure. More recently, the structural determination of the LtnA1 and LtnA2 peptides of lacticin 3147 was aided by a novel method that involves nickel boride (Ni2B), an in situ generated hydrogenation and desulfurization catalyst (23). First, the N-terminal 2-oxobutyryl residue and the Dhb at position 2 of LtnA2 (see Fig. 3b) were removed by treatment with 1,2-diaminobenzene in aqueous acetic acid. Subsequent treatment with Ni2B in the presence of NaBD4 in CD3OD/D2O resulted in the incorporation of a single deuterium at the β-carbon of each residue that was involved in a (methyl)lanthionine linkage. Deuterium atoms are introduced at both the α- and β-carbon of dehydro amino acids, which allows distinction between (Me)Lan and Dha/Dhb that was not possible with the ethanethiol method. Both approaches are unable to determine the thioether ring topology, however, and modern NMR techniques and in some select cases tandem mass spectrometry have been used for this task. It should be noted that the thioether linkages restrict the use of tandem mass spectrometry methods to linear stretches of amino acids within the lantibiotics, because fragmentation within a ring does not result in fragment peptides (17). However, when tandem MS techniques are used in combination with site-directed mutagenesis to remove one or more rings from the parent structure, it offers a highly valuable approach to establish ring topology and requires much less material than NMR studies (16). None of these techniques can establish the stereochromy of the methyl/lanthionines with absolute confidence.
Lantibiotics, Biosynthesis and Mode of Action of

Figure 4  Posttranslational modifications in the biosynthesis of the class II lantibiotic lacticin 481. Both the dehydration and the cyclization events are catalyzed by the bifunctional protein LctM. The unmodified leader sequence is removed by the cysteine protease domain of LctT concomitant with transport of the mature lantibiotic outside the cell.

In early studies, this stereochemistry was established for a select few lantibiotics by comparison to chemically synthesized standards in combination with gas chromatography using chiral stationary phases. The stereochemistry determined in these early studies is assumed to be the same for lantibiotics that were discovered and characterized subsequently, but for most members this supposition has not been confirmed experimentally.

Mode of action

Chemical approaches have also contributed to the current understanding of the mode of action of lantibiotics. Prior to the discovery that nisin and other lantibiotics bind to lipid II, biophysical techniques, such as analysis of efflux of analytes from liposomes and other vesicles, as well as electrophysiology studies had established that many lantibiotics punch holes into membranes (4). More recently, NMR studies have elucidated the details of the molecular recognition used by nisin to sequester lipid II (8), and modification of nisin with fluorescent markers using synthetic chemistry showed that nisin removes lipid II from its functional location in the cell (7). Conversely, fluorescent labeling of lipid II showed that after initial docking of nisin with lipid II a higher order structure is assembled consisting of 4 lipid II molecules and 8 nisin molecules that make up the pores in the bacterial membranes (6).

Biomimetic studies

Several biomimetic studies involving the nonenzymatic, intramolecular cyclization of synthetic peptides that contain dehydroamino acids and unprotected cysteine residues have been used to investigate the cyclization step of lantibiotic biosynthesis. Short peptides that contain a single Dha or Dtb residue and a free cysteine undergo stereoselective cyclization to produce the naturally occurring isomers of (Me)Lan (24–26). These results indicate that the peptides have an innate propensity for Si-face attack on the dehydro amino acids and Re-face protonation of enolates to give the D-configuration at the u-carbons. This selectivity has a kinetic origin and is not caused by thermodynamic control (27). An attempt to prepare biomimetically the nisin A-ring (Lan) and B-ring (MeLan) was unsuccessful because of the much higher reactivity of dehydroalanines compared with dehydrobutyrines (27). Collectively, these studies show that although enzymatic control may not be required for the stereoselectivity of (Me)Lan formation, it is absolutely essential to govern the regioselectivity.

Mechanistic studies on the biosynthetic enzymes

The in vitro reconstitution of the enzymatic activities of LctM, NisC, and EpiD has paved the way to detailed mechanistic studies. The dehydration reaction catalyzed by the bifunctional LctM requires ATP for phosphorylation of the serine and the threonine residues that undergo dehydration (28). Usually, the phosphorylated intermediate is not observed with the wild-type substrate, but incorporation of certain nonproteinogenic amino acids into the substrate peptide by using expressed protein ligation resulted in the accumulation of phosphorylated peptides. Preparation of authentic phosphorylated substrate by ligation of synthetic phosphopeptides to a recombinant truncated LctA peptide that carries a C-terminal thioester confirmed that LctM catalyzes the elimination of the phosphate group to generate dehydro amino acids (28). Furthermore, high resolution Fourier transform mass spectrometric analysis of the dehydration reaction has shown that...
the dehydration reaction takes place by a processive mechanism (29). Mechanistic investigations of the EpiD enzyme that catalyzes the oxidative decarboxylation of the C-terminal cysteine in epidermin biosynthesis to produce the AviCys structure revealed that a (Z)-enethiol is generated that presumably adds to an unsaturated residue in a reaction catalyzed by EpiC (18).

Lantibiotic engineering

The cloning of the gene clusters involved in the biosynthesis of many lantibiotics laid the foundation for genetic protein engineering aimed at in vivo production of novel compounds with potentially interesting properties. Many studies have indicated the feasibility of changing the molecular structures of lantibiotics by mutagenesis of the prelantibiotic genes (30). In these investigations, not only the biosynthetic machinery, but also the immunity factors had to be considered to generate successful expression systems. Collectively, these studies have demonstrated the low substrate specificity of the biosynthetic enzymes involved. Disadvantages of reprogramming the structures of lantibiotics through in vivo engineering include a limited structural and functional space that can be sampled through mutagenesis and the potential for breakdown of immunity of the producing strain in cases where more active compounds are actually generated. The in vitro reconstitution of the complete biosynthetic pathways for lacticin 481 and haloduracin has provided the opportunity to explore the substrate specificity of their biosynthetic machinery in more detail. Using the expressed protein ligation technology, a series of Ser and Thr analogs embedded in the substrate peptide have been probed as potential substrates for the dehydration reaction, which demonstrates that the dehydration domain of lacticin 481 synthase has remarkable substrate promiscuity (31). Similarly, the cyclization domain was shown to tolerate Cys analogs such as selenocysteine and to convert them into the unnatural occurring (Me)Cys crosslinks (32). In addition to tolerating changes in the structure of the posttranslationalally modified amino acids, the enzyme also displays remarkable substrate promiscuity with respect to the position of the dehydrated residues as well as substitutions with nonproteinogenic amino acids at nonmodified positions. Similarly, the oxidative decarboxylase EpiD exhibits a very broad substrate scope (33). These investigations demonstrate that the engineering of novel structures by using the biosynthetic enzymes has great promise.

The Future of Lantibiotics

Many questions still remain with respect to both the biosynthesis and the mode of action of lantibiotics. Whereas the targets of the nisin, mersacidin, and cinnamycin groups are now known, the mechanism of action of many other lantibiotics (e.g., lacticin 481, Pep5, and sublancin) is still unclear. Similarly, the biosynthetic pathways still hold many unresolved questions, which include the molecular recognition that allows the synthetases their high level of substrate promiscuity and at the same time provides exquisite control of the regioselectivity of cyclization. Similarly, the biosynthetic enzymes for the majority of the 15 currently known posttranslational modifications in lantibiotics have not yet been identified, with the rate of discovery of new lantibiotics and new modifications exceeding the characterization of new biosynthetic enzymes.

Currently, the most common commercial application for lantibiotics is as a preservative in the food industry to combat foodborne pathogens and spoilage bacteria, but many other uses are under active investigation. Duramycin has been shown to increase chloride transport in nasal epithelial cells of cystic fibrosis patients (33), which in turn increases the fluidity of mucus in the lungs and airway and decreases the patient’s susceptibility to infections. Furthermore, several lantibiotics have shown potent activities against multidrug-resistant pathogenic bacterial strains. Combined with the development of new techniques to alter the structures of the lantibiotics, the future will likely see detailed SAR studies that may result in improved variants.

References

Lantibiotics, Biosynthesis and Mode of Action


Further Reading


See Also

Antibiotics, Biomimetic Synthesis of Antibiotics, Biosynthesis of Antibiotics from Microorganisms Antibiotics, Mechanism of Action
Oceans provide enormous and diverse habitats for marine life. The distinct feature of marine life is the domination of invertebrates, which account for more than 95% of marine animals. Most marine invertebrates are sessile and soft-bodied and lack obvious physical defenses. Instead, they have evolved to defend by chemical means against predation and overgrowth by other fouling organisms. In fact, their secondary metabolites have unusual structural features and potent biologic activities, many of which are not found in terrestrial natural products. This review focuses on bioactive metabolites isolated mainly from marine invertebrates with a special emphasis on the uniqueness of marine natural products.

The world’s oceans cover more than 70% of the earth’s surface and represent greater than 95% of the biosphere. Species ranging from 3 to 100 million are estimated to inhabit in the oceans. All but 1 of the 35 principal phyla in the animal kingdom are represented in aquatic environments; 8 phyla are exclusively aquatic. Most of them are sessile and soft-bodied organisms, most of which have evolved by chemical means to defend against predators and overgrowth by competing species. As expected, a variety of bioactive metabolites were found in marine invertebrates (see a series of reviews on marine natural products published in Natural Product Reports since 1984).

Bergmann’s revolutionary discovery of arabinose-containing nucleosides [e.g., spongouridine (1)] from the Caribbean marine sponge Cryptotheca crypta was a driving force for the birth of a new research field, Marine Natural Products Chemistry, in the early 1970s (1, 2). Nearly 18,000 new compounds, including polyketides, peptides, alkaloids, terpenoids, shikimic acid derivatives, sugars, and a multitude of mixed biogenesis metabolites, have been discovered during the last 40 years according to MarinLit (University of Canterbury, New Zealand). Many marine natural products have structural features previously unreported from terrestrial sources (3). Halogenated, especially brominated, and sulfated compounds are often encountered as marine natural products. Arsenic compounds, in particular arsenobetaines and arsenosugars, are distributed widely in marine algae and invertebrates. Recently, the first polyarsonic organic compound from Nature, arsenicin A (2), was reported from the sponge Echinostallina bargibanti (4). Also, several highly bioactive compounds with interesting modes of action have been discovered from marine invertebrates. A considerable percentage of these compounds was suggested or found to be derived from microorganisms. Indeed, certain bi-valves and c pajean fish become poisonous by feeding toxic dinoflagellates or herbivorous c pajean fish (5), whereas some cytotoxic metabolites of sponges and tunicates are produced by symbiotic microorganisms (6).

This review describes bioactive compounds isolated from marine algae and invertebrates with an emphasis on their uniqueness. Because of limited space, metabolites from bacteria, cyanobacteria, and fungi cannot be included, although some from cyanobacterial and endosymbiotic bacterial origins are described (some reviews on metabolites of marine bacteria, cyanobacteria, and fungi are provided in the “Further Reading” section). Structures and bioactivities are described for compounds that represent natural product classes, but steroids and carotenoids are not included.

Polyketides and Fatty Acid Derivatives

A diverse array of polyketide metabolites are found in marine organisms, which range from simple oxylipins to highly complex polyethers and macrolides. Particularly intriguing are “ladder-shaped” polyethers of dinoflagellate origin and sponge macrolides.

Fatty acid derivatives

Oxylipins, which are a major metabolite of fatty acids, are observed regularly in seaweeds and opisthobrachs that feed on seaweeds. Aplydilactone (3), which is an unusual oxylipin isolated from the sea hare Aplysia kurodai, activates phospholipase A₂ (7). More than 50 halogenated C₁₅ acetogenins, which are derived from fatty acids, have been isolated from red algae of the genus Laurencia; laurenin (4) was first isolated from L. nipponica (8). These compounds were reported to be antifeeding and insecticidal. Certain species of soft corals...
are known to contain prostanoids, of which the most intriguing are punaglandins (punaglandin 1 [(9), the first halogenated protaglandins isolated from the soft coral Telesto riei (3). They are antiviral and cytotoxic.

Fatty acid-derived cyclic peroxides are often found in marine sponges of the genus Plakortis; the first example was chondrin (6). 1,2-Dioxygen and 1,2-dioxolane carboxylates were also discovered from Plakortis sponges (9). These cyclic oxides show a wide range of biologic activities, for example, antimicrobial, cytotoxic, and antimalarial activities.

In addition, a variety of bioactive fatty acid derivatives have been isolated from marine organisms, including glyceroceramides (10). Perhaps the most unusual example is a highly chlorinated sulfolipid (11) isolated from the mussel Mytilus galloprovincialis (10).

Polycyclic lipids

The most prominent source of marine polycyclic lipids are sponges (11). Although the chain lengths vary from C14 to C49, likely they are to be derived from C16 fatty acids. Polypropionates show a wide range of bioactivities, including antimicrobial, cytotoxic, antiviral, and enzyme inhibitory.

Corticatic acid A (8), which is a C16 polypropionate carboxylic acid isolated from Petrosia corticata, is not only antimicrobial, but also it inhibits geranylgeranytransferase I (12).

Polyketides

Linear polyacetylenes are found frequently in marine sponges of the order Haplosclerida (11). Although the chain lengths vary from C4 to C24, likely they are to be derived from C18 fatty acids. Polyacetylenes show a wide range of bioactivities, including antimicrobial, cytotoxic, antiviral, and enzyme inhibitory.

Corticatic acid A (8), which is a C16 polypropionate carboxylic acid isolated from Petrosia corticata, is not only antimicrobial, but also it inhibits geranylgeranytransferase I (12).

Polycyclic lipids

The most prominent source of marine polycyclic lipids are sponges (11). Although the chain lengths vary from C14 to C49, likely they are to be derived from C16 fatty acids. Polypropionates show a wide range of bioactivities, including antimicrobial, cytotoxic, antiviral, and enzyme inhibitory.

Corticatic acid A (8), which is a C16 polypropionate carboxylic acid isolated from Petrosia corticata, is not only antimicrobial, but also it inhibits geranylgeranytransferase I (12).

Polyacetylenes

Linear polyacetylenes are found frequently in marine sponges of the order Haplosclerida (11). Although the chain lengths vary from C4 to C24, likely they are to be derived from C18 fatty acids. Polyacetylenes show a wide range of bioactivities, including antimicrobial, cytotoxic, antiviral, and enzyme inhibitory.

Corticatic acid A (8), which is a C16 polypropionate carboxylic acid isolated from Petrosia corticata, is not only antimicrobial, but also it inhibits geranylgeranytransferase I (12).

Polyketides

Dinoflagellates of the genus Amphidinium produce highly oxygenated, long-chain polyketides named amphidinolins (amphidinol 3 [(20)), that are highly antifungal and hemolytic (5). Polyacetylenes (13) are another class of highly unusual polyketide polyethers originally isolated from Irish seaweed that contains a polyether compound, namely highly antifungal gambieric acid (9). Isolated from Siphonaria grisea, whereas such unusual pyrrole-containing metabolites as siphonarin A (10) were isolated from S. zelandica. Marine polypropionates not only play defensive roles in mollusks, but also they show antimicrobial, antiviral, and cytotoxic activities.

Polyethers

Dinoflagellates are unique, aquatic photosynthesizing eukaryotes, that are classified in the kingdom Protoctista. They produce a variety of unusual polyketides with potent bioactivities (14).

The most unique metabolites are “ladder-shaped” polyethers: brevetoxin B (11), which was isolated from Gymnodinium breve (current). Karenia brevis is the first example of this group, and more than 15 brevetoxins have been isolated from the same source. Ciguatoxin (12) is a causative agent of ciguatera, a fish poisoning in subtropical and tropical regions, first isolated from a moray eel. It was later found to be originated from the dinoflagellate Gambierdiscus toxicus that contains a more complex polyether, maitotoxin (13), which is also involved in DSP (14). G. toxicus produces a variety of polyether compounds, namely highly antifungal gambieric acid (14). Maitotoxin, which is the most toxic natural product (LD50 value of 50 mg/kg in mice), increases membrane permeability to Ca2+, but the detailed mechanism remains unknown. Ciguatoxin is also highly toxic (1/10 potency of maitotoxin) and activates voltage-gated Na+ channels, whereas brevetoxins are much less toxic, although their mode of action is similar to that of ciguatoxin.

A different class of “ladder-shaped” polyethers, yesositoxins [e.g., yesostoxin 1 (25)], was isolated from toxic scallops; again its producer is the dinoflagellate Proteroceratium reticulatum (5). The mode of action seems to be different from that of brevetoxins and ciguatoxins.

The sponge Halichondria okadai contains a polyether metabolite named okadac acid (16) that was also isolated as a causative agent of diarrheal shellfish poisoning, and also isolated from mussels and other bivalves (14). However, the real producers are dinoflagellates of the genus Dinophysis. It is a potent cancer promoter that was found to be caused by inhibition of protein phosphatases 1 and 2A at nanomolar levels. Pectenotoxins are also involved in DSP and are produced by Dinophysis spp.; pectenotoxin 2 (17) inhibits actin polymerization (14).

Azaspiracids (azaspiracid-1 (18)) are another class of highly unusual polyketide polyethers originally isolated from Irish seaweed that contains azaspiracid shellfish poisoning (5). They are produced by the dinoflagellate Protoperidinium crassipes. A similar class of polyether toxins named pinnatoxins (pinnatoxin A (19)) were reported from the bivalve Pina pectinata; a closely related species P. attenuata is known to cause food poisoning in China. Pinnatoxins are likely of dinoflagellate origin and activate Ca2+ channels (15).

Open-chain polyketides

Dinoflagellates of the genus Amphidinium produce highly oxygenated, long-chain polyketides named amphidinolins (amphidinol 3 (20)), that are highly antifungal and hemolytic (5). Polyacetylenes (13) are another class of highly unusual polyketide polyethers originally isolated from Irish seaweed that contains a polyether compound, namely highly antifungal gambieric acid (9). Isolated from Siphonaria grisea, whereas such unusual pyrrole-containing metabolites as siphonarin A (10) were isolated from S. zelandica. Marine polypropionates not only play defensive roles in mollusks, but also they show antimicrobial, antiviral, and cytotoxic activities.

Polyethers

Dinoflagellates are unique, aquatic photosynthesizing eukaryotes, that are classified in the kingdom Protoctista. They produce a variety of unusual polyketides with potent bioactivities (14).

The most unique metabolites are “ladder-shaped” polyethers: brevetoxin B (11), which was isolated from Gymnodinium breve (current). Karenia brevis is the first example of this group, and more than 15 brevetoxins have been isolated from the same source. Ciguatoxin (12) is a causative agent of ciguatera, a fish poisoning in subtropical and tropical regions, first isolated from a moray eel. It was later found to be originated from the dinoflagellate Gambierdiscus toxicus that contains a more complex polyether, maitotoxin (13), which is also involved in DSP (14). G. toxicus produces a variety of polyether compounds, namely highly antifungal gambieric acid (14). Maitotoxin, which is the most toxic natural product (LD50 value of 50 mg/kg in mice), increases membrane permeability to Ca2+, but the detailed mechanism remains unknown. Ciguatoxin is also highly toxic (1/10 potency of maitotoxin) and activates voltage-gated Na+ channels, whereas brevetoxins are much less toxic, although their mode of action is similar to that of ciguatoxin.

A different class of “ladder-shaped” polyethers, yesositoxins [e.g., yesostoxin 1 (25)], was isolated from toxic scallops; again its producer is the dinoflagellate Proteroceratium reticulatum (5). The mode of action seems to be different from that of brevetoxins and ciguatoxins.

The sponge Halichondria okadai contains a polyether metabolite named okadac acid (16) that was also isolated as a causative agent of diarrheal shellfish poisoning, and also isolated from mussels and other bivalves (14). However, the real producers are dinoflagellates of the genus Dinophysis. It is a potent cancer promoter that was found to be caused by inhibition of protein phosphatases 1 and 2A at nanomolar levels. Pectenotoxins are also involved in DSP and are produced by Dinophysis spp.; pectenotoxin 2 (17) inhibits actin polymerization (14).

Azaspiracids (azaspiracid-1 (18)) are another class of highly unusual polyketide polyethers originally isolated from Irish seaweed that contains azaspiracid shellfish poisoning (5). They are produced by the dinoflagellate Protoperidinium crassipes. A similar class of polyether toxins named pinnatoxins (pinnatoxin A (19)) were reported from the bivalve Pina pectinata; a closely related species P. attenuata is known to cause food poisoning in China. Pinnatoxins are likely of dinoflagellate origin and activate Ca2+ channels (15).
its primary target is G-actin. Recently, it was discovered from cyanobacteria, although its eubacterial origin was predicted (6). Another class of macrolides that inhibit actin polymerization is the tris-oxazole-containing macrolides, namely kabiramides and mycalicides (1). Kabiramide C (26) was isolated from eggmasses of a nudibranch of the genus Hexabranchus, whereas mycalicides are of sponge origin, which indicates kabiramides were sequestered by nudibranchs from sponges. These macrolides are potently antifungal and cytotoxic. They bind G-actin in a molar ratio of 1:1, which thereby inhibits actin polymerization. Similar macrolides named alyponines were isolated from the sea hare Aplysia kurodai; alyponine A (27) shows remarkable antitumor activity (T/C > 500% at 0.08 mg/kg against P388 leukemia cells) as well as a similar mode of action to that of kabiramides (7). Presumably, alyponines are of cyanobacterial origin.

Halichondrin B (28), which is a polyether macrolide isolated from H. okadai, shows promising antitumor activity and has entered phase I clinical trials (2, 15). It inhibits polymerization of tubulin by binding to the colchicine domain. The macrocyclic portion seems to be essential for the activity (2). Halichondrins have been found in several species of sponges, which indicates their microbial origin (6). Spongistatins/hyrtiostatins/cinachyrolide A [spongistatin-1 (29)] are highly unusual, 35-membered macrolides isolated from several sponges (3). They inhibit growth of tumor cells at sub-nM levels by binding to the vinca domain of tubulin. Their low yields and occurrence in several different sponges suggest their microbial origin. Spirastrellolide B (30) was reported recently as an antimitotic agent from the sponge Spirastrella coccinea. It is actually a potent and selective inhibitor of protein phosphatase 2A (12).
Peloruside A (31), which is a 16-membered, highly oxidized macrolide from the sponge Mycale hentschelii, induces tubulin polymerization (2). 13-Deoxytedanolide (32) isolated from Japanese sponges of the genus Mycale shows promising antitumor activity. It inhibits protein synthesis by binding to a 70 S large subunit of eukaryotic ribosome (17).

Bryostatins are medicinally important macrolides discovered from the cosmopolitan bryozoan Bugula neritina (2). Twenty bryostatins, which all possess a 20-membered ring, are known to date. Bryostatin 1 (33) showed good antitumor activity; it selectively modulates protein kinase C (12).
Phorboxazole A (34) is an unusual oxazole-containing macrolide isolated from a sponge Phorbas sp. (18). It is highly antifungal and cytostatic.

Salicylihalamide A (35) is a salicylic acid-containing macrolide enamide isolated from a sponge Haliclona sp., and it inhibits V-ATPases at a low nM concentration. Members of this family have been isolated from sponges, tunicates, and bacteria (12).

PKS/NRPS metabolites
Bengamides were isolated originally as anthelmintic agents from a sponge Jaspis sp. Later, bengamide A (36) was found to inhibit growth of tumor cells as well as methionine aminopeptidases (2).

Marine natural products of the pederin class (mycalamides, onnamides, and theopederins) isolated from sponges are mixed biogenesis metabolites of polyketide synthase and nonribosomal synthase (1). In fact, biosynthesis gene clusters of this class have been cloned recently using metagenomic techniques from the marine sponge Theonella swinhoei (6). These compounds are...
potently cytotoxic; theopederin A (37) inhibits protein synthesis in a similar mode of action to that of 13-deoxytedanolide.

Calyculin A (38) is an extraordinary metabolite composed of C28 fatty acid and two γ-amino acids isolated from the sponge *Discodermia calyx*. It is not only highly antifungal and antitumor but also a potent cancer promoter that was found to be caused by potent inhibition of protein phosphatases 1 and 2A (1). More than 15 calyculin derivatives were isolated from several marine sponges, which indicates the involvement of symbiotic microorganisms in the production of calyculins (6).

Pateamine (39) is a macrolide isolated from a marine sponge *Mycale* sp. Its potent cytotoxicity is attributed to inhibition of transcriptional initiation (19).

### Aromatic polyketides and enediyne

Aromatic polyketides are rare metabolites in marine invertebrates. Naphthoquinone and anthraquinone derivatives were reported as echinoderm pigments in the earlier stage of marine natural products research. The most interesting aromatic polyketides are the pentacyclic polyketides of the halenaquinone/halenaquinol class. Halenaquinone (40) was first isolated from the sponge *Xestospongia exigua*. Compounds of this class show a variety of biologic activities, which include inhibition of tyrosine kinase (12, 19).

Only two groups of enediyne have been found from marine organisms; namenamycin (41) was first isolated from the *Fijian tunicate* *Polysyncraton lithostrotum*, whereas shishijimicins A (42)–C, β-carboline-containing enediyne, were isolated from the Japanese tunicate *Didemnum proliferum*, along with 41, which
thus suggests the involvement of symbiotic microorganisms in their production (1). As expected, these compounds inhibited growth of tumor cells at pM levels.

**Nonribosomal Peptides and Amino Acid Derivatives**

In addition to ribosomal peptides, some of which exhibit interesting bioactivities as is the case of conotoxins, marine organisms, in particular, sponges and tunicates contain a wide variety of nonribosomal peptides, many of which contain unusual or unprecedented amino acids. It should be noted that these peptides show a range of biological activities.

**Amino acid derivatives**

Microalgae and macroalgae often contain UV-absorbing amino acids collectively dubbed "mycosporines." Mycosporine-glycine (43) is most widely distributed in marine organisms that use it for protection from UV irradiation. Domoic acid (44) is not only a helminthic agent originally isolated from a red alga, but also it is a causative agent of amnesic shellfish poisoning (14). It is a potent agonist of glutamate receptors. Dysiherbaine (45), which is a novel betaine isolated from the sponge Dysidea herbacea, is a potent non-NMDA-type agonist with very high affinity for kainate receptors (20).
Giroline (46), which is a potent cytotoxin isolated from the sponge Pseudaxinyssa cantharella, inhibits protein synthesis (2).

**Bromotyrosine derivatives**

Bromotyrosine-derived metabolites are often encountered in marine sponges of the families Aplysinidae and Pseudoceratidae, in particular Pseudoceratina (= Psammaplysilla) purpurea. They show a variety of biological activities, which include antimicrobial, enzyme inhibitory, and antifouling activities. Psammaplysin A (47) is antimicrobial, cytotoxic, and antifouling, whereas psammaplin A (48) is an inhibitor of histone deacetylase (2). The marine sponge Ianthella basta synthesizes at least 25 bastadins that are linear or cyclic peptides composed of four bromotyrosine residues [bastadin 5 (49)] and show antimicrobial, cytotoxic, and enzyme inhibitory activities as well as interaction with Ca²⁺ channels (21).

**Linear peptides**

Sponges and tunicates frequently contain unusual linear and cyclic peptides; those from the former source were reviewed in 1993 and 2002 (22, 23). To avoid duplication, this review focuses on bioactive peptides isolated from other sources and new findings. Gymnanjamide (50) is the first described peptide from a hydroid (24). This pentapeptide from Gymnanium regae contains three previously unknown amino acids. Although modestly cytotoxic, its structure is reminiscent of dolastatin 10 (33), which is a powerful antibitin agent of the sea hare Dolabella auricularia (2). Actually, dolastatins are of cyanobacterial origin. Dysinosin A (52) is a novel inhibitor of factor VIIa and thrombin, which was isolated from a new genus and species of an Australian sponge of the family Dysideidae (12). Perhaps the most intriguing linear peptide is polytheonamide B (53), which is a potent cytotoxin discovered from T. swinhoei.
(25). It is composed of 48 amino acid residues, most of which are unusual amino acids. More significantly, it has the sequence of alternating D- and L-amino acids.

Cyclic peptides

Most cyclic peptides have been isolated from sponges and tunicates. Azumamides A (54) through E, which are cyclic tetrapeptides isolated from the sponge *Mycale laevis*, are the most recent addition to the list of cyclic peptides. They strongly inhibit histone deacetylase (12). Dolastatin 11 (55), which is a cyclic depsipeptide isolated from *D. auricularia*, stabilizes actin filaments as in the case of jaspakinolide/jaspamide. Kahalalide F (56) is a cyclic depsipeptide isolated from the sacoglossan mollusk *Elysia rufescens*. It shows promising antitumor activity and has entered Phase II clinical trials, but its mode of action is not clear (2). Dolastatins and kahalalides likely are sequestered by the mollusks from cyanobacterial diets. A scidean peptide often contain thiazole/thiazoline and oxazole/oxazoline amino acids as represented by patellamide A (57) isolated from *Lisoclinum patella* (3). These cyclic peptides show cytotoxic activity. Probably the most well-known ascidian peptide is didemnin B (58) isolated from *Trididemnum solidum*. This depsipeptide showed remarkable antitumor activity and entered clinical trials, but it was dropped because of side effects (2). It inhibits protein synthesis. Aplidine (dehydrodidemnin B) (59) isolated from *Aplidium albicans* is more promising as an anticancer drug, although it shows multiple modes of action (2).
Marine Natural Products: Chemical Diversity of O

H2N

O

O

N

Me2

39

O

OO

O

40

41: R = OH

O

SMe

HO

NHCO2Me

SSSMe

H

R

42: R = OH

N

43

N

44

45

N

46

N

47

48

Marine organisms produce a wide range of alkaloids with potent bioactivities, which include such specific classes as 3-alkylpiperidine, guanidine, indole, polyanine, pyridoacridine, and pyrrole-imidazole alkaloids. Their biological activities vary from antimicrobial to neurological.

3-alkylpiperidines

A variety of 3-alkylpiperidine-derived compounds have been obtained from sponges belonging to five families of the order Haplosclerida (26). They show a range of bioactivities, for example, cytotoxic, antimalarial, and antifouling. It is likely that 3-alkylpiperidines are produced by sponge cells but not by symbiotic microorganisms.

The first 3-alkylpiperidine derivative reported is halitoxin (60), which was isolated from Haliclona sp. Similar polymeric alkylpyridines are also known from several sponges. In addition to the polymers, various types of metabolites of 3-alkylpyridines or 3-alkylpiperidines have been isolated, namely, macrocyclic bis-3-alkylpiperidine [telomerase-inhibitory cyclostellettamine A (61)] (12). Bis-quinolizadine [petrosin A (62)] and bis-1-oxaquinolizadine macrocycles [xestospongin C (63)] isolated from Petrosia seriata and Xestospongia spp., respectively, also belong to this group. The former is ichthyotoxic, whereas the latter is a potent vasodilator as well as an inhibitor of IP3 receptor. Halicyclamine (64), which was isolated from Haliclona sp., is another group of macrocyclic bis-3-alkylpiperidines, whereas sarain A (65), which was isolated from Reniera sarai, has a more complex polycyclic core. These compounds are moderately cytotoxic.

The most well-known group of 3-alkylpiperidine alkaloids is the manzamines. Manzamine A (66), which is the first member of this group, was isolated from an Okinawan Haliclona sp. More than 20 manzamines have been isolated from sponges of eight different genera. Manzamine A is highly cytotoxic,
antituberculosis, and antimalarial, but its mode of action remains unknown (27).

**Pyrrole-imidazoles**

The pyrrole-imidazole alkaloids are found exclusively in marine sponges, in particular in the families Agelasidae, Axinellidae, and Halichondridae (28). Oroidin (67), which was discovered from *Agelas oroides*, is the building block of about 100 metabolites of this family, which can be divided into those derived from different modes of cyclization, for example, (-)-dibromophakelin (68) isolated from *Phakellia flabellate* and ageladine A (69) from *Agelas nakamurai*, and those derived from different modes of dimerization, namely sceptrin (70), which is isolated from *Agelas sceptrum*, palau’amine (71) from *Stylotella aurantium* and massadine (72) from *Stylissa massa*. Quite recently, a dimer of massadine (oridin tetramer) named stylissadine A was isolated from *Stylissa caribica*. The pyrrole-imidazole alkaloids show a range of bioactivities, which include antimicrobial, cytotoxic, antagonistic to receptors, immunosuppressive, enzyme inhibitory, and antifouling.

**Pyridoacridines and related alkaloids**

Pyridoacridines are highly colored polycyclic alkaloids mainly isolated from sponges and tunicates (29). They are divided into four groups, the simplest of which is styelsamine D (73), which was isolated from the ascidian *Eusynstyela latericius*. The pentacyclic pyridoacridines are also classified into two groups as represented by amphimedine (74), which was isolated from a sponge *Amphimedon* sp., and ascididemin (75), which was isolated from a tunicate *Didemnum* sp., respectively. Dercitin (76) is a thiazole-containing pentacyclic alkaloid isolated from a sponge *Dercitus* sp., whereas cyclodercitin (77), which was
isolated from a sponge Stelletta sp., is a member of the hex acyclic group. The most complex octacyclic pyridoacridines are represented by eudistone A (79), which was isolated from Eudistoma sp.

Alkaloids related to pyridoacridines are known also from sponges and tunicates. Discornin B (79), which is the first marine pyridoquinoline alkaloid, was isolated from the sponge Latruncula cf. bocagai. More than 20 alkaloids of this class are known at present (30). A different example of this family is batzelline A (80) from a deepsea sponge of the genus Bathylamella and wakayin (81) from an ascidian Clavelina sp.

Many pyridoalkaloids show a wide range of biological activities, namely antimicrobial, antiviral, antiparasitic, insecticidal, antitumor, and enzyme inhibitory (12, 29, 30).

Indole alkaloids

Highly substituted pyrroles are often found in sponges, ascidians, and mollusks (30). Lamellarins A (82–D) are the first metabolites of this group that were reported from a mollusk Lamellaria sp., followed by the discovery of more than 50 alkaloids of this class from sponges and ascidians. Polycitone A (83) was isolated from a tunicate Polycitor sp., whereas storniine A (84) was isolated from a sponge of the genus Cistella. The most recent addition to this class is dictyodendrins A (85) through E isolated as telomerase inhibitors from the sponge Dicytostelidella serpulaforma (12). Various bioactivities have been reported for lamellarins and related compounds, which include antitumor, antiviral, and enzyme inhibitory.

Indole alkaloids

Many indole-containing metabolites have been reported from marine organisms, some of which were already mentioned. This section focuses on important indole-containing substances that belong to structural classes not mentioned above. These alkaloids show antimicrobial, antiparasitic, antitumor, and enzyme inhibitory activities (27). Dragmacidin (86) is a novel bis-indole isolated from a deepsea sponge Dragmacidon sp., whereas another deepsea sponge Spongosorites ruetzleri contains a similar bis-indole, topsentin (87).

β-carboline-containing metabolites are known from sponges and tunicates. Eudistomins were the first β-carboline alkaloid isolated from marine organisms. Eudistomin K (88), which is a novel oxahexitrione ring containing β-carboline, was isolated from the tunicate Eudistoma olivaceum, whereas a guanidine-containing grossularine 1 (89) was from the tunicate Dendrosoma grossularia (27).

Physisostigmine alkaloids are often contained in bryozoans. The highly unusual constituent is secureine A (90) isolated from Sarcinifustra secernifrons from the North Sea. Another interesting physisostigmine is urochordamine A (91) isolated as a larval settlement inducer from the ascidian Ciona savignyi (30). Neosurugatoxin (92), which is a causative agent of “ivory shellfish poisoning” isolated from the gastropod Babylonia japonica, is a reversible nicotinic acetylcholine antagonist (14). Its bacterial origin has been suggested. Finally, diazonamide A (93) is a highly unusual cytotoxic metabolite of the tunicate Dialeana angulata. It inhibits tubulin polymerization (2).

Guanidine alkaloids

Many guanidine-containing compounds have been reported from diverse marine organisms (31). No doubt exists that the most well-known guanidine alkaloid is tetrodotoxin (94), which is a causative agent of puffer fish poisoning (14). It is highly toxic because of inhibition of voltage-gated Na⁺ channels. Production of tetrodotoxin by bacteria of such genera as Pseudalteromonas and Vibrio was reported. Similarly, saxitoxin (95) causes paralytic shellfish poisoning. Bivalves accumulate the toxin in dinoflagellates, for example, Alexandrium catenella, A. tamarense, and Gymnodinium catenatum. Its mode of action is similar to that of tetrodotoxin. More than 30 saxitoxin derivatives are known to date.

Pillomycolin A (96) is a novel polycyclic guanidine alkaloid isolated from the sponge Piliocaulis spiculifer (31). It is highly cytotoxic, antifungal, and antiviral. The related alkaloids were reported from the Mediterranean sponge Crambe crambe that also contains other types of guanidine alkaloids such as...
crambescin A (97). Batzelladines are a similar class of alkaloids isolated from a sponge Batzella sp.; batzelladine A (98) shows anti-HIV activity.

Variolin B (99), which is a pyrido[3,4-d]pyrimidine alkaloid isolated from the sponge Kirkpatrickia variafoxa, is strongly cytotoxic, antifungal, and antiviral (30). It inhibits cyclin-dependent kinases (12).

**Polyamine alkaloids**

Stellettadine A (100), which is isolated from a sponge Stelleta sp., represents the first polyamine alkaloid with an arcaine backbone (32). Similar alkaloids (e.g., stellettazole A (101)) are also known from the same source. These compounds show larval settlement-inducing, antifungal, and enzyme inhibitory activities. Isoquatenated polyamines are encountered occasionally as metabolites of soft corals as represented by sinulamide (102), which inhibits H,K-ATPase (12).

**Quinolines and isoquinolines**

Aaptamine (103), which is a cytotoxic benzonaphthyridine alkaloid isolated from the sponge Aaptos aaptos, induces differentiation in chronic leukemia cells (33). Schulzeine A (104) is a novel dihydroquinoline alkaloid isolated from the sponge Penares schulzei that inhibits glycosidases (12).

Sponges and tunicates contain tetrahydroisoquinoline alkaloids (30). Renieramycin A (105) from a sponge Reniera sp.
represents the first example of this particular class of alkaloids isolated from marine organisms. Ecteinascidins, in particular ecteinascidin 743 (106), which was isolated from the tunicate Ecteinascidia turbinata, are promising as anticancer leads. Ecteinascidin 743 cleaves DNA chains and is in clinical trials (2).

Quinolizidines and indolizidines

Only a few compounds of these classes of alkaloids have been reported from marine organisms. Clavepictine B (107), which was isolated from the ascidian Clavelina picta, is marginally cytotoxic. Stellettamine A (108), which is an indolizidine derivative from a sponge Stelletta sp., is antifungal and cytotoxic. It also inhibits calmodulin (12). Lepadiformine (109), which has a similar structural feature isolated from the tunicate Clavelina inoluccensis, inhibits K+ channels (33).

Steroidal alkaloids

Steroidal alkaloids belong to a rare class of marine natural products. Plakinamine A (110), which is an antimicrobial metabolite from a sponge Plakina sp., is the first steroidal alkaloid isolated from marine organisms. Cephalostatins and ritterazines are unprecedented dimeric steroidal alkaloids isolated from the hemichordate Cephalodiscus grichristii and the tunicate Ritterella tokioka, respectively (1, 30). They are specific to marine metabolites. Cephalostatin 1 (111) and ritterazine B (112) are
Marine Natural Products, Chemical Diversity of

highly cytotoxic, but their mode of action remains to be elucidated.

Although unlikely steroids or triterpenoids, zoanthamines found in zoanthids of the genus Zoanthus should be mentioned here. Zoanthamine (113) represents the first example of 10 metabolites of this class. Norzoanthamine has been best studied and found to suppress the production of IL-6 (15).

Terpenoids

Although those similar to terrestrial terpenoids are found commonly in marine organism, in particular in algae, several terpenoids with new or modified skeletons have been isolated frequently from algae, sponges, and crinidarians. Halogenated terpenoids are often contained in algae, whereas sulfated terpenoids and steroids are distributed widely in sponges.

Monoterpenoids and Sesquiterpenoids

Red algae contain highly halogenated monoterpenoids such as 114, which is an antifeeding constituent isolated from Plumaria cartilagineum (34, 35).

Several halogenated sesquiterpenes with various skeletal types were reported from red algae of the genus Laurencia, for example, elatol (115), which is a chamigrane sesquiterpene that has antifeeding and antifouling activities from L. obtusa. A series of linear and cyclic sesquiterpenes with an 1,4-diacyctobybutadiene functionality [e.g. caulerpenyne (116)] have been isolated as antifeeding agents from green algae of
the order Caulerpales (35). Coelenterates are a rich source of sesquiterpenes of 20 skeletal types, which include \( \Delta^{9(12)} \)-capnellane (117), an antifeeding metabolite isolated from Capsella imbricata (36).

Sponges produce furanosesquiterpenes of various skeletal types, such as furodysinin (118) from Dysidea herbacea and nakafuran-8 (119) from D. etheria, the latter of which is an antifeeding. Similarly, mixed shikimate-mevalonate metabolites are often encountered as sponge metabolites; the first example is avarol (120), which was isolated from Dysidea avara. Several related compounds have been isolated from dictyoceratid sponges. They show a wide range of bioactivities, for example, antimicrobial, antitumor, anti-inflammatory, and enzyme inhibitory.

Diterpenoids

Brown algae are rich in diterpenoids, which can be classified into three groups, namely "xenicanes," "extended sesquiterpenes," and "dolabellanes," which are represented by dictyotadial (121) from Dictyota crenulata, spatul (122) from Spatoglossum schmittii, and amijidictyol (123) from Dictyota linearis, respectively.
These terpenes are involved in chemical defense. Red algae contain brominated diterpenes of several established skeletal types. Spongian diterpenes are a chemical marker for dictyoceratid and dendroceratid sponges (37). The first example of a spongian diterpene is isoagatholactone (124) isolated from Spongia officinalis. In addition to those possessing a basic spongian skeleton, a wide variety of rearranged spongian diterpenoids have been reported, which include gracillin B (125) from Spongionella gracilis. Many spongian diterpenes are antimicrobial and cytotoxic.

Again, soft corals and gorgonians are a rich source of diterpenoids of 19 structural classes, some of which are specific to them (35, 36). Besides tobacco plants, cembranoid diterpenes are limited to soft corals. Lophotoxin (126) isolated from sea whips of the genus Lophogorgia is a sodium channel inhibitor (3). Xenicin (127) from the soft coral Xenia elongata and briarein A (128) from the gorgonian Briareum asbestinum represent non-cembranolide diterpenes. Diterpenoids of these classes show antimicrobial, cytotoxic, and insecticidal activities.
Pseudopterosin A (129), which is a diterpene glycoside isolated from the sea whip Pseudopterogorgia elisabethae, shows anti-inflammatory activity by inhibiting release of leukotriene B\(_2\) from leukocytes (2, 35). A more interesting class of diterpenes includes sarcodictyn A (130) and eleutherobin (131), which were isolated from the soft corals Sarcodictyon roseum and Eleutherobia sp., respectively. These diterpenoids show potent cytotoxicity by stabilizing microtubules (2, 35).

Ageladine A (132), which is a 9-methyladenine derivative of a diterpene isolated from a sponge Agelas sp., is antimicrobial and inhibitory against Na, K-ATPase (12). A novel, chlorinated diterpenoid, chlorolissoclimide (133), which was isolated from the ascidian Lissoclinum forskali, inhibits eukaryotic protein synthesis (38).

Isocyanoterpenes and related terpenoids

Isocyanide-containing natural products are rare; they have been reported only from cyanobacteria, Penicillium fungi, marine sponges, and nudibranchs (35, 39). Sesquiterpenoid and diterpenoid isocyanides are found in a limited species of sponges and nudibranchs that prey on these sponges. Axisonitrile-1 (134) isolated from the sponge Axinella cannabina is the first isocyanide-containing marine natural product. Isocyanopupukeanane (135) was isolated originally from the nudibranch Phylidia verrucosa and later from a sponge Ciocalypta sp. Kalihinol A (136) and 7,20-diisocyanoadocane (137) were isolated from the sponges Acanthella carvenosa and Adocia sp., respectively.
respectively. Isocyanoterpenes are often accompanied by thio-
cyanates, isothiocyanates, and formamides. These terpenoids
display a wide range of bioactivities, which include antimicro-
bial, cytotoxic, ichthyotoxic, antifouling, and antimalarial. The
carbonimidic dichloride group is considered to be equivalent
to isocyanide, and the first sesquiterpenoid that contains this
moiety was isolated from the sponge *Pseudoaxinyssa pitys.*

Terpenoids that contain carbonimidic dichloride show similar
bioactivities to those of isocyanide counterparts.

**Sesterterpenoids, triterpenoids, and steroids**

Variabilin (**139**), which is a C25 tetronic acid isolated from
the sponge *Ircinia variabilis,* represents the class of furano-

Several unusual polysaccharides have been isolated from sponges (34). Bioactivities include antimicrobial, cytotoxic, antimalarial, and antitumor.

Mannolide (140) is perhaps the most well-known marine sesquiterpene isolated from the sponge *Luffariella variabilis* (40). A number of sesquiterpene glycosides have been isolated from the sponge *Axinella infundibula* (12). It resembles bacterial 13-deoxytedanolide, a marine sponge-derived antitumor macroclide, which is the first natural product inhibitor of cdc25A (12). Mycaperoxide A (143), which is a norsesterepene peroxy isolated from *Mycale* sp., showed promising antimalarial activity (9).

Probably one of the most interesting marine triterpenoids is the squareni-14,18-diene derived polyethers found in red algae of the genus *Laurencia*. Among the nearly 30 metabolites of this class, thyrone (141), which has been isolated recently as a potent telomerase inhibiter from *Axinella infundibula* (12). Mucaperoxide A (143), which represents the first example of this class. A dicuslate-2 (141) is a member of the triterpenoid hydroquinone sulfates isolated from an ascidian (aka Ascidia) sp. that is the only known natural product inhibitor of kinase (12, 19).

Didemneketal C (147) is an unusual heptaprenoid isolated from an ascidian *Didemnum sp.*; its methanolysis product, didemnaketale B (148), inhibited HIV-1 protease (12).

**Sugars**

Several unusual polysaccharides have been isolated from sponges (11), among which the most unusual is asirennolide A (149), which has been isolated recently as a potent telomerase inhibitor from *Axinella infundibula* (12). It resembles bacterial lipopolysaccharides.

**Acknowledgments**

I would like to thank Otto Hensens for editing the manuscript.

**References**


21. Costadopuros EA, Pitsinos EN, Moutsos VI, Sarakinos G. A general method for the synthesis of battoxanes and isobattoxanes:


24. Milanowski DJ, Gustafson KR, Rashid MA, Rannell LK, McMa-
hoon JB, Boyd MR. Gymnangiamide, a cytotoxic pentapeptide from the Micronesian marine sponge Gymnangium regae.

25. Hamada T, Matsunaga S, Yano G, Fusetani N. Polythenamides A and B, highly cytotoxic, linear polypeptides with unprecedented

26. Andersen RJ, van Soest RWM, Kong F. 3-Alkylpiperidine alkaloids isolated from marine sponges in the order Haplosclerida.

27. Gul W, Hamann MT. Indolic alkaloid marine natural products: an established source of cancer drug leads with considerable promise

28. O’Malley DP, Li K, Mau M, Zografi A, Libar PS. Total syn-
thesis of dimeric pyrrole-imidazole alkaloids: sceptrim, aegiferin, nagatamide E, oxysceptrin, nakamuranic acid, and the azelina-

37.


33. Folmer F, Houssen WE, Scott RH, Jaspars M. Biomedical re-

34. Faulkner DJ. Interesting aspects of marine natural products chem-

2006;5:115–141.

36. Col J. The chemistry and chemical ecology of echinoderms (Ooe-

37. Keyzers RA, Northcote PT, Davies-Coleman MT. Spongin dian-

38. Robert F, Gao HQ, Donia M, Merrick WC, Hamann MT, Pel-


Further Reading

Allingham JS, Krulich VA, Rayme I. A hit-to-lead in marine natural prod-

Antunes EM, Copp BR, Davies-Coleman MT, Samaa T. Pyrroliomoi-

Bandaranayake WM. The nature and role of pigments of marine inver-

Blunt JW, Copp BR, Hu WP, Munro M-HG, Northcote PT, Princep M.

Boerman EJ, Bowman BJ. V-Antipsides as drug targets. J. Bioenerget.

Burja AM, Barajai, A. Bou-Mansour E, Burgess JD, Wright PC.


Crews P, Hunter LM. The search for antiparasitic agents from marine
animals. In: Marine Biotechnology, vol 1. Pharmaceutical and Bioac-


Dambbtsky VM, Levitsky DD. Amanotoxins. Prog. Lipid Res. 2004;43:
403–446.

348.

and earlier reviews cited within.

Fenical W, Jensen PR. Developing a new resource for drug discovery:

Gerwick WH, Bernat MW. Eicosanoids and related compounds from
marine algae. In: Marine Biotechnology, vol 1. Pharmaceutical and Bioac-

Glickman GI. The diversity of naturally occurring organobromine com-

Hoy D, Moore CG, Murphy PJ. The guanidine metabolites of Pol-
cauls spiculifer and related compounds; isolation and synthesis.

Ireland CM, Copp BR, Foster MP, McDonal LA, Radaski DC, Swaray
JC. Biomedical potential of marine natural products. In: Marine Bio-
technology, vol 1. Pharmaceutical and Bioactive Natural Products,


Kompert JM, Sillanenae C, Barnathan G. Sulfated compounds from

See Also
Natural Product Discovery
Natural Products in Microbes
Natural Products in Plants
Natural Products, Common Biological Targets for
Natural Products: an Overview
The carbon skeleton of isoprenoids is derived from the branched C_5 skeleton of isoprene. Isopentenyl diphosphate and dimethylallyl diphosphate represent the biologic equivalents of isoprene. From research on cholesterol biosynthesis in liver tissues and on ergosterol in yeast, mevalonate was accepted as the universal precursor of isoprenoids. However this assertion is inaccurate. Incorporation of labeled acetate and glucose isotopomers into bacterial isoprenoids and into diterpenes of ginkgo embryos indicated fortuitously the existence of an alternative mevalonate-independent route. Its full elucidation required experiments using ^13C- and ^2H-labeled precursors followed by extensive nuclear magnetic resonance analyses as well as a combination of biochemical and molecular biology methods. These additional studies revealed a complete set of novel unsuspected enzymes.

Isoprenoids represent the largest family of natural products, with an exceptional structural diversity. Isoprenoids are present in all living organisms. This group includes essential metabolites, such as sterols of the eukaryotic plasma membranes, prenyl chains of the quinones and carotenoids from electron transport chains, and carotenoids from the photosynthetic apparatus in the plant chloroplasts, or in the phototrophic bacteria. Isoprenoids also include secondary metabolites of a more restricted distribution and with a less obvious physiologic significance. Their carbon skeleton can be derived from the combination of C_5 subunits with the branched skeleton of isoprene.

Steps for the Formation of Isoprene Units

The biologic precursors of isoprene units are isopentenyl diphosphate and dimethylallyl diphosphate. These precursors can be obtained by two different metabolic pathways: the mevalonate (MVA) pathway, which was the first one to be elucidated, and the long overlooked methylerythritol phosphate (MEP) pathway.

Mevalonate (MVA) pathway

Labeling experiments, performed mainly with liver tissues and yeast, elucidated the biosynthesis of cholesterol and ergosterol, leading to the discovery of the mevalonate pathway leading to IPP and DMAPP. In this pathway, isoprene units are derived, like fatty acids, from acetyl-coenzyme A. The key intermediate is mevalonate, which results from the reduction of hydroxymethylglutaryl-coenzyme A catalyzed by the HMG-CoA reductase, and represents the committed step of the pathway. Confirmation of this pathway was obtained for the biosynthesis of plant sterols, triterpenes, and sesquiterpenes.

Discrepancies with the MVA pathway and the discovery of an alternative route: the MEP pathway

The MVA pathway was accepted as the unique biosynthetic pathway for the formation of all isoprenoids in all living organisms. Discrepancies with this general assertion appeared, however, as early as the 1950s. For instance, ^14C-labeled MVA was not incorporated into chloroplast isoprenoids (e.g., carotenoids and phytol from chlorophylls; Fig. 6), whereas it was well incorporated into phytosterols synthesized in the cytoplasm. Unexpected labeling patterns were found in the prenyl chain of ubiquinone. In Escherichia coli at incorporation of ^14C-labeled acetate, finally, the labeling pattern in an isoprene unit from the sesquiterpenic pentalenene series from a Streptomyces species at incorporation of uniformly labeled acetate led to the discovery of the mevalonate pathway leading to IPP and DMAPP. In this pathway, isoprene units are derived, like fatty acids, from acetyl-coenzyme A. The key intermediate is mevalonate, which results from the reduction of hydroxymethylglutaryl-coenzyme A catalyzed by the HMG-CoA reductase, and represents the committed step of the pathway. Confirmation of this pathway was obtained for the biosynthesis of plant sterols, triterpenes, and sesquiterpenes.
Methylerythritol Phosphate Pathway for the Formation of Isoprene Units

Figure 1 | Mevalonate pathway for the biosynthesis of isopentenyl diphosphate 7 and dimethylallyl diphosphate 8.

Figure 2 | Methylerythritol phosphate 12 pathway for the biosynthesis of isopentenyl diphosphate 7 and dimethylallyl diphosphate 8.

Labeled [U-13C6]glucose was not in accord with what was expected from the MVA pathway.

The discovery of an alternative mevalonate-independent route to isoprene units was a fortuitous, nonprogrammed, and non-programmable side product of two independent biosynthetic investigations: on triterpenoids of the hopane series 20 in bacteria (1) and on diterpenes of the ginkgolide 26 and bilobalide series in the higher plant Ginkgo biloba (3). In this pathway (Fig. 2), isoprene units are derived directly from carbohydrate metabolism, with pyruvate 9 and glyceraldehyde 3-phosphate 10 as starting materials and with MEP 12 as the committed intermediate.

Significance of the MEP pathway

The MEP pathway remained overlooked for nearly 40 years. The MEP pathway is present in most eubacteria, including pathogens and opportunistic pathogens, as shown first by biochemical evidence and later by comparison of gene sequences. It is present in all phototrophic organisms, but its presence is restricted to the chloroplasts; the MVA pathway operates in the cytoplasm. It is also found in phylogenetically related nonphotosynthetic taxa, such as the Plasmodium spp., which are the parasites responsible for malaria possessing apicoplasts (plastid-derived organelles).
The MEP pathway revealed a set of seven novel enzymes that correspond to unannotated genes with unprecedented reaction mechanisms, especially the last two enzymes (GCP and LytB). These two enzymes are characterized by a Fe/S cluster, which catalyzes the conversion of a 1,2-diol derivative into an olefin and the conversion of an allylic alcohol into an olefin.

The MEP Pathway

Several aspects of the MEP pathway reaction sequence are unexpected, including the conversion of two highly oxidized carbohydrate derivatives, pyruvate 9 and glyceraldehyde phosphate 10 (GAP), into two monounsaturated alcohol diphosphates, IPP 11 and DMAPP 12. 1-deoxy-D-xylulose 5-phosphate (DXP), the first C5 Intermediate and DXP Synthase

The first step of the MEP pathway is to form DXP 11, by the DXP synthase (DXS) from pyruvate 9 and GAP 10 (2). This reaction could be postulated from the knowledge of the origin of the carbon atoms in isoprene units. Incorporation of deuterium-labeled isotopomers into the prenyl chain of ubiquinone 22 and menaquinone 23 (Fig. 6) from E. coli provided the first evidence for the involvement of deoxyxylulose (DX) in isoprenoid biosynthesis. This observation was extended to plant isoprenoids in plant systems. Free DX is usually well incorporated into the MEP pathway, with the perulose being phosphorylated by a nonspecific xylulose kinase in bacteria as well as in plants (4). The DXS gene of this enzyme was observed directly, but synthetic methylerythrose phosphate 18 derived from the synthetic ME. Free ME added to the culture medium was later confirmed by the incorporation of DX isotopomers with multiple labeling into the prenyl chains of the quinones 22 and 23 (Fig. 6) in E. coli or phytol 24 from Catharanthus roseus cell cultures (2). This reaction was first written as an acid-catalyzed rearrangement of an α-ketol (Fig. 4a), followed by the NADPH-dependent reduction of the resulting carbonyl derivative. An analogy was made between the corresponding reaction found in the biosynthesis of branched chain amino-acids. A 2-C-methyl-d-erythritol 4-phosphate 12 derivative is the product of this rearrangement. Deuterium-labeled free 2-C-methyl-d-erythritol (ME) was incorporated into the prenyl chain of ubiquinone and menaquinone from wild-type E. coli, but with a much lower yield than DX. This result suggests its correlation with the biosynthesis of isoprene units. In E. coli mutants with a dxs deletion, this deletion was rescued by the addition of synthetic deuterated ME to the culture medium. Incorporation was quantitative, as all isoprene units were derived from the synthetic ME. Free ME added to the culture medium is used only by a few bacteria. In Salmonella enterica, the sorbitol phosphotransferase system is responsible for the phosphorylation and transport of free ME. An analysis of E. coli mutants auxotrophic to ME enables the identification of the DXP reducto-isomerase gene (dxr). As expected, this enzyme catalyzes the rearrangement of DXP into methylenetetrahydrofolate 4-phosphate 18 (Fig. 4) and catalyzes the concomitant NADPH-dependent reduction of the intermediate α-diketone into MEP. The enzyme requires a divalent cation, such as Mg$^{2+}$ or Mn$^{2+}$. The postulated alpha-ketol intermediate 18 has never been observed directly, but synthetic methylenetetrahydrofolate 18 is converted by DXX into DXP in the presence of NADP$^+$ and into MEP in the presence of NADPH. The addition of NADPH
and DXP follows an ordered mechanism with NADPH binding before DXP. The stereochemistry of the reduction step is known. The pro-5 hydrogen of NADPH is transferred onto the re face of the aldehyde intermediate (6). 1-Fluoro DXP is a poor substrate of DXR. Although no definitive conclusion could be drawn from this conversion, a careful kinetic analysis of the conversion of this substrate analog suggests an alternative retro-aldol/aldol reaction for the rearrangement step (Fig. 4(a)) which was confirmed by the secondary isotope effects observed upon incubation of DXP with deuterium labeling at C-3 or C-4 (8).

Fosmidomycin (Fig. 4) and its analogs are strong inhibitors of the MEP pathway, and they inhibit strongly the DXR. According to the X-ray structure of the E. coli DXR, which crystallizes in the presence of fosmidomycin, the antibiotic acts as an analog of the DXP substrate rather than as an analog of the intermediate methylerythrose phosphate (9).

From methylerythritol phosphate to methylerythritol cyclodiphosphate

No obvious indications were available to identify the next steps after the formation of MEP (2, 10, 11). The incubation of synthetic trimethyl-labeled MEP with a crude cell-free system from E. coli and with a cocktail of all plausible cofactors and the analysis of the radioactive metabolites suggested the formation of a ME/nucleotide adduct, which could not be characterized fully.

A systematic search in gene libraries for a gene encoding an enzyme that uses a polyol phosphate and a nucleotide triphosphate as substrates led to the acsi gene from H. hemophilus influenzae encoding an enzyme coupling ribitol 5-phosphate with cytidine 5’-triphosphate (CTP) and yielding the 5’-diphosphocytidyl adduct of ribitol. This gene presented high homologies with the unannotated ygbP gene from E. coli. The corresponding protein was tested in the presence of MEP 12 and nucleotide triphosphates, with the most efficient one being CTP, and catalyzed the formation of the 4-diphosphocytidyl adduct of ME (Fig. 2). In E. coli, the ygbP gene was accompanied by two additional genes, ycbB and ygbR, which together constitute a small cluster. In fact, these two genes coded for the next two enzymes of the pathway. The YcbR protein catalyzed the ATP-dependent phosphorylation of the tertiary hydroxy group of 4-diphosphocytidyl MEP 13 yielding 4-diphosphocytidyl MEP 2-phosphate 14. The YgbP protein converted the latter intermediate into ME cyclodiphosphate 15 with the elimination of cytidine 5'-monophosphate (Fig. 2). Interestingly, in many bacteria, the ygbP and the ygbR genes yield a fusion protein, which corresponds to a bifunctional enzyme catalyzing the two steps, which are performed by two distinct enzymes in E. coli or in plants (12).

From ME Cyclodiphosphate to IPP and DMAPP

In the conversion of MEP 12 into ME cyclodiphosphate 15, the oxidation state of the C5 branched carbon skeleton does not change. The main modification is the introduction of a good leaving group at C-2 on the ME skeleton, which implies that the additional steps include elimination and reduction to afford the unsaturated alcohol diphosphates, IPP and DMAPP, from a tetrol derivative, which is unprecedented in enzyme reactions. Only two additional genes, gcpE and lytB, accompanied all above-mentioned genes of the MEP pathway, making them candidates for the catalysis of the final steps. The amino acid sequence of the two proteins encoded by the gcpE and lytB genes reveals the signature of [Fe₄S₄] proteins, i.e., with three conserved cysteines involved in the fixation of the cluster in the active sites (2, 13). Most of these proteins are oxygen...
Methylerythritol Phosphate Pathway for the Formation of Isoprene Units

\[
\begin{align*}
\text{Enolpyruvyl Shikimate} & \quad \text{(a)} \\
\text{pathway} & \quad \text{(b)} \\
\end{align*}
\]

**Figure 5** Origin of carbon atoms from glucose in the isoprene units of (a) the hopanoids of *Zymomonas mobilis*, (b) the diterpenoids of *Ginkgo biloba*.

**Figure 6** Isoprenoids from bacteria (bacteriohopanepolyol, 20; pentenolactones, 21; ubiquinone, 22; menaquinone, 23) and from higher plants (phytol, 24; β-carotene, 25; ginkgolide, 26; sterols, 27).

Sensitive, and enzyme tests must be performed under an inert argon atmosphere. GcpE converts ME cyclodiphosphate 15 into 4-hydroxy DMAPP 16, which is in turn converted into IPP 7 and DMAPP 8 in an estimated 5:1 ratio (**Fig. 2**). The first reaction corresponds with the conversion of a diol derivative into an olefin; the [Fe₄S₄] cluster in a reduced state acts as an electron donor and probably acts as a Lewis acid for the elimination of the C-3 hydroxy group of ME cyclodiphosphate. The second reaction is the conversion of an allylic alcohol into an olefin. Again the reduced [Fe₄S₄] cluster of LytB is involved in a reductive one-electron transfer process and acts as a Lewis acid for the elimination of the C-4 hydroxy group of 4-hydroxy DMAPP 16 (**Fig. 2**).

Both GcpE and LytB must be associated with a reducing system to convert the oxidized [Fe₄S₄]²⁺ cluster into the active reduced [Fe₄S₄]⁺⁺ form (**Fig. 3**), which can be performed with the isolated enzymes from E. coli either by the biologic system flavodoxin/flavodoxin reductase/NADPH, by the semiquinone radical of 5-deazaflavin, or by dithionite. In cyanobacteria, and in plant chloroplasts where flavodoxin is absent, this reducing
role can be performed by ferredoxin as shown for the GcpE enzyme, the electron flow coming directly from photosynthesis in the light or from catabolic metabolism in the dark (14).

Elucidation of the MEP Pathway: Key Experiments

The first results from studies the MEP pathway were obtained from in vivo incorporation experiments of 13C labeled glucose isotopomers with bacteria or ginkgo embryos, which determined the origin of the carbon atoms of the isoprene units and which resulted in a reasonable biogenetic scheme. This step was followed by biochemical experiments that were designed to validate the aforementioned scheme and were led to the identification of enzymes, deoxyxylulose phosphate (DXP) and MEP, and the identification of enzymes, deoxyxylulose phosphate synthase (DXS) and DXP reducto-isomerase (DXR). Later steps were identified by using a combination of molecular biology and biochemical methods, including enzyme tests as mentioned in the previous paragraph. Most experiments were cited for the elucidation of the pathway. Three sets of key experiments are described in detail.

Incorporation of 13C-labeled glucose into the hopanoids of Zymomonas mobilis and into the diterpenoids of ginkgo embryos: the origin of carbon atoms in isoprene units

Incorporation studies of 13C-labeled precursors require extensive nuclear magnetic resonance (NMR) measurements. The signals of 13C-NMR-spectra of the analyzed metabolites must be assigned fully. The 13C-NMR spectra of the reference compound of natural abundance (1.1%) and of the labeled metabolite must be recorded in the same conditions. A comparison between the relative signal intensities of the two spectra indicates that the carbon atom is labeled and indicates the magnitude of the isotope enrichment.

The incorporation of 13C-labeled acetate into bacterial hopanoids (Fig. 6) revealed unambiguously a labeling pattern that was not compatible with the MVA pathway. The incorporation of 13C-labeled glucose isotopomers into the isoprene units of the hopanoids from the bacterium Zymomonas mobilis determined the origin of all the carbon atoms of the isoprene units (Fig. 5a, 1, 2). This bacterium uses glucose as a carbon and energy source via the Entner-Doudoroff pathway. It has a tricarboxylic acid cycle and does not convert pyruvate into glyceraldehyde phosphate. These metabolic peculiarities facilitated a retrobiosynthetic analysis of the labeling pattern, which suggested that the C5 isoprene skeleton was formed from a C2 moiety derived from pyruvate decarboxylation and from a C3 subunit derived from a triose phosphate. The incubation of doubly labeled (4,5-13C2)glucose showed that the coupling of the labeled carbon atoms was preserved in the isoprene units, which indicates that they were introduced together via a single precursor and suggests that an intramolecular rearrangement performed the insertion of the C2 pyruvate-derived subunit between the carbon atoms derived from C-4 and C-5 of glucose. This key experiment excluded the MVA pathway for the formation of the bacterial isoprene units and was the signature of an alternative biosynthetic route.

Similar experiments performed on ginkgo embryos also revealed an unexpected labeling pattern of the isoprene units of the diterpenoid skeletons. In this case, glucose is metabolized via glycolysis and a retrobiosynthetic analysis was in accordance with the formation of isoprene units from pyruvate and from a triose phosphate derivative (Fig. 5b) (3).

Incorporation of 2H-labeled deoxysylulose and methlyerythritol into terpenoids from bacteria and from plant plastids

The incorporation of 13C-labeled glucose isotopomers showed that a pyruvate and a triose phosphate derivative were the precursors of isoprene units in the alternative route. The triose phosphate derivative was identified as glyceraldehyde phosphate using E. coli mutants; each lacks a single enzyme of the triose phosphate metabolism (allowing the interconversion of pyruvate and glycerol) and grows either in the presence of no-labeled pyruvate and 13C-labeled glycerol or in the presence of a 13C-labeled pyruvate and nonlabeled glycerol. These initial results led to the identification of the first two C2 intermediates of the pathway.

1-D-Deoxy-D-xylulose is a known natural product. It was first isolated from the fermentation broth of a Streptomyces and later shown to be a precursor of pyridoxol. Its structure can be deduced biogenetically from pyruvate and glyceraldehyde phosphate (Figs. 2 and 3). Two 2H-labeled DX isotopomers were synthesized chemically. They were incorporated efficiently into the prenyl chains of ubiquinone and menaquinone by wild-type E. coli, indicating that a DX derivative is an isoprenoid precursor (2, 3).

Free methlyerythritol is a widespread poliol in plants. The formation of the isoprene skeleton via the alternative route involves an intramolecular rearrangement. The branched carbon skeleton of ME can be deduced from the rearrangement of the straight chain DX (Fig. 4). Deuterium-labeled ME isotopomers were synthesized chemically and were incorporated into the prenyl chains of the E. coli quinones (2).

E. coli constructs using MVA for the identification of MEP pathway genes

Many isoprenoids are essential metabolites in living cells. Inhibition of their biosynthesis is lethal, which implies that any deletion or any major mutation of a gene of the MEP pathway will be lethal and must be rescued by another source of precursor for IPP and DMAPP. In E. coli, this can be performed by introducing the genes of the MVA pathway, which is absent in this bacterium (15).

For instance, the deletion of the ddx gene, which encodes the enzyme catalyzing the conversion of DXP into MEP, is lethal. This deletion can be rescued two ways by adding
synthetic ME, which is phosphorylated in vivo and enters the MEP pathway, or by introducing the genes of enzymes which allow the conversion of MVA into IPP. For this reason, these genes must be introduced in the construct, encoding, respectively, mevalonate kinase, phosphomevalonate kinase, and dihydrophosphomevalonate decarboxylase (Fig. 1).

This approach has been used to check that ggE and lyB are involved in the MEP pathway. These genes accompanied regularly the other known genes of the pathway in the completely sequenced bacterial genomes. The deletion of either of these genes was lethal, which indicates that they are essential but was rescued by the insertion of the three above-mentioned genes of the MVA pathway, indicating that they are involved in the biosynthesis of IPP and DMAPP (2).

Future Developments
Inhibition of the MEP pathway: toward novel antibacterial and antiparasitic drugs
The MEP pathway is the only pathway involved in the biosynthesis of essential isoprenoids in pathogenic bacteria and in parasites, and it is absent in animals and in humans. Therefore, any enzyme of this pathway is a potential target for a novel type of antimicrobial drugs (16, 17). This concept has been validated by the mode of action of fosmidomycin, a natural antibiotic that inhibits the second step of the MEP pathway catalyzed by the deoxyxylulose phosphate reducto-isomerase (DXR).

Overexpression of the MEP pathway: toward improved production of plant terpenoids of economic value
The MEP pathway is the starting material of major plant terpenoids of economic value, e.g., monoterpene s from essential oils, diterpenoids from pathogenic bacteria and in parasites, and it is absent in animals and in humans. Therefore, any enzyme of this pathway is a potential target for a novel type of antimicrobial drugs (16, 17). This concept has been validated by the mode of action of fosmidomycin, a natural antibiotic that inhibits the second step of the MEP pathway catalyzed by the deoxyxylulose phosphate reducto-isomerase (DXR).

Cross-talk between the MVA and the MEP pathway: A Novel Aspect for the Regulation of Terpenoid Biosynthesis in Plants
Finally, the dichotomy between the cytoplasmic mevalonate pathway and the plastidial MEP pathway is not strict. An exchange of metabolites occurs between the two compartments at the level of C5, C10, and C15 prenyl diphosphates. The inhibition of one pathway (e.g., by mevinolin for the MVA pathway or by fosmidomycin for the MEP pathway) can be complemented by the other route either partially, as in most tested plant systems, or even completely, in the case of the tobacco Bright yellow-2 cell cultures (20). In addition, a regulation of the two pathways by the nuclear cycle is suggested: the MEP pathway is activated in the presence of light and is activated by photosynthesis (21). Cross-talk and light regulation lead to new findings in the regulation of isoprenoid biosynthesis in plants.

References
Methylerythritol Phosphate Pathway for the Formation of Isoprene Units


Further Reading


See Also

Antibacterial Drugs, Design of
Iron-Sulfur World
Isoprenoids
Steroid and Triterpene Biosynthesis
Terpenes, Biosynthesis of Terpenoids in Plants;
Natural Products: An Overview
Jon Clardy and Michael A. Fischbach, Harvard Medical School, Massachusetts
doi: 10.1002/9780470048672.wecb377

Natural products, the remarkable collection of small molecules made by living organisms in an idiosyncratic manner, have made important contributions to organic chemistry, biology, and medicine. Although natural products contain a mother lode of molecular diversity, all of their structures reflect an evolutionary history and a biosynthetic production from a limited set of starting materials and reactions. The structures of natural products differ quantifiably from the small molecules usually found in the screening libraries used in high-throughput assays. Deeper structural analysis often reveals that natural products bind important biological targets with an exquisite degree of shape and charge complementarity. In addition, natural products often employ a variety of strategies to disguise reactive functional groups until they are needed. Biosynthetic pathways for natural products resemble those used to construct more common biomolecules; the main themes are reactions that couple monomers into larger molecules and the lavish use of redox chemistry not just to create skeletons but also to cross-link them and decorate them with oxygen-based functionality. Although the historic contributions of natural products to chemistry, biology, and medicine are clear, their current use, especially in drug discovery, seems less clear. However, efforts to explore new areas of biological diversity, mine genomes, and study the natural roles of natural products are rejuvenating natural products research and illuminating its future.

What are Natural Products?
"Natural product" is such a familiar term to most chemists and biologists that it is easy to forget that it describes a group of compounds with breathtaking molecular diversity and biological activity. Two natural products easily can seem to have little, if anything, in common, so what distinguishes natural products from other small molecules? In the past, describing a natural product was fairly straightforward: It was an organic molecule obtained from a living (or formerly living) organism that had a limited distribution in nature. Biosynthetic production and sporadic occurrence were the key defining features. But as we have learned more about natural products, this seemingly clear definition has become more ambiguous as illustrated by three molecules: tryptophan, psilocin, and serotonin (Fig. 1).

All prokaryotes and some eukaryotes biosynthesize tryptophan, and all organisms incorporate tryptophan into their proteins. Its widespread distribution excludes it from being a natural product, and it is usually called a primary metabolite. Psilocin, which is biosynthesized from tryptophan, is best known as the hallucinogenic principle of "magic mushrooms," and its limited distribution in a group of small mushrooms, not to mention its role in pre-Columbian religious rituals in Mexico, make it a...
classic natural product (2). Serotonin, which is also biosynthesized from tryptophan, is a well-known human neurotransmitter, and its regulation is an important therapeutic target for depression and other conditions. Serotonin and serotonin bind some of the same receptors in the human brain, although to very different effect (3). But many organisms, including bacteria and plants, make serotonin for various purposes, and its classification depends on context. In humans, it is a neurotransmitter; in barnacle larvae looking for a place to settle, it is a natural product (or settling pheromone); and the difference has more to do with the scientists conducting the study than the molecule being studied (4). Is there a fundamental difference between these three molecules that have very similar structures, share large parts of their biosynthetic pathways, and bind common cellular targets? In addition to the ambiguities introduced by the limited-distribution test, our increased understanding of natural products has made even "coming from a living organism" ambiguous. Is the product of a biosynthetic pathway reconstituted in a test tube rather than a living organism a natural product? If a pathway found in nature were altered in some way, say by introducing a different enzyme for one step, then does it still produce a natural product? If enzymes from many different pathways are combined, does that new pathway produce a natural product? It seems likely that in the future, the term "natural product" will embrace any small molecule produced by a (genetically encoded) biosynthetic pathway whether the pathway is common or rare in nature or even created in the laboratory.

Even the usual classifications of natural products by their producing organism—plant, microbial, and marine natural products are the major divisions—are breaking down under genetic infection and other conditions. Psilocin and serotonin bind some of the same receptors in the human brain, although to very different effect (3). But many organisms, including bacteria and plants, make serotonin for various purposes, and its classification depends on context. In humans, it is a neurotransmitter; in barnacle larvae looking for a place to settle, it is a natural product (or settling pheromone); and the difference has more to do with the scientists conducting the study than the molecule being studied (4). Is there a fundamental difference between these three molecules that have very similar structures, share large parts of their biosynthetic pathways, and bind common cellular targets? In addition to the ambiguities introduced by the limited-distribution test, our increased understanding of natural products has made even "coming from a living organism" ambiguous. Is the product of a biosynthetic pathway reconstituted in a test tube rather than a living organism a natural product? If a pathway found in nature were altered in some way, say by introducing a different enzyme for one step, then does it still produce a natural product? If enzymes from many different pathways are combined, then does this new pathway produce a natural product? It seems likely that in the future, the term "natural product" will embrace any small molecule produced by a (genetically encoded) biosynthetic pathway whether the pathway is common or rare in nature or even created in the laboratory.

Even the usual classifications of natural products by their producing organism—plant, microbial, and marine natural products are the major divisions—are breaking down under genetic infection and other conditions. Psilocin and serotonin bind some of the same receptors in the human brain, although to very different effect (3). But many organisms, including bacteria and plants, make serotonin for various purposes, and its classification depends on context. In humans, it is a neurotransmitter; in barnacle larvae looking for a place to settle, it is a natural product (or settling pheromone); and the difference has more to do with the scientists conducting the study than the molecule being studied (4). Is there a fundamental difference between these three molecules that have very similar structures, share large parts of their biosynthetic pathways, and bind common cellular targets? In addition to the ambiguities introduced by the limited-distribution test, our increased understanding of natural products has made even "coming from a living organism" ambiguous. Is the product of a biosynthetic pathway reconstituted in a test tube rather than a living organism a natural product? If a pathway found in nature were altered in some way, say by introducing a different enzyme for one step, then does it still produce a natural product? If enzymes from many different pathways are combined, then does this new pathway produce a natural product? It seems likely that in the future, the term "natural product" will embrace any small molecule produced by a (genetically encoded) biosynthetic pathway whether the pathway is common or rare in nature or even created in the laboratory.

The Structures of Natural Products

At the most fundamental level, natural products differ from other small molecules because of their evolutionary history. Natural products—or, more accurately, the genes that produce them—have undergone countless rounds of alteration, selection, and amplification that have honed their biological activities to better fulfill evolutionary selection criteria. This evolutionary recycling of their structural features has created what one recent commentator described as "the extraordinary advantages of small molecule natural products as sources of agents, which interject themselves in a helpful way in various physiological processes" (9).

Synthetic molecules, especially those prepared by medicinal chemists, also have selection criteria, but they usually involve synthetic accessibility, drug-like properties, or elaborations on a successful model. Whereas natural products generate their molecular diversity from a relatively limited set of starting materials and reaction types, as will be discussed shortly, synthetic molecules can stem from much larger sets of starting materials and reactions. Several studies have compared natural products and synthetic molecules and concluded that natural products contain relatively more carbon, hydrogen, and oxygen and less nitrogen and other elements than synthetics (10). In contemporary synthetic molecules, drug-like properties (how they are absorbed, distributed, metabolized, and eliminated) are enhanced by having molecular weight cutoffs (500 daltons is popular), limited hydrogen bonding potential, and solubility criteria (11). Most natural products actually conform to these rules, but notable exceptions of clinically useful compounds that greatly exceed the molecular weight guideline and have greater water solubility than recommended can be found (Fig. 2).

For any small molecule to possess sufficient binding energy to be a potent ligand (a dissociation constant in the 10^-9 to 10^-4 M range), its structural features must organize the molecule's electrostatic, hydrogen-bonding, and hydrophobic interactions to complement those of its binding partner. Successful natural and synthetic molecules achieve these conformationally restricted structures in rather different ways. Natural products often achieve this rigidification by forming macrocycles, having polycyclic structures, and minimizing gauche interactions (Fig. 2). Synthetic molecules favor possessing aromatic rings and minimizing the number of rotatable bonds. Natural products generally have more sp3-hybridized carbons and sterogenic centers (Fig. 2). Morphin, a molecule with an ancient if occasionally troubled involvement with humans, illustrates the exquisite level of three-dimensional conformational control to project hydrophobic surfaces and hydrogen bonding partners in all directions that can be achieved by natural products. In more elaborate comparisons with many molecular parameters, natural products and synthetics occupy rather different regions of “chemical space.”

A less appreciated aspect of natural products is their extraordinary repertoire of strategies employed to carry out their biological roles. They are, for example, masters of disguising reactive functionality. Sometimes the disguise can be as simple as converting a reactive aldehyde into an intramolecular hemiacetal, but it can also involve much more elaborate schemes. Some spectacular examples of this “wolf in sheep’s clothing” approach can be found in a structurally diverse family of natural products called DNA-targeting agents. These molecules must move through a biological milieu—including the cell and nucleus—and clear membranes—with many molecules that are much more reactive than DNA—and then degrade DNA beyond easy repair. Calicheamicin initiates a complex set of reactions to carry out this destructive mission by unmasking a thiol, which undergoes an intramolecular conjugate addition, which sets up the aromatization of the enediyne fragment, which generates an aromatic diradical, which damages both DNA strands (Fig. 2). Although calicheamicin’s molecular gymnastics might seem unduly complicated, calicheamicin linked to an antibody that directs it to cancer cells has shown therapeutic efficacy as an anticancer agent in several clinical trials (13). Not all
DNA-targeting agents rely on masking their reactivity. CC-1065 uses shape complementarity with specific DNA sequences to enhance its reactivity to carry out a disabling alkylation (Fig. 2) (14). It accelerates the reaction in the appropriate biological context through the exquisite alignment of the reacting partners.

**Biosynthetic Pathways**

Biochemistry has revealed nature's two main themes in carrying out the chemistry of life: 1) coupling reactions that join similar (or identical) small molecules into the larger molecules needed to store information, form cellular structures, and carry out catalytic functions and 2) redox reactions such as the largely oxidative reactions our bodies use to get energy from small molecules and the largely reductive reactions our bodies use to build up small molecules. Not surprisingly, the biosynthesis of natural products reprises these two reaction types. Although many natural product biosynthetic pathways employ only one of these two themes—either coupling or redox chemistry—some use both. These two-stage pathways typically begin with the coupling-mediated assembly of a multimer followed by the redox-mediated maturation of the initially formed product. In these hybrid pathways, redox-mediated maturation reactions decorate the core structure assembled by the coupling pathway. Nature also uses redox reactions to build core structures, an approach rarely employed for the same purpose in laboratory synthesis. These general observations are expanded in the examples below.

**Coupling reactions**

A coupling reaction, broadly defined, joins two fragments with the accompanying loss of a small molecule (or ion). In one well-known coupling reaction, the nucleophilic amino group of one amino acid reacts with the electrophilic carboxyl carbon of a second amino acid to form an amide bond with the loss of water (Fig. 3). Proteins, RNA/DNA polymers, lipids, and polysaccharides are all produced by different monomer-joining coupling reactions. Major classes of natural products—terpenes, polysaccharides, nonribosomal peptides, and polyketides—are also produced by coupling reactions, most of which feature a phosphate in some form as the ejected fragment. The details of the coupling reaction for these different classes vary depending on the subunits to be joined and the nature of the ejected fragment. The ejected fragment is a good leaving group, meaning that the bonds being formed are stronger than those being broken, and its release provides an entropy compensation for joining the two substrates.

Terpenes are formed by coupling double-bond isomers of the same monomer, isopentenyl pyrophosphate (IPP), with loss of a pyrophosphate ion (15, 16). Three such coupling reactions join four IPPs, each with five carbon atoms, to make geranylgeranyl-pyrophosphate, the C20-precursor of diterpenes (Fig. 3). Linear terpene precursors, which differ only in the number of IPP units joined, can be assembled by a few polyprenyl synthase enzymes, as the monomers are identical and the only variable is how many monomers to link together. In subsequent steps, these linearpolyenes undergo cation-induced cyclizations to create the starting frameworks for terpenes and their derivatives.

The coupling reaction used to build polysaccharides, such as the sugar fragment in vancomycin (Fig. 2), from hexose and pentose monomers is conceptually similar to the terpene pathway, except the leaving group is a nucleoside diphosphate (a ribosylpyrophosphate) rather than inorganic pyrophosphate, and the sugars to be joined can have considerable variety. This variability in the substrates requires a larger set of coupling enzymes, or glycosyltransferases, each of which selects specific glycosyl donor and acceptor cosubstrates (17). Genomic analyses highlight the consequences of this increased specificity.
Streptomyces avermitilis, a bacterium with a completely sequenced genome (18, 19), has three times as many predicted glycosyltransferases than predicted polyphenyl synthases.

The amino acid monomers of nonribosomal peptides (20) like vancomycin (21) (Fig. 2) are also linked using a phosphate-displacing coupling reaction with an adenosine monophosphate (AMP) leaving group. Unlike terpene and polysaccharide biosynthesis, the overall coupling process has two distinct reactions connected by a stable intermediate (Fig. 3). In the first reaction, the aminoacyl-adenylate is transferred to a "carrier protein" to which it is tethered through an aminoacyl thioester bond, and in the second reaction, the amine of an aminoacyl co-substrate reacts with the thioester to form a peptide bond (22). Both reactions have good leaving groups—either AMP or a thiolate. This two-step process resembles the ribosomal assembly of peptides with the amino acid bound to a carrier protein by a thioester bond in the nonribosomal case or a tRNA by an ester bond in the ribosomal case. Tethering substrates to intermediate carrier protein domains may be required for proper directionality of chain assembly, a role that is fulfilled by elongation factors during ribosomal peptide synthesis.

Nonribosomal peptide synthetases (NRPSs) function similarly to the ribosome in a second way: They act processively to build a peptide chain. In both cases, the amine group of the incoming aminoacyl monomer attacks the (thio)ester bond that links the nascent peptide to its carrier, which translocates the chain while elongating it by one monomer (22).

The polyketides (23) form the only major class of natural products that has monomer couplings that do not involve the loss of phosphate; like the nonribosomal peptides, the monomer couplings use a thioester intermediate. These molecules, which have biosynthetic pathways much like those of fatty acids, are assembled from the alpha-carboxylated two- and three-carbon metabolites malonyl-CoA and methylmalonyl-CoA (22). In these monomers, the alpha-carboxyl group is a caged form of carbon dioxide, and decarboxylation yields a thioester enolate...
that attacks a downstream thiolster to form a carbon-carbon bond (Fig. 3). In this coupling reaction, decarboxylation not only provides the attacking nucleophile, but also it serves as the driving force for unidirectional assembly. Although the coupling reactions for polyketides and nonribosomal peptides differ, the enzymes that assemble them function by similar processive logic: The incoming monomer attacks the nascent chain, which translocates it while elongating it by one monomer. Unlike many of today’s software designers, nature prioritizes compatibility across platforms, which leads to numerous molecules that are formed by hybrid pathways, just like the lipoproteins, glycoproteins, and glycolipids from primary cellular pathways, hybrid natural products originate from variants of these chemically compatible coupling reactions in which a donor monomer of one type gets linked to an acceptor monomer of a different type. For example, vancomycin (Fig. 3) is a hybrid between a nonribosomal peptide and a polysaccharide in which the key bond is formed by a coupling enzyme that links the peptide residue 4-hydroxyphenylglycine to the hexose monomer UDP-glucose (24).

**Redox reactions**

Redox chemistry occurs in natural product biosynthesis in at least three ways: building core structures, cross-linking core structures, and adding functionality to core structures. Just as the coupling reactions in natural product pathways resemble those from macromolecular synthesis, redox transformations in natural product pathways have counterparts in primary metabolism, which indicates that the genes in natural product pathways and the genes in primary metabolic pathways have common ancestors. Examples to illustrate the distinctions between building cores, cross-linking cores, and decorating cores are discussed below.

Podophyllotoxin (Fig. 1) was the basis for a successful anticancer drug because of its ability to prevent tubulin polymerization, although etoposide, the successful drug developed from it, actually works by inhibiting topoisomerase II and thereby preventing DNA replication (23). Podophyllotoxin’s skeleton is assembled through a very common redox reaction, the one-electron oxidation of coniferyl alcohol and a coupling of two such radicals (26, 27) (Fig. 4). This basic structure then undergoes a reduction, aromatic and aliphatic hydroxylations, and modifications of its aromatic substituents to give podophyllotoxin (27). The initial one-electron oxidation of a phenolate anion is also used to build the skeleton of morphine (Figs. 2 and 4) (28).

An example of oxidative cross-linking takes place during vancomycin biosynthesis; three cytochrome P450 enzymes—which use the same heme cofactor as the cytochromes from the primary metabolic electron transport chain—cross-link aromatic rings in its scaffold to form its rigid Cup-like shape (29, 30). Cytochromes P450 that decorate cores with oxygen-based functionality are commonly found in biosynthetic pathways. For example, these enzymes hydroxylate the reduced polycyclic scaffold of the diterpene taxol and install functional groups that are required for target binding and increase its hydrophilicity (31).

**Modifying naturally occurring pathways**

The most frequently encountered natural product biosynthetic pathways—those producing polyketides, nonribosomal peptides, polysaccharides, and terpenes—are all highly “evolvable” because they have a biosynthetic logic that is highly susceptible to small changes and rearrangements. François Jacob, one of the fathers of molecular biology, described these changes as “molecular tinkering” to emphasize the way nature cobbles together a workable solution through small alterations and rearrangements of previously existing materials (32). The “evolvability” of natural product biosynthetic pathways takes on two forms: modularity, in the case of polyketides, nonribosomal peptides, and polysaccharides, and plasticity, in the case of terpenoids. Researchers have taken advantage of these same features to engineer new pathways that produce new products. As mentioned above, polyketides are assembled processively by logic that resembles ribosomal peptide synthesis. But a fundamental difference is found: Whereas the ribosome is a “factory on wheels” that translocates along an mRNA template, polyketide synthases (PKSs) are multidomain enzymatic assembly lines that function both as the template and the synthesizer. Like assembly lines in factories, PKSs are modular, and each module (a set of domains) adds a single monomer to the growing polyketide chain. Researchers have exploited this modularity to alter or replace modules in the PKS that produces the backbone of erythromycin and have generated a library of “programmed” derivatives that differ at specific positions (33). A complementary approach, in which a library of artificial two-module PKSs was constructed, led to the production of new polyketide fragments and set the stage for future efforts to construct “synthetic” polyketide pathways from a standard set of biosynthetic building blocks (34). The former approach resembles the mutational component of the evolutionary approach used by nature, and efforts to introduce the component of selection are likely to improve the future prospects of “synthetic” pathway construction.

The coupling reaction described above for terpene pathways results in a linear C9n chain, whereas terpenoid natural products have a staggering array of shapes and topologies. The cyclic skeletons characteristic of terpenes are formed from these methylated precursors by enzymes called terpene cyclases (35). Terpene pathways are plastic in two ways: Relatively few mutations in these cyclase enzymes can lead to skeletons of different topology, and the redox enzymes that add oxygen-based functionality to these skeletons are often promiscuous. By exploiting the plasticity of terpene cyclases, seven mutants of the gamma-humulene synthase have been designed in which 1-5 amino acid substitutions led to the production of seven new products with a wide range of topologies (36).

**The Future of Natural Products Chemistry**

Although the historical contributions of natural products to the development of chemistry, biology, and medicine are clear, their role in the future development of these fields is cloudy,
Natural products chemistry has, like most things, evolved in fits and starts—periods of rapid advances followed by consolidation. The periods of rapid advance have always been associated with the opening up of new ecological niches to natural products research. In recent times, the opening up of the marine environment to natural products research spurred a few decades of truly heroic natural product discoveries. Marine natural products such as halichondrin B and ecteinascidin 743, analogs of which are in anticancer trials, were identified (37). Today, the most impressive remaining frontiers are largely microbial, and exploring the marine environment with suitably modified microbiological techniques has led to an increasing flow of molecules that are both novel and potentially useful. One example is the proteasome inhibitor salinosporamide, which was isolated from the marine actinomycete Salinispora tropica CNB-440 (38–40). The recently sequenced genome of this organism reveals a wealth of new molecules, which highlights the potential of marine microenvironments as a source of molecular diversity and showcases the biosynthetic prowess of this new actinomycete genus (41). Likewise, the discoveries of new molecules like jamicamidamide (42) and curacin (43) demonstrate the promise of marine cyanobacteria as a source of natural products (44). Some of the most impressive molecular diversity may not be found in the ocean but in the soil beneath our feet. Myxobacteria, the source of the tubulin-binding anticancer drug epothilone (45), showcases the biosynthetic prowess of this new actinomycete genus (41). Likewise, the discoveries of new molecules like jamicamidamide (42) and curacin (43) demonstrate the promise of marine cyanobacteria as a source of natural products (44). Some of the most impressive molecular diversity may not be found in the ocean but in the soil beneath our feet. Myxobacteria, the source of the tubulin-binding anticancer drug epothilone (45),

**New geographical and phylogenetic niches**

On the negative side, most large pharmaceutical companies have either greatly reduced or completely oxidized their natural product drug discovery programs; but on the positive side, researchers are finding new ecological niches containing dramatically new natural products, developing methods to mine genomic sequences for new natural products, and discovering new ways to connect natural products with ecology (Fig. 5).

**Figure 5** Biosynthetic pathways. (a) In the terpenoid coupling reaction, isomers of isopentenyl pyrophosphate are joined with the loss of pyrophosphate, leading to a linear intermediate that is cyclized to a terpene skeleton, as shown for the diterpene taxol. (b) In the polysaccharide coupling reaction, the loss of carbon dioxide from a two or three-carbon monomer yields a thioester enolate that attacks a carrier protein-tethered intermediate, forming a carbon-carbon bond as shown for the polyketone precursor of enterocin. (c) In the first step of the nonribosomal peptide coupling reaction, an aminoacyl adenylate is transferred to a carrier protein or "thiolation" leading to a linear intermediate that is cyclized to a terpenoid skeleton, as shown for the diterpene taxol. (d) In the polyketide coupling reaction, the loss of carbon dioxide from a two or three-carbon monomer yields a thioester enolate that attacks a carrier protein-tethered intermediate, forming a carbon-carbon bond as shown for the polyketone precursor of enterocin.

**Genome mining**

The ability to connect natural products to the genes that produce them has had two important consequences. The first consequence was stimulating genetic and biochemical studies into natural product biosynthesis, which increased our knowledge of the inner workings of these pathways dramatically and set the stage for the modification of existing pathways and the construction of new pathways.

On the negative side, most large pharmaceutical companies have either greatly reduced or completely oxidized their natural product drug discovery programs; but on the positive side, researchers are finding new ecological niches containing dramatically new natural products, developing methods to mine genomic sequences for new natural products, and discovering new ways to connect natural products with ecology (Fig. 5).

**New geographical and phylogenetic niches**

Natural products chemistry has, like most things, evolved in fits and starts—periods of rapid advances followed by consolidation. The periods of rapid advance have always been associated with the opening up of new ecological niches to natural products research. In recent times, the opening up of the marine environment to natural products research spurred a few decades of truly heroic natural product discoveries. Marine natural products such as halichondrin B and ecteinascidin 743, analogs of which are in anticancer trials, were identified (37). Today, the most impressive remaining frontiers are largely microbial, and exploring the marine environment with suitably modified microbiological techniques has led to an increasing flow of molecules that are both novel and potentially useful. One example is the proteasome inhibitor salinosporamide, which was isolated from the marine actinomycete Salinispora tropica CNB-440 (38–40). The recently sequenced genome of this organism reveals a wealth of new molecules, which highlights the potential of marine microenvironments as a source of molecular diversity and showcases the biosynthetic prowess of this new actinomycete genus (41). Likewise, the discoveries of new molecules like jamicamidamide (42) and curacin (43) demonstrate the promise of marine cyanobacteria as a source of natural products (44). Some of the most impressive molecular diversity may not be found in the ocean but in the soil beneath our feet. Myxobacteria, the source of the tubulin-binding anticancer drug epothilone (45), showcases the biosynthetic prowess of this new actinomycete genus (41). Likewise, the discoveries of new molecules like jamicamidamide (42) and curacin (43) demonstrate the promise of marine cyanobacteria as a source of natural products (44). Some of the most impressive molecular diversity may not be found in the ocean but in the soil beneath our feet. Myxobacteria, the source of the tubulin-binding anticancer drug epothilone (45), are proving to be one of the richest sources of new natural products (46, 47).

**Genome mining**

The ability to connect natural products to the genes that produce them has had two important consequences. The first consequence was stimulating genetic and biochemical studies into natural product biosynthesis, which increased our knowledge of the inner workings of these pathways dramatically and set the stage for the modification of existing pathways and the construction of new pathways.
The second consequence was to create a new approach to natural product discovery based on bioinformatics. Instead of screening culture extracts to identify hits in an assay, a function-based approach, a research group scanned the genome of a bacterial strain to identify loci that encode natural product biosynthetic genes, a sequence-based search, and then isolated a new natural product from one of these loci, ECO-02301 (48). This effort was notable not only because the new molecule was identified by genomics rather than by screening but also because bioinformatics played an important role in accelerating the process of structure elucidation. Given the rapidly decreasing cost of genome sequencing and the increasingly powerful ability of bioinformatic analysis to predict natural product structure from gene sequence, genomics and bioinformatics are likely to play a more prominent role in future natural product discovery efforts.

**Clues from ecology**

The potential of natural products to serve as human medicines has always been the major driver of natural products research, and the relentless emphasis on their roles in treating human diseases has had the effect of largely separating natural products from the biological contexts in which they evolved and functioned. We know much more about the unnatural uses of natural products than we do about their intended uses. For example, the number of articles retrieved by searching for “taxol and cancer” (∼9500) greatly exceeds the number retrieved by searching for “taxol and ecology” (three, two of which deal with the possible extinction of the taxol producer to make anticancer medications). Chemical ecology, the subfield of chemical biology that studies molecular interactions between organisms, places natural products in their evolutionary biological context, and natural products chemistry and chemical ecology have extensive overlaps. Both ecology and natural products chemistry are becoming increasingly focused on microbiology. As the great insect ecologist E. O. Wilson wrote in Naturalist, his autobiography, “If I could do it all over again and relive my vision in the twenty-first century, I would be a microbiologist” (49) As ecologists discover new types of microbes and study their roles in interspecies interactions...
interactions, natural products chemists will acquire the producers of fascinating new molecules. One example is the role of actinomycetous bacteria in maintaining an insect–fungus mutualism. Leafcutter ants, New World ants that strip foliage from plants, take the vegetation to large underground colonies where the partially chewed leaves are fed to a fungus (50). In this mutualistic relationship, the fungus eats the leaves, and the ants eat the fungus. Another fungus, which attacks the food fungus, poses a grave threat to the ant colony. A rts protect their food fungus in several ways, including carrying a symbiotic bacteria that produces a selective fungicide, a natural product that selectively kills the antagonistic but not the food fungus (51). Each of the 39 species of leafcutter ants has its own species of food fungus, its own species of antagonistic fungus, and its own symbiotic bacteria to make antifungal compounds with appropriate activity. It is likely that exploring this mutualistic relationship, and possibly even other insect-bacteria mutualisms, will lead to a deeper understanding of the roles of microbes in maintaining symbioses, new natural products and their biosynthetic pathways, and possibly even new compounds with therapeutic efficacy.

References

45. Bollag DM, et al. Epothilones, a new class of microtubule-

stabilizing agents with a taxol-like mechanism of action. Cancer

44. Simmons TL, et al. Biosynthetic origin of natural products isolated

43. Chang Z, et al. Biosynthetic pathway and gene cluster analy-
sis of curacin A, an antitubulin natural product from the trop-

42. Edwards DJ, et al. Structure and biosynthesis of the ja-
amicamides, new mixed polyketide-peptide neurotoxins from

41. Udwary DW, et al. Genome sequencing reveals complex sec-
tary metabolome in the marine actinomycete Salinispora trop-

40. Feling RH, et al. Salinosporamide A: a highly cytotoxic pro-
tease inhibitor from a novel microbial source, a marine bac-


37. McDaniel R, et al. Multiple genetic modifications of the ery-
 thromycin polyketide synthase to produce a library of novel “un-

36. Bohlmann J, Meyer-Gauen G, Croteau R. Plant terpenoid syn-

35. Yoshikuni Y, Ferrin TE, Keasling JD. Designed divergent evolu-

34. Seidel V, et al. Biosynthesis of podophyllotoxin in Linum album

33. McDaniel R, et al. Multiple genetic modifications of the ery-
 tromycin polyketide synthase to produce a library of novel “un-


31. Jennewein S, Croteau R. Taxol: biosynthesis, molecular genetics,
2001;57:13–19.

by OxyB: a cytochrome P450 from the vancomycin producing

by OxyB: a cytochrome P450 from the vancomycin producing

28. Zenk MH, Gerardy R, Stadler R. Phenol oxidative coupling of ben-

27. Seidel V, et al. Biosynthesis of podophyllotoxin in Linum album

26. Buchanan BB, Gruissem W, Jones RL. Biochemistry & Molecular
Biology of Plants. 2000. American Society of Plant Physiologists,
Rockville, MD


24. Buchanan BB, Gruissem W, Jones RL. Biochemistry & Molecular
Biology of Plants. 2000. American Society of Plant Physiologists,
Rockville, MD


by OxyB: a cytochrome P450 from the vancomycin producing

21. Wong ST, et al. Microbial genomics as a guide to drug dis-
covery and structural elucidation. ECO-02301, a novel antifungal


19. Currie CR. A community of ants, fungi, and bacteria: a mul-


Further Reading

Clardy J, Walsh C. Lessons from natural molecules. Nature 2004;432:
829–837.

Fischbach MA, Walsh CT. Assembly line enzymology for polyketide
and nonribosomal peptide antibiotics: logic, machinery, and mecha-

Dewick PM. Medicinal Natural Products, A Biotechnological Approach.
2002. Chichester, UK: John Wiley & Sons Ltd.

Wyatt TD. Pheromones and Animal Behaviour. 2003. Cambridge, UK:
Cambridge University Press.

See Also

Marine Natural Products, Chemical Diversity of
Natural Products in Plants
Pharmaceuticals: Natural Products and Natural Product Models
Natural Products as Anticancer Agents
Chemical Ecology: An Overview
Plants contain numerous natural products (secondary metabolites) that may not participate directly in their growth and development but play an important role in ecological interactions with other organisms. Despite immense chemical diversity, which originates from simple carbohydrates produced because of photosynthesis, plant natural products are formed from only a few biosynthetic building blocks that consist of acetate, mevalonate, and shikimate. These basic building blocks undergo a variety of biosynthetic transformations and combinations that lead to numerous classes of plant natural products including, but not limited to, carbohydrates, fatty acids and their esters, aromatic polyketides (phenols and quinones), terpenoids and steroids, phenyl propanoids (lignans and lignin, coumarins, flavonoids, and isoflavonoids), and alkaloids.

Summarized in this article are representative members of these important classes of plant natural products with special emphasis on their chemical diversity. The article concludes with a brief discussion on recent methods for the maximization of chemical diversity and the production of natural products from plants.

The number of different plant species on the surface of the earth has been estimated to be over 250,000 (1, 2), and only a fraction of these have been investigated for their constituent natural products (3). Plants are known to produce over 100,000 natural products (4). However, according to Verpoorte (5) “extrapolations of the number of species studied and the number of compounds known suggests that, from all plant species, at least a million different compounds could be isolated”. The vast majority of these compounds, commonly referred to as secondary metabolites, does not seem to participate directly in the growth and development of plants (6). In their natural environments, plants coexist and interact with other organisms in a variety of ecosystems (7), and the possible roles that these natural products play in plants, especially in the context of ecological interactions, are being speculated about, appreciated, and debated (8, 9). Although the functional role that secondary metabolites play in the producing organism is a matter of controversy (10, 11), the chemical diversity of plant natural products is well recognized, and it has been suggested that the chemical diversity of plant natural products is far greater than their functional diversity (9, 10, 12, 13).

Two models exist to explain the abundant chemical diversity of plant natural products. In the first model, secondary metabolites produced by plants are believed to be involved in physiological responses during the interactions with their biotic and abiotic environments, especially as elements of their defense arsenals. In this model, the diversity of compounds produced by plants is explained by considering the great diversity of plant life strategies and the vast number of accompanying defense strategies (14–16). The second model is an evolutionary model that makes the assumption that potent biological activity is a rare property for any natural product to possess and therefore is of no value to the producer organism (8). This model, based on the Screening Hypothesis of Jones and Firn (10, 17, 18), suggests that organisms that make and “screen” many chemicals will have an increased likelihood of enhanced fitness simply because the greater the chemical diversity the greater the chances of producing the rare metabolites with useful and potent biological activities.

**Origin of Natural Products in Plants**

It is intriguing that despite the immense chemical diversity exhibited by them, plant natural products are derived from only a few building blocks: acetate (which contains two carbon atoms), mevalonate (which contains five carbon atoms), and shikimate (which contains nine carbon atoms). These building blocks are derived in turn from simple carbohydrates produced...
because of the light-catalyzed reduction of atmospheric carbon dioxide by higher (green) plants during photosynthesis. The products formed by the condensation of the above building blocks (small biosynthetic units) are additionally elaborated (“tailored” or “decorated”) by numerous enzyme-catalyzed reactions such as cyclization, elimination, rearrangement, reduction, oxidation, methylation, and so forth. Chemical diversity that results from these “decoration” reactions will be considered under each biosynthetic class of plant natural products.

Carbohydrates

Although not as structurally diverse as other classes of natural products, carbohydrates are among the most abundant chemical constituents of plants. All animals and most microorganisms depend on plant-derived carbohydrates for their nourishment and survival. Simple carbohydrates (aldoses and ketoses), the first formed products of photosynthesis, are used by plants to make their food reserves, as starter units for the synthesis of plant secondary metabolites, and to make sugar derivatives (glycosides) of products of secondary metabolism. Plant carbohydrates consist of monosaccharides (pentoses and hexoses), disaccharides, oligosaccharides, and polysaccharides. Monosaccharides of plant origin include the stereoisomeric forms of hexose sugars (D-glucose (A2), β-D-galactose (A2), β-D-mannose (A3), α-L-rhamnose (A4), β-D-xylose (A5), β-D-ribose (A7), and β-D-fructose (A8) (Fig. 1)). L-ascorbic acid (A9), commonly known as vitamin C and occurring in most fresh fruits and vegetables, is a monosaccharide derived from D-glucose. As the name implies, disaccharides are formed from two monosaccharide units and contain a C-C glycosidic bond. Among the common plant disaccharides, maltose (A10) and lactose (A11) contain respectively C(1)–O–C(4) and C(1)–O–C(4) links formed between two D-glucose (A2) units, whereas sucrose (A12) contains the C(1)–O–C(2) link formed between D-glucose (A2) and D-fructose (A8) units (Fig. 1). Polysaccharides that are polymeric monosaccharides perfom two major functional roles, namely, the food reserves and structural elements in plants. Amylose (A13), an example of a storage polysaccharide, is a linear polymer that contains 1000–2000 C(1)–C(4) linked glucopyranose units. Cellulose is an example of a structural polysaccharide composed of a linear chain of ca. 8000 residues of C(1)–C(4) linked glucopyranose units. It is noteworthy that cellulose, the main constituent in plant cell walls, is the most abundant organic material on earth.

Products of Acetate Pathway

The two-carbon precursor, acetyl coenzyme A (acetyl-CoA), is the initial substrate for synthesis of the carbon backbone of plant polyketides. As the name implies, polyketides are naturally occurring polymers of ketene (CH2=O) and contain alternating carbonyl and methylene groups derived from the acetate pathway. Polyketides and their derivatives are ubiquitous and are found in all organisms known to produce secondary metabolites. Because of their immense structural diversity, a unified classification of polyketides has yet to emerge (19). Plant polyketides are represented by two major classes of metabolites: fatty acids and aromatic compounds.

Fatty acids and their esters

Plants contain both saturated and unsaturated fatty acids mostly as esters of the trihydroxy alcohol, glycerol. As they are derived from the linear combination of acetate (C2) units, common fatty acids possess an even number of carbon atoms and contain a straight chain. Thus, fatty acids that contain an odd number of carbons are rare in nature. Over 300 fatty acids belonging to 18 structural classes occur in plants (20). Among these, the more common are saturated fatty acids that contain 16 or 18 carbon atoms such as lauric acid, myristic acid, palmitic acid, and stearic acid, and the unsaturated analogs of stearic acid, namely, oleic acid and linoleic acid. Oils produced by many plants constitute glycerol esters of both saturated and unsaturated fatty acids. The other classes of fatty acids are defined by the number and arrangement of double or triple bonds and various other functional groups.

Aromatic polyketides in plants

Aromatic natural products of polyketide origin are less prevalent in plants compared with microorganisms. The majority of the plant constituents that contain aromatic structures are known to arise from the shikimate pathway (see below). Unlike those derived from the shikimate pathway, aromatic products of the polyketide pathway invariably contain a meta oxygenation pattern because of their origin from the cyclization of polyketides. Phenolic compounds such as chrysophanol-anthrone (B1), and emodin-anthrone (B2), and the anthraquinones, aloes-emodin (B3) and emodin (B4) (Fig. 2), are products of the polyketide pathway and are found to occur in some plants of the genera Cassia (Leguminosae) (21), Rhamnus (Rhamnaceae) (22), and Aloe (Liliaceae) (23). The dimer of emodin-anthrone (B2), namely hypericin, (B5) is a constituent of the antidepressant herbal supplement, St. John’s wort (Hypericum perforatum, Hypericaceae) (24).

Products of Mevalonate Pathway

Mevalonic acid, a six-carbon building block, is made up from three molecules of the most basic two-carbon precursor, acetyl-CoA. The mevalonate pathway, which involves the intermediary of mevalonic acid, directs acetate into a series of natural products different from those derived directly from the acetate pathway and includes terpenoids and steroids. Terpenoids constitute the most chemically diverse and one of the largest groups of plant natural products, and therefore a detailed discussion on this group of natural products is warranted.

Most terpenoids are derived from mevalonic acid (MVA) through the universal precursor isopentenyl diphosphate (IPP) and its allylic isopentenyl diphosphate (DMAPP). Thus, the vast majority of terpenoids contain the basic structural residue 2-methylbutane, often less precisely referred to as isoprene units. These C5 hemiterpene units combine with
natural products in plants, chemical diversity of

Figure 1. Chemical diversity of plant carbohydrates.

A1 β-D-Glucose A2 β-D-Galactose A3 β-D-Mannose A4 α-L-Rhamnose A5 β-D-Xylose

A6 α-L-Arabinose A7 β-D-Ribose A8 β-D-Fructose A9 L-Ascorbic acid

A10 Maltose A11 Lactose A12 Sucrose

A13 Amylose A14 Cellulose

Figure 2. Examples of some aromatic polyketides in plants.

B1 Chrysophanol-anthrone B2 Emodin-anthrone B3 Aloe-emodin

B4 Emodin B5 Hypericin

each other in a variety of ways leading to mono- (C10), sesqui- (C15), di- (C20), sester- (C25), tri- (C30), tetra- (C40), and poly- (C5n (n > 8)) terpenes. The primary products of condensation undergo more elaboration (reduction, oxidation, derivatization, etc.) and “decorations” that lead to terpenoid hydrocarbons, alcohols and their glycosides, ethers, aldehydes, ketones, and carboxylic acids and their esters, which makes terpenoids the most diverse class of plant natural products. It is noteworthy that over 40,000 different terpenoids have been isolated and characterized from natural sources including plants (25, 26).

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Monoterpene precursors undergo a variety of cyclization and rearrangement reactions leading to diverse monoterpenes and bicyclic monoterpene derivatives that contain irregular monoterpene carbon skeletons and bicyclic monoterpenoids that consist of hydrocarbons, alcohols, and ketones. Compared with cyclopentane and cyclopropane analogs, cyclobutane and cyclohexane monoterpenoids that contain irregular monoterpene carbon skeletons are rare in nature. (+)-Trans-chrysanthemic acid (D1) and (+)-trans-pyrethric acid (D2) esters, which are known to occur in flower heads of Chrysanthemum cinerariifolium (Compositae), are two important examples of cyclocopropane monoterpenoids. Noteworthy examples of cyclobutane monoterpenoids are (15.25)-frangol (D3), which occurs in the roots of Artemisia frangrans (Asteraceae), and junonione (D4), which occurs in the fruits of the juicy tree (Juniperus communis, Cupressaceae) (Fig. 4).

To date about 200 cyclopentane monoterpenoids are known (32), and the majority of these in plants occur as iridoids and seco-iridoids that contain the iridane carbon skeleton fused to a six-membered oxygen heterocycle. The simplest iridoid, (+)-nepetalactone (D5), is a constituent of the volatile oil of Nepeta cataria (Labiatae), which is known to be a powerful cat attractant and stimulant. Other well-known plant-derived iridoids consist of a diverse array of valepotriates known to occur in the popular herbal supplement valerian (Valeriana officinalis, Valerianaceae). Most of these valepotriates, including (+)-valerian (D6), the constituent responsible for the tranquilizing properties of the valerian, contain several hydroxyl groups centered with the C3, C4, and C5 moieties of iridoids also occur as plant constituents. Important examples are (+)-aspiruloside (D7) with insect antifeedant activity in Asperula odorata (Galium odoratum, Rubiaceae) and many other plants and (+)-loganin (D8) from the fruits of Strychnos nux vomica (Loganiaceae). Although not as prevalent as iridoids, the seco-iridoids, (+) -oleuropin (D9), (+)-jasmonamide A (D10), and (+)-secologanin (D11), Fig. 4, have been isolated from many parts of the olive tree (Olea europaea, Oleaceae), Jasminum multiflorum (Oleaceae) and Strychnos nux vomica (Loganiaceae), respectively.

Cyclohexane monoterpenes are a chemically diverse group of monoterpenoids that occur in the plant kingdom mainly as hydrocarbons, alcohols, ketones, aromatic hydrocarbons, and carboxylic acids (Fig. 5). The saturated hydrocarbon trans-ρ-methane (E1) is a constituent of the oil of turpentine and the resin of pine (Pinaceae) trees. Its unsaturated analogs, namely (R)-(+)-limonene (E2) [present in oil of orange (Citrus aurantium) and mandarin (Citrus reticulata, Rutaceae) peel oil]; ω-terpine (E3) and terpinolene (E4) in some Citrus, Juniperus, Mentha and Pinus species; (R)-(-)- u-pheanthrene (E5) in Eucalyptus pheanthra (Myrtaceae); and (S)-(+)- p-phenanthrene (E6) in water fern Wateria (Phalangium aquilicum, Umbelliferae), are components of many plant volatile oils. The rich chemical diversity of cyclohexane monoterpenes is apparent from the natural occurrence of all four pairs of ρ-methane-3-ol enantiomers, for
example, (−)-menthol (E7) [a major component of peppermint (Mentha piperita, Labiatae) oil], (+)-neomenthol (E8) [a constituent of Japanese peppermint (Mentha arvensis) oil], and (−)-neoisomenthol (E9) in geranium (Pelargonium roseum, Geraniaceae) oil. The unsaturated versions of p-menthol, namely p-menthenol, exhibit extensive regioisomerism and are represented by (−)-pulegol (E10) (a constituent of several peppermint (Mentha gentilis and M. spirata, Labiatae) oils), (−)-isopulegol (E11) in Mentha arvensis (Labiatae), (−)-α-pinene (E12) in several Mentha and Eucalyptus species, (−)-α-terpineol (E13) in Artemisia, Eucalyptus, Juniperus, and Mentha species), and (−)-carveol (E14). Oxidation products of both saturated and unsaturated cyclohexane monoterpenoids also occur in nature. Of these, the most abundant in the plant kingdom are (−)-menthone (E15) in peppermint.
Figure 5 Chemical diversity of monocyclic cyclohexane and aromatic monoterpenes. (Mentha x piperita) oil), (−)-isopulegone (E16) (in oil of Mentha pulegium), (−)-piperitone (E17) (in Eucalyptus oil), and (−)-carvone (E18) (in ripe fruits of dill (Anethum graveolens, Umbelliferae) and caraway (Carum carvi, Umbelliferae)).

Aromatic versions of cyclohexane monoterpenes (benzenoid menthanes or cymenes) are also found in nature and are constituents of some plants frequently used as spices. The hydrocarbon p-cymene (E19) has been found to occur in the oils of cinnamon (Cinnamomum zeylanicum), cypress, eucalyptus, thyme, and turpentine, whereas m-cymene (E20) is a constituent of the oil of blackcurrant (Ribes nigrum, Saxifragaceae). The corresponding phenols, thymol (p-cymen-3-ol) (E21) and carvacrol (p-cymen-2-ol) (E22), have been found to occur in many plants. Thymol (E21) is a constituent of thyme (Thymus vulgaris, Labiatae) and Orthodon angustifolium (Labiatae). Carvacrol (E22) has been found to occur in oils of thyme, marjoram, origanum, and summer savoy.

A different chemical diversity of monoterpenes is apparent from the natural occurrence of their bicyclic analogs that bear cyclopropane (carane and thujane types), cyclobutane (pinane type), and cyclopentane (camphane/bornane, isocamphane and fenchone types) rings (Figs. 6 and 7). The carane type of bicyclic monoterpenoids in plants is represented by (−)-3-carene...
Figure 6  Bicyclic cyclopropane and cyclobutane monoterpenoids.

Figure 7  Bicyclic cyclopentane monoterpenoids.

(F1) that occurs in Pinus longifolia (Pinaceae) and the related carboxylic acid, (+)-chaminic acid (F2), in Chamaecyparis nootkatensis (Cupressaceae). Compared with caranes, the thujane type of monoterpenoids is more abundant in plants. The hydrocarbon analog (−)-3-thujene (F3) has been found to occur in the oils of coriander (Coriandrum sativum), Eucalyptus, and Thuja occidentalis (Cupressaceae). Its regiosomer, (+)-sabinene (F4), occurs in Juniperus sabina (Cupressaceae). The hydration product of (−)-3-thujene, namely (−)-thujol (F5), occurs in plants belonging to the genera Thuja, Artemisia, and...
chemically diverse group of terpenoids known in nature. Like acyclic sesquiterpenes, about 10,000 sesquiterpenoids are known (32), and in the plant kingdom they commonly occur as hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, lactones, and oxiranes. The acyclic sesquiterpene hydrocarbons are the most thoroughly studied sesquiterpenoids, occurring in a large number of species, and in the plant kingdom they commonly occur as hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, lactones, and oxiranes. The acyclic sesquiterpene hydrocarbons, which are the most chemically diverse group of sesquiterpenoids known in nature, are included in this review.

Sesquiterpenoids

The C15 terpenoids known as sesquiterpenoids are the most chemically diverse group of terpenoids known in nature. Like monoterpenoids, many sesquiterpenoids contribute to the flavor and fragrance of a variety of plants. To date about 10,000 sesquiterpenoids are known (32), and in the plant kingdom they commonly occur as hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, lactones, and oxiranes. The acyclic sesquiterpene hydrocarbons are the most thoroughly studied sesquiterpenoids, occurring in a large number of species, and in the plant kingdom they commonly occur as hydrocarbons, alcohols, aldehydes, ketones, carboxylic acids, lactones, and oxiranes. The acyclic sesquiterpene hydrocarbons, which are the most chemically diverse group of sesquiterpenoids known in nature, are included in this review.

Figure 8: Acyclic plant sesquiterpenoids and their derivatives.
Natural Products in Plants, Chemical Diversity of

Figure 9  Cyclization modes of farnesane skeleton that lead to diverse monocyclic and bicyclic sesquiterpenoids.

respectively. Some elemene-type sesquiterpenoids present in plants are represented by (−)-bicycloelemene (I8) from peppermint (Mentha piperita) and M. arvensis and β-elemenone (I9) from Commiphora abyssinica. The humulane type of monocyclic sesquiterpenes, which contain an 11-membered macrocyclic ring, is also found in plants and includes regioisomeric α- and β-humulenes (I10 and I11, respectively) from Lindera strychnifolia (Lauraceae), and (−)-humulol (I12), all of which are important constituents of the essential oils from clove (Caryoppylli flos, Caryophyllaceae), hops (Humulus lupulus, Cannabaceae), and ginger (Zingiber zerumbeticum, Zingiberaceae). Bicyclic sesquiterpenoids are formed because of two
carbon–carbon bonds, each linking two carbon atoms of the farnesane skeleton together.

**Diterpenoids**

The diterpenoids, which contain 20 carbon atoms, are represented by acyclic, monocyclic, bicyclic, tricyclic, and tetracyclic structures. Over 5,000 naturally occurring diterpenoids, many of which frequently occur in plant families Araliaceae, Asteraceae, Cistaceae, Cupressaceae, Euphorbiaceae, Leguminosae, Labiatae, and Pinaceae, are known (32). The acyclic diterpenoid alcohol phytol (J1) (Fig. 11) is a part of the structure of chlorophyll. A group of monocyclic diterpenoids with a 14-carbon macrocyclic ring called cembranes [e.g., cembrene A (J2)] also occurs in plants and is represented by over 100 members (32). Among the other cyclic diterpenoids, the most abundant in plants are the bicyclic labdanes [e.g., (−)-forskolin (J3)], tricyclic abietanes [e.g., abietic acid (J4)], and tetracyclic kauranes [e.g., (−)-kaurane (J5)], represented respectively by 500, 200, and 100 members (32). Baccatin III (J6), which is the diterpenoid part of the well-known anticancer drug paclitaxel...
Triterpenoids and steroids

Triterpenoids and steroids are groups of natural products that contain about 30 carbon atoms. They have a common origin, and their structures can be considered as being derived from that of squalene. Triterpenoids are found mostly in the plant kingdom, whereas steroids occur in plants, animals, and microorganisms. The chemical diversity of plant triterpenoids results from the ability of the C30 precursor, squalene, to undergo various modes of cyclization and subsequent "decoration" reactions. The plant triterpenoids belong to two main groups, the tetracyclic and pentacyclic. The tetracyclic triterpenoids, which consist of dammarane (K1) and tricyclane (K2) among others, are regarded by some authors as methylated steroids. The group of pentacyclic triterpenoids is by far the most diverse and is divided into five main groups: friedelane (K3), lupane (K4), ursane (K5), oleanane (K6), and houpe (K7). (Fig. 12). The steroids are modified triterpenoids that contain the tetracyclic ring system present in lanosterol. Chemical diversity represented by steroids depends mainly on the nature of the side chain attached to the steroid nucleus. Most prevalent in the plant kingdom is stigmastane (K8) and cycloartane (K9) classes of steroids (Fig. 12).

Triterpenoids and steroids frequently occur in many plant species as their glycosides called saponins (33). The chemical diversity of saponins is dependent therefore on both the nature of the 30-carbon moiety and the carbohydrate residue. Some saponins contain carbohydrate residues attached to several different positions of the aglycone (triterpenoid or the steroid) skeleton. Saponins are classified into 11 main structural classes based on the carbon skeletons of their aglycone moiety (33). In addition to their soap-like behavior in aqueous solution because of this combination of polar (carbohydrate) and nonpolar (aglycone) structural elements, saponins exhibit a diverse range of pharmacological and medicinal properties. Within the plant kingdom saponins are present in two major taxonomic classes, Magneliopsida (dicot) and Liliopsida (monocot) (33). Some important examples of saponins include glycyrrhizic acid (K10) from licorice and digitoxin (K11) from foxglove (Digitalis purpurea).

Products of Shikimate Pathway

The shikimate pathway links the metabolism of carbohydrates to the biosynthesis of aromatic natural products via aromatic amino acids. This pathway, which is found only in plants and microorganisms, provides a major route to aromatic and phenolic natural products in plants. To date, over 8,000 phenolic natural products are known, which accounts for about 40% of organic carbon circulating in the biosphere. Although the bulk of plant phenolics are components of cell wall structures, many phenolic natural products are known to play functional roles that are essential for the survival of plants.

It has been noted that the chemical diversity of plant phenolics is as vast as the plant diversity itself. Most plant phenolics are derived directly from the shikimic acid (simple benzoic acids), shikimate (phenylpropanoid) pathway, or a combination of shikimate and acetate (phenylpropanoid-acetate) pathways.

Products of each of these pathways undergo additional structural elaborations that result in a vast array of plant phenolics such as simple benzoic acid and cinnamic acid derivatives, monolignols, lignans and lignin, phenylpropanoids, coumarins, stilbenes, flavonoids, anthocyanidins, and isoflavonoids.

Benzoic acid derivatives

As apparent from their structures, many benzoic acid derivatives are directly formed from shikimic acid by dehydration, dehydrogenation, and enolization reactions. Gallic acid is a component of gallotannins common in some plants that are used in the tanning of animal hides to make leather. Astringency of some foods and beverages, especially coffee, tea, and wines, is because of their constituent tannins. Other benzoic acid derivatives that occur in plants include protocatechual acid, 4-hydroxybenzoic acid, and salicylic acid.

Cinnamic acid derivatives

Cinnamic acid and its derivatives found in plants originate from the aromatic amino acids L-phenylalanine and L-tyrosine by the elimination of ammonia. Some common natural cinnamic acid derivatives include p-coumaric acid, caffeic acid, ferulic acid, and sinapic acid.

Monolignols, lignans, and lignin

The alcohols formed from some cinnamic acid derivatives, namely p-coumaryl alcohol, coniferyl alcohol (L1), and sinapyl alcohol (L2), commonly known as monolignols, undergo dimerization reactions that yield lignans such as (+)-pinoresinol (L3), (-)-sesamin (L4), (-)-matairesinol (L5), and podophyllotoxin (L6). (Fig. 13). Several thousand lignans are found to occur in nature. Lignins, the structural components of plant cell walls, are polymers of monolignols and/or lignans.

Phenylpropenes

Phenylpropenes are derived from cinnamic acid and its derivatives by a series of reductions and other transformations. Cinnamaldehyde, the first product of the reduction of cinnamic acid, occurs in the bark of cinnamon (Cinnamomum zeylanicum, Lauraceae). Several hydrocarbon analogs are also known to occur in plants. Anethole is the main constituent of oils from anise (Pimpinella anisum, Umbelliferae), fennel (Foeniculum vulgare, Umbelliferae), and star anise (Illicium verum, Illiciaceae). Eugenol is a major constituent of cinnamon leaf, whereas myristicin occurs in nutmeg (Myristica fragrans, Myristicaceae).

Coumarins

Coumarins derive their name from their precursor, α-coumaric acid. They occur widely in plants both in the free form and as...
Triterpenoid and steroid skeletons common in plants and structures of some saponins. Glycosides and are commonly found in families such as the Umbelliferae and Rutaceae. The parent compound coumarin (M1) is found in sweet clover (Melilotus alba, Leguminosae) and its hydroxyl derivative umbelliferone (M2) has been isolated from several Ferula spp. (Umbelliferae). Coumarins with complex structures also occur in plants and are formed by incorporating additional carbons derived from the mevalonate pathway. Alkylation of umbelliferone (M2) with dimethylallyl diphosphate (DMAPP) leads to demethylsuberosin (M3), which undergoes cyclization yielding marmesin (M4), the precursor of naturally occurring furanocoumarins, psorolen (M5), and bergapten (M6) (Fig. 14).

Stilbenes, flavonoids, anthocyanidins, and isoflavonoids

In contrast to other plant phenolics, the basic carbon skeleton of stilbenes, flavonoids, anthocyanidins, and isoflavonoids
indole, and imidazole ring systems. Alkaloids are also known to originate from aromatic amino acids, tyrosine, phenylalanine, tryptophan, anthranilic acid, and histidine, and thus contain pyrrolidine, pyrrolizidine, piperidine, quinolizidine, indolizidine, pyridine, quinoline, isoquinoline, indole, and imidazole ring systems. Alkaloids are also known to originate from mixed biosynthetic pathways, the most important of which include terpenoid and steroidal alkaloids. A limited number of alkaloids that contain a purine ring (e.g. caffeine) also occur in plants. Of the large number and variety of plant alkaloids, only a few are considered here for the purpose of illustration of their chemical diversity.

**Alkaloids derived from aliphatic amino acids and nicotinic acid**

Alkaloids that contain pyrrolidine and pyrrolizidine ring systems are derived from the nonprotein amino acid, L-ornithine. Cocaine (N1) and L-3,4-dihydroxyphenylalanine (L-DOPA), the two important pyrrolidine alkaloids that contain a piperazine ring system, have been found to occur in coca (Erythroxylon coca, Erythroxylaceae) leaves and whole plant of the deadly nightshade (Atropa belladonna, Solanaceae). The hepatotoxic alkaloid senecionine (N2) contains a bicyclic pyrrolizidine skeleton derived from two molecules of L-ornithine. Piperidine alkaloids, for example, piperine (N3), and pseudopelletierine, are known to be derived from the amino acid, L-lysine. Piperine is responsible for the pungency of black pepper (Piper nigrum, Piperaceae), whereas pseudopelletierine is a constituent of the bark of pomegranate (Punica granatum, Puniceae). The bicyclic ring system in quinolizidine alkaloids such as (+)-sparteine (N4) in the broom plant (Cytisus scoparius, Leguminosae) is derived from two molecules of L-lysine in a manner similar to the L-ornithine-derived pyrrolizidine ring system. Indolizidine alkaloids derived from L-lysine via the cyclic amino acid, L-3-piperidone, contain fused six- and five-membered rings with a nitrogen atom at the ring fusion. An important example of an indolizidine alkaloid is swainsonine (N5), which occurs in the leguminous plant Swainsona canescens. Alkaloids that contain a pyridine ring also occur in the plant kingdom. Two common plant-derived pyridine alkaloids, nicotine (N6) and anabasine, both of which are found in tobacco (Nicotiana tabacum, Solanaceae), contain a pyridine and a pyrrolidine or a piperidine ring, respectively (Fig. 13).

**Alkaloids derived from aromatic amino acids**

Aromatic amino acids that originate from the shikimate pathway also act as precursors to many alkaloids. A alkaloids that contain a phenylethylamine moiety are derived from L-tyrosine or its oxidation product L-DOPA. Mescaline (N7), originating from the latter amino acid is known to occur in several cacti and is responsible for the hallucinogenic activity of peyote (Lophophora williamsii, Cactaceae). Lophocerine is a tetrahydroisoquinoline alkaloid derived from L-dopa, and found to occur in a different Lophophora species, L. schottii. Condensation of two phenylethyl units derived independently from the same or different aromatic amino acid(s) leads to a variety of benzyltetrahydroisoquinolines, which, with additional structural modifications, produce a diverse range of alkaloids. (-)-Reticuline occurring in several plant species of Amonoea is an important benzyltetrahydroisoquinoline alkaloid that acts as a precursor to several pharmacologically active.
alkaloids such as papaverine (N8), (+)-tubocurarine (N9), and morphine (N10). Papaverine and morphine are known to occur in opium (Papaver somniferum, Papaveraceae) and are responsible for its narcotic activity, whereas (+)-tubocurarine (N9) is a muscle relaxant obtained from the arrow poison of the South American Indians, curare (Chondrodendron tomentosum, Menispermaceae). Phenethylisoquinoline alkaloids are similar structurally to benzylisoquinolines but as the name implies contain a phenylethyl moiety instead of a benzyl moiety as the pendant aromatic ring. Both (S)-autumnaline and the cyclized analog colchicine (N11) belonging to this class have been found to occur in the seeds of autumn crocus (Colchicium autumnale, Liliaceae).

Maximization of Chemical Diversity and Production of Natural Products in Plants

As is apparent from the foregoing discussion, plants produce a huge array of natural products, many of which are specialized secondary metabolites associated with particular plant species and/or having to play important ecological roles. It is likely that for diversification and survival of the plant kingdom, individual plants had to develop the ability to perform in vivo combinatorial chemistry by mixing and matching and evolving the genes required for different secondary metabolite biosynthetic pathways (34, 35). With the elucidation of several secondary metabolic pathways in plants together with the advent of techniques for the introduction of genes into plants and the availability of an increasing number of genes, it has become possible to modulate and diversify secondary metabolite production in transgenic plants and plant cell cultures.

Two general approaches for the production of long-chain polyunsaturated fatty acids usually found in fish oil have been employed, both of which used 18 carbon fatty acids endogenous to plants as the starting substrates (36). Soybean and canola, the oilseed plants rich in omega-6 fatty acids, have been engineered to produce omega-3 polyunsaturated fatty acids such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) (37, 38). Chalcone synthase (CHS), the first plant natural product polyketide synthase (PKS) to be characterized at the molecular level (39), catalyzes the condensation of 4-coumaryl-CoA with three molecules of malonyl-CoA to afford naringenin chalcone, a precursor of the major classes of plant flavonoids. The cloning of a novel type III polyketide chalcone synthase (PCS) from aloe (Aloe arborescens, Liliaceae) rich in aromatic polyketides, especially quinones such as aloe-emodin and emodin, resulted in...
PCS-catalyzed condensation of five molecules of malonyl-CoA to produce 5,7-dihydroxy-2-methyl chromone new to this plant (40). Another novel Aloe arborescens type III PKS that produces two hitherto unknown aromatic octaketides, SEK4 and SEK4b, has recently been reported (41). The application of plant cell cultures for the production of the polyketide hypericin from St. John’s wort (Hypericum perforatum, Hypericaceae) has been investigated (42).

To date over 30 plant terpenoid synthases have been cloned as cDNAs, and many of these were found to encode enzymes of secondary metabolism (43). Isolation and analysis of six genomic clones encoding monoterpene (\((-\)-pinene and \((-\)-limonene), sesquiterpene (\(\alpha\)-\(\delta\)-selinene) and diterpene (abietadiene) synthases from Abies grandis, and a diterpene (taxadiene) synthase from Taxus brevifolia have been reported (44). Overexpression of a cotton farnesyl diphosphate synthase (FPPS) in transgenic Artemisia annua has resulted in 3- to 4-fold increase in the yield of the sesquiterpenoid anti-malarial drug, artemisinin, in hairy roots (45).

Plant cell culture, an environmentally friendly and renewable alternative for the production of plant natural products, has also been investigated to obtain taxane diterpenoids from Taxus sp. (46–48) and terpene indole alkaloids from the Madagascar periwinkle (Catharanthus roseus) (49). A recent study has provided evidence for the production of novel terpene indole alkaloids using both differentiated Catharanthus roseus (seedlings) and hairy root culture (50).

References


Further Reading


Hartmann T, Dierich B. Chemical diversity and variation of pyrrolizidine alkaloids of the senecionine type: biological need or coincidence? Planta 1998;206:443–452.


See Also

Natural Product Discovery, Molecular Biological Approaches to
Natural Products in Microbes, Chemical Diversity of
Natural Products: An Overview
Terpenoids in Plants
Pharmaceuticals: Natural Products and Natural Product Models

Sheo B. Singh, Merck Research Laboratories, Rahway, New Jersey

doi: 10.1002/9780470048672.wecb444

Natural products have played a vital role in the treatment of human ailments for thousands of years and continue to play a big role in the modern discovery of new agents for the treatment of diseases today. In certain therapeutic areas, natural products account for almost all key modern medicine used today. Many drugs are formulated and used directly as they are found in nature, some are derived directly from natural products by semisynthesis, and others are modeled after natural products. In this overview, examples of the pharmaceutically important natural products have been summarized.

Introduction

Natural product preparations have played a vital role in empiric treatment of ailments for thousands of years in many advanced civilizations and continue to play a significant role even today in various parts of the world. The use of plants and preparations derived from plants has been the basis for the sophisticated medical treatments in Chinese, Indian, and Egyptian civilizations for many thousands of years. These medical applications have been documented in the Chinese 萩经 (1100 BC), in the Indian Ayurveda (1000 BC), and in Egyptian medicine as early as 2900 BC. Plant preparations continued to be the basis of medical treatments in the ancient Western world as well. This knowledge migrated through Greece to Western Europe, including England, during the ancient period and led to its formal codification in the United Kingdom and to the publication of the London Pharmacopoeia in 1618.

The isolation of strychnine (1), morphine (2), atropine (3), colchicine (4), and quinine (5) in the early 1800s from the commonly used plants and their use for the treatment of certain ailments might constitute the early idea of "pure" compounds as drugs. E. Merck isolated and commercialized morphine (2) as the first pure natural product for the treatment of pain (1–3). Preparations of the Willow tree have been used as a painkiller for a long period in traditional medicine. Isolation of salicylic acid (6) as the active component followed by acetylation produced the semisynthetic product called "Aspirin" (7) that was commercialized by Bayer in 1899 for the treatment of arthritis and pain (4).

The World Health Organization estimates that herbal and traditional medicines, derived mostly from plants, constitute primary health care for ~89% of the world population even today. The compounds produced by plants play significant roles in the treatment of diseases for the rest of the 20% of populations that are fortunate to use modern medicine. About 50% of the most prescribed drugs in the United States consist of natural products or their semisynthetic derivatives, or they were modeled after natural products. "Curare," the crude extract from the South American plant, Chondodendron tomentosum, and the derived purified compound tubocurarine, has been used as anesthetic in surgery until recently. Purified digoxin, as well as the crude extracts that contain digitalis's glycosides from foxglove plant, Digitalis lanata, is used as cardiotoxic even today.

Drugs derived from microbial fermentations have played perhaps a bigger role in the modern drug discovery and have revolutionized the practice of medicine, which leads to saving human lives. Although the contribution of purified natural products as single agent drugs is significant in almost all therapies, their contribution in the treatment of bacterial infection is perhaps most critical (5). Natural products constitute drugs or leads to all but three classes of antibiotics. The discovery of microbial natural products-based antibiotics began with the serendipitous observation by Fleming in 1929 that bacterial growth was prevented by the growth of Penicillium notatum. Although this discovery was highly publicized and very important, it took over 10 years before the active material, penicillin, was purified and structurally elucidated by Florey and Chain in early 1940s. Subsequent commercialization was very quick, driven largely by the medical needs of World War II. Penicillin was one of the first broad-spectrum antibiotics that treated bacterial infection and saved millions of lives. Fleming, Florey, and Chain were awarded the Nobel Prize in 1945 for their efforts on penicillin. The success of penicillin led to unparalleled efforts by government, academia, and the pharmaceutical industry to focus drug-discovery efforts based on the newfound "microbial"
sources for the discovery of natural products beyond plants. However, initial efforts were mostly focused on the discovery of antibiotic compounds from fermentations of a variety of microorganisms of not only fungal origin but also soil-dwelling prokaryotes (e.g., Streptomyces spp.), which led to the discovery by 1962 of almost all novel classes of antibiotic scaffolds that are being used today. The antibiotic discovery effort was performed largely by Fleming’s method of detection of antibacterial activity on petri plates. Zones of inhibition of bacterial strains on agar plates were measured after applying whole broth or extracts obtained from microbial ferments (5). As newer biological assays and screening techniques became available in the 1960s, microbial sources, along with plant and marine sources, started to be used for screening against other therapeutic targets, which led to the discovery of leads and drugs in those areas. Examples of these discoveries will be discussed with the target area. As time progressed, improved technologies in biology and chemistry helped with the popularization of natural products; natural product extracts became part of the screening resource in most large pharmaceutical houses from 1960 through the 1980s until their de-emphasis in the early 1990s. Therefore, natural product extracts became popular sources for the screening against purified enzymes and receptors, an occurrence that led to the discovery of leads and drugs in those areas. Examples of these discoveries will be discussed with the target area. As time progressed, improved technologies in biology and chemistry helped with the popularization of natural products; natural product extracts became part of the screening resource in most large pharmaceutical houses from 1960 through the 1980s until their de-emphasis in the early 1990s. Therefore, natural product extracts became popular sources for the screening against purified enzymes and receptors, an occurrence that led to the identification of many nonantibiotic natural products that have revolutionized the practice of medicine, saved countless human lives, improved quality of life, and perhaps helped increase life expectancy for humans.

**Antibacterial Agents**

Natural products contribute to over 80% of all antibiotics that are in clinical practice today. Natural products contribute to all but three classes of antibiotics (3).

Penicillin was the first β-lactam and the first broad-spectrum antibiotic discovered that started the “Golden age” (1940–1962) of antibiotics. The structure of penicillin contains a thiazolidine ring that is fused to a β-lactam ring. The existence and stability of the β-lactam ring was highly controversial at the time despite the availability of a single crystal X-ray structure of one of the penicillins. Penicillin G (8) was the first penicillin that was clinically used. Penicillin G was converted easily by chemical or biochemical means to β-amino-penicillanic acid (9), which became the lead for the semisynthetic modifications that led to the synthesis of various penicillin derivatives. Some early derivatives (e.g., amoxicillin 10) are still in clinical use. Penicillin (general structure 11) bind to penicillin-binding proteins and inhibit the bacterial cell wall. Penicillins became targets of β-lactamases without having significant intrinsic antibacterial effectiveness of these compounds (11, 12). Cephalosporin C (12), a second class of β-lactam antibiotics, was first discovered from Cephalosporium acremonium, isolated from a sewer outfall of Sardinia, Italy in 1948. Although Cephalosporin C was less active than penicillin G, it was less prone to β-lactamase action and therefore attracted a lot of attention that led to the development of five generations of orally active clinical agents (e.g., cephalexin, 13 and general structure, 14, 5, 6).

Continued search for even better antibiotics led to the discovery of highly potent and broad-spectrum antibiotic thienamycin (15), the third class of the β-lactams, called carbapenems, in which the sulfur atom of the thiazolidine ring was replaced by a methylene group. Thienamycin was produced by Streptomyces cattleya (17). The primary amine group of thienamycin self-catalyzes the opening of the β-lactam ring, which leads to the concentration-dependent instability that poses a serious challenge for the fermentation-based production of the compound. The Merck group stabilized the compound by replacing the primary amine with an aminomethylideneamino group and synthesized imipenem (16) (8). They developed a highly efficient total synthesis that remains in commercial use today. Imipenem was approved for clinical use 23 years ago in 1985, but it remains one of the most important broad-spectrum hospital antibiotics in the market today. Like other β-lactams, several generations of carbapenems (general structure 17) have been approved for clinical use in recent years (9). As resistance to β-lactam antibiotics increased because of the expression of a variety of β-lactamases, many groups focused their efforts on discovering compounds that could be more reactive to β-lactamases without having significant intrinsic antibacterial activity of their own and pharmacokinetic properties that would be similar to β-lactam antibiotics. This focus led to the discoveries of clavulanic acid (18) and monobactam sulfazecins (19). Nature effectively stabilized the latter monobactam structure by the addition of a 4-thiulamic acid. The β-lactamase inhibitor clavulanic acid was combined with amoxicillin, which...
led to the development of a potent and successful antibacterial agent, Amoxicillin (10) (GSK, Surrey, UK) (10). Chemical modifications of the monobactam produced aztreonam (20), a clinical agent with a narrow spectrum but significantly improved activity against Gram-negative pathogens, particularly Pseudomonas aeruginosa (11).

Immediately after the discovery of penicillin, Waksman started efforts on soil-dwelling bacteria and discovered the first of the aminoglycosides, streptomycin (22) from Streptomyces griseus, in 1943 (6). Subsequently, a series of aminoglycosides was isolated. These aminoglycosides are potent broad-spectrum antibiotics and are potent inhibitors of protein synthesis. Unfortunately, nephrotoxicity limited their wider use, and they are used mainly for treatment of infections caused by Gram-negative bacteria. Continued efforts to screen prokaryotic organisms led to the discovery of the phenylpropanoids (chloramphenicol, 23) and tetracyclines. The latter is a major class of tetracyclic polyketides that were discovered from various species of Streptomyces spp. Although the parent tetracycline (22) was not used as an antibiotic to a great extent, the chloro derivative (Clotrimocycline 24), oxytetracycline (25), and minocycline (26) are clinical agents. This class of compound suffered from the selection for rapid resistance via efflux mechanism that limited their use (6). Recently, however, chemical modifications of the A-ring yielded compounds that overcame the efflux pump and lead to the development of tigecycline (27), as an effective broad-spectrum antibiotic (12).

Another large class of orally active protein synthesis inhibitor antibiotics that were produced by Streptomyces spp. is represented by 14-membered lactones generally called macrolides, exemplified by the first member, erythromycin (28) (6). Chemical modifications of this class of compounds led to many clinical agents such as the aza derivatives, azithromycin (29) and ketolide (telithromycin, 30) (13, 14). Mupirocin (pseudomonic acid, 31) is another protein synthesis inhibitor that was isolated from Pseudomonas fluorescens and is used only as a topical agent (6).

Vancomycin (32), a glycopeptide produced by Streptomyces orientalis, is a key Gram-positive antibiotic, originally discovered in 1954, and remains a critical antibiotic in clinical practice even today for the treatment of Gram-positive bacterial infections (6). Teicoplanin (33), a related glycopeptide produced by Streptomyces teicomyceticus, is a newer antibiotic that complements vancomycin in the clinic but is not effective against vancomycin-resistant bacteria. Ramoplanin (34) represents another glycopeptide that is larger in molecular size and structurally different from vancomycin and teicoplanin; it is in the late stages of clinical development for treatment of Gram-positive bacterial infections. Glycopeptides inhibit the bacterial cell wall. Daptomycin (35), a cyclic lipopeptide produced by Streptomyces roseosporus, is one of the newest members of antibiotics approved for the clinical practice as a broad-spectrum Gram-positive agent. It works by depolarization of the bacterial cell membrane (14). Streptogramins were discovered in the early 1960s but were used for humans only recently when a 70/30 mixture of dalfopristin (36) and quinupristin (37) with the trade name Synercid (King Pharmaceuticals, Bristol, N.J.) was developed for the treatment of drug-resistant Gram-positive bacterial infections (15).

Antibiotics from natural sources range from compounds with small molecular size (e.g., thienamycin) to large peptides (e.g.,
ramoplanin). They generally possess complex architectural scaffolds and densely deployed functional groups, which affords the maximal number of interactions with molecular targets and often leads to exquisite selectivity for killing pathogens versus the host. This function is nicely illustrated by vancomycin binding to its target. Vancomycin has five hydrogen bond contacts with the D-Ala-D-Ala terminal end of peptidoglycan. Resistant organisms modify the terminal D-Ala with D-lactate, which leads to loss of one hydrogen bond and a 1000-fold drop in binding affinity and loss of antibiotic activity (see Fig. 1).  

Antifungal Agents
Significant similarities in the fungal and mammalian cellular processes result in very few fungal-specific drug targets that lead to the development of only a few quality and safe antifungal agents. Amphotericin B (38), a natural product that consists of a polyene lactone, is a highly effective broad-spectrum antifungal agent unfortunately with a very limited safety margin. Recently, glucan synthesis was identified as a fungal-specific target that could be inhibited by a series of cyclic peptides called echinocandins, which were identified in the 1970s as having potent antifungal activities (16, 17). This identification provided impetus for the discovery and development of new related lipopeptides that led to the identification of pneumocandins from Glarea lozoyensis (e.g., pneumocandin Bo, 39). Chemical modifications at two sites of pneumocandin Bo led to the synthesis of caspofungin (40), which was the first in the class of glucan synthesis inhibitors; it is a "potent", highly effective, and safe antifungal agent approved for serious fungal infections in hospitals. Side chain replacements of the related cyclic peptide...
FR901379 (isolated from the fungus Coleophoma empetri) and echinocandin B (isolated from Aspergillus nidulans) led to two additional clinical agents of this class, micafungin (41) and anidulafungin (42), respectively. (16, 17)

**Antimalarials**

Quinine (5) isolated from Cinchona bark was one of the first antimalarials discovered, and it became a model for the discovery and development of some of the most successful antimalarial agents, chloroquine and its successors. However, the development of resistance by the malarial parasite *Plasmodium falciparum* for these drugs has rendered them ineffective. Artemisinin (43), a sesquiterpene peroxide originally isolated from a Chinese herb *Artemisia annua* in 1972 as an antimalarial agent, was chemically modified to a derivative, artemether (44), which is a very effective and widely used antimalarial agent (18). Unfortunately, limited supply of this plant-derived compound rendered it inaccessible for wider use. Recently, biosynthetic genes of artemisinin have been identified and successfully transfected to an heterologous host, *Escherichia coli*. This method has allowed the production of an intermediate, amorphadiene (45) and artemisinic acid (46), which could be
transformed chemically to artmether and potentially could relieve the strain of supply and could provide wider availability (19–21).

**Antivirals**

Most antiviral agents are based on nucleoside structures and have their origin from spongouridine (47) and spongothyminidine (48) that were isolated from marine sponges in the 1950s by Bergmann and his coworkers (22–24). These natural nucleosides possessed sugars other than ribose and deoxyribose and provided rationale for the substitution of the sugars in the antiviral nucleosides with various sugar mimics, including linear polar groups that led to the synthesis of Ara-A (49) and acyclovir (50). HIV protease inhibitors were developed from pepstatin (51), a pepsin inhibitor produced by various fungal species. Pepstatin possesses as the structural component statine a β-hydroxyγ-amino acid that mimics the transition-state intermediate of the hydrolytic reaction catalyzed by the proteases (25). This structure became the foundation for the rational peptidomimetics and design of all HIV protease inhibitors, for example, indinavir (Crixivan®, Merck & Co., Inc., Whitehouse Station, NJ; 52) and others (26).
Antipain Agents

Use of the opium poppy (Papaver somniferum) to ameliorate pain dates back thousands of years, and the active metabolite morphine (2) was isolated first from its extracts in 1806 followed by codeine (53) in 1832 (27, 28). Morphine and its derivatives are agonists of opiate receptors in the central nervous system and are some of the most effective pain relievers known and prescribed for postoperative pain. Morphine and codeine differ by substitution by methyl ether. Unfortunately, addictive properties of these compounds limit their use. Efforts have been made to reduce the addictive properties of morphine, which resulted in a semisynthetic derivative buprenorphine (54) (29). This compound is 25 to 50 times more potent than morphine with lower addictive potential and has been indicated for use by morphine addicts.

Conotoxins, a class of 10 to 35 amino acid-containing peptides produced by cone snails to intoxicate their prey, were isolated and characterized by Olivera and coworkers (30). They are a novel class of analgesics that helped identify the target and blocking of N-type Ca\(^{2+}\) channels. One compound, Ziconotide (55), was synthesized and developed as a treatment for severe chronic pain (31, 32).

The alkaloid epibatidine (56) was discovered from the skin of an Ecuadorian poison frog (Epipedobates tricolor), and its...
potent analgesic activity was demonstrated as early as in 1974 by Daly and coworkers (33–35). The paucity of the material delayed the structure elucidation and was only accomplished after the invention of newer and more sensitive NMR techniques in 1980s. Once the structure was elucidated as chloronicotine derivative, it was synthesized and was shown to antagonize nicotinic acetylcholine receptor, is about 50-fold more potent than morphine without addictive properties, and was under advanced clinical development until its discontinuance in 2003 (36–39).

Antineoplastic Agents

Natural products have played a much bigger role, perhaps second only to antibacterial agents, in the discovery and development of anticancer agents either directly as drugs or leads to drugs (48). In fact, they contribute 64% of all approved cancer drugs (3). Taxol® (paclitaxel) is unarguably the most successful anticancer drug in clinical use. It is a taxane diterpenoid that was isolated from the Pacific yew Taxus brevifolia in 1967 as a cytotoxic agent but not pursued as a development candidate until its novel mode of action was determined in 1979 as a stabilizer of microtubule assembly (49–52). The discovery of the mechanism of action led to the United States National Cancer Institute (NCI) committing significant resources to the large-scale production, eventual clinical development, and approval of Taxol® (paclitaxel) by U.S. Food and Drug Administration in 1992 for treatment of breast, lung, and ovarian cancer. Another important plant product, camptothecin (62), an alkaloid from Camptotheca acuminata, was discovered in 1966, also by the Wall and Wani group (53). The development of this compound was also hampered by the lack of knowledge of the mechanism of action and most importantly by its extremely poor water solubility. Determination of the mechanism of action as an inhibitor of topoisomerases-I led to significant efforts both by NCI and the pharmaceutical industry, which resulted in chemical modification of the structure, introduction of water-solubilizing groups (e.g., amino), and development of several derivatives as anticancer agents exemplified by topotecan (63) and irinotecan (64) (54, 55).

Alzheimer’s Agents

Galantamine (60) is a tetracyclic alkaloid that was isolated originally from G. nivalis and subsequently from Narcissus spp. (45). It was approved by trade name Reminyl® by Johnson & Johnson, New Brunswick, NJ for the treatment of Alzheimer’s disease. That galantamine, a selective inhibitor of acetylcholinesterase, was confirmed by X-ray structural characterization of galantamine bound to a plant acetylcholinesterase (46, 47).
compounds bind to tubulin and inhibit cell division by inhibiting mitosis; they were perhaps the best-known anticancer agents before Taxol. Chemical modifications of vinblastine led to the clinical agents vinorelbine (66) and vindesine (67) (48 one, 59–61). Podophyllotoxin (68) was isolated from various species of the genus Podophyllum spp. as an anticancer agent. Chemical modifications of the naturally occurring epimer, epipodophyllotoxin (69), led to the synthesis and development of etoposide (70) and teniposide (71) as clinical agents (48, 54, 62–65).

Combretastatin A4 phosphate (72) is a phosphate prodrug of combretastatin A4, a cis-dilene, isolated from Combretum caffrum (66). Combretastatin A4 is one of the many combretastatins that inhibits tubulin polymerization (67), shows efficacy against solid tumor, is a vascular targeting agent that blocks the blood supply to solid tumors, and is in Phase III clinical development for the treatment of various types of tumors as a vascular targeting agent (68–71).

Microbial sources have been a very rich source for cancer chemotherapeutic agents. Of particular note is the Streptomyces spp., which has been responsible for the production of many approved anticancer agents that are in clinical practice. These agents are represented by highly diverse structural classes exemplified by the anthracycline family (e.g., doxorubicin, 73) (72–74), actinomycin family (e.g., dactinomycin, 74), glycopeptides family (e.g., bleomycins A2 and B2, 75 and 76), and mitomycin family (e.g., mitomycin C, 77) (72, 76). All these compounds specifically interact with DNA for their mode of action.

Staurosporine (78) produced by Streptomyces spp. is a potent inhibitor of protein kinase C (77–79). This compound inhibits many other kinases with almost equal potency and has become a great tool for the study of kinases. Lack of selectivity for protein kinase C has significantly hampered the development of this compound. Recently, however, several compounds derived from this lead have entered into the clinic for potential treatment of cancer. These include 7-deoxystaurosporine (79) and CGP41251 (80, 81). CGP41251 shows multiple modes of action including inhibition of angiogenesis in vivo.

Microbial sources other than Streptomyces spp. have also provided highly interesting and structurally diverse compounds. Discovery of epothilones from myxobacterial strains by a German group (82) and the Merck group (83, 84) constitute a breakthrough discovery. The Merck group used an assay that mimicked Taxol at the active site for the screening of natural products that led to the isolation of epothilones A (81) and B (82). The discovery of a unique structural class, interfering biological activity, and clinically proven mode of action drew significant attention from the scientific community and led to a variety of approaches, including combinatorial biosynthesis, chemical modifications, and total synthesis, that permitted preparation of many derivatives with improved potency and drug-like properties. A series of these compounds have entered human clinical trials, and many are in the late stages of development. Epothilone discovery and development has been recently reviewed (85).

Recent pursuit of marine microbial sources led to the isolation of salinosporamide A (83). It is a β-lactone produced by the marine bacteria Salinispora tropica and is a proteasome inhibitor (86). Mechanistically, it works by specific covalent modification of the target. This compound has entered human clinical development for treatment of multiple myeloma (87–89).

Although use of marine microbial sources for the discovery of natural products is a somewhat recent phenomenon, marine natural products from higher species have contributed tremendously to the discovery of novel architecturally complex compounds as anticancer agent leads with one, Ecteinascidin 743, now approved in the European Union for treatment of sarcoma. The discovery of natural products derived from marine sources exploded in the 1970s not only because of increased...
level of NCI funding but also because of technological advancements in the techniques for collection of specimens, chemical isolation, and structural elucidation of low amounts of compounds initially isolated. Because of the fear of a limited supply of marine sources for large-scale production, marine natural products have remained the exclusive purview of academia except for a small Spanish pharmaceutical company, PharmaMar, which collaborates closely with academia and governments. Bryostatins are among the most interesting marine natural products known. They were isolated from the bryozoan Bugula neritina. They are a series of polyketide macro lactones represented here by the major congener bryostatin I (84), which is a modulator of protein kinase C, and has been subjected to several human clinical trials (90, 91). Recently, combination studies have been recommended for Phase I and Phase II trials, mainly under the auspices of the NCI. Modeling studies along with a diligent chemical design approach has led to the synthesis of a simplified analog (85) that has been shown to be equally active as bryostatin I in most in vitro studies (92, 93). This compound stands a better chance of being produced at larger scale by total synthesis.

Dolastatins are a class of peptides comprised of mostly nonribosomal amino acids. They were isolated from a sea hare Dolabella auricularia (94). Dolastatin-10 (86) is one of the most potent and the best-studied members (95, 96). It exerts antitumor effect by inhibiting tubulin polymerization and binds at the vinca alkaloid binding site (97, 98). Dolastatin-10 has been studied in Phase II human clinical trials but was discontinued because of lack of efficacy (99). Auristatin PE (77), a synthetic analog, seems to be more promising and is being studied in Phase II human trials (100).

Discodermolide (88) was isolated from Discodermia disso-
lute by using a P388 cell line toxicity bioassay; later it was determined that it stabilized microtubule assembly better than Taxol® and it drew a lot of attention as an anticancer agent (101, 102). Its development, like that of many other complex marine natural products, was hampered because of the lack of ample supply of the material required for the clinical studies. In this case, the supply problem was overcome by the synthetic efforts of the Novartis process group. They synthesized it on a large enough scale to allow clinical studies. Unfortunately, its development seems to have been halted because of toxicity at Phase I (103, 104).

Several other novel, structurally and mechanistically diverse marine natural products have entered various preclinical and clinical studies. One of these products is ecteinascidin-743 (89), ET-743), isolated from the tunicate Ecteinascidia turbinata (105, 106) and recently approved for the treatment of sarcoma in the European Union as the first “direct-from-the-sea” drug. Total synthesis and methods developed during the total synthesis allowed the preparation of a simpler analog phthalascidin (90) with comparable activities (107–109).

Hemiasterlin (92), a tripeptide isolated from a sponge that was chemically modified to HTI-286 (92), which binds to the vinca binding site of tubulin, depolarizes microtubules; it entered into clinical development but was apparently dropped (110, 111).

One difficulty with the cytotoxic agents that are used for the treatment of cancer is the differentiation of cytotoxicity between target tumor cells and normal cells. In an innovative approach, the Wyeth group took advantage of a tumor cell-specific drug delivery mechanism of antibodies. They conjugated calicheam-
icin (93), perhaps the best described member of the ene-dyne
Immunosuppressant Agents

Natural products represent essentially all clinically used immunosuppressant agents. These agents collectively have made organ transplant possible. Cyclosporin (95) is an N-methyl cyclic peptide and originally was isolated from the fungus Trichoderma polyergus as an antifungal agent; almost immediately, the inhibition of T-cell proliferation and in vivo immunosuppressive properties were discovered and led to the development and approval of this molecule as a highly effective immunosuppressive agent (116–118). Natural products isolation of related compounds allowed the discovery of new congeners with reduced or no immunosuppressive activity in favor of antifungal and various other biological activities (e.g., antiparasitic activity) and asthma (119, 120). FK506 (96), a macrocyclic lactone, discovered from Streptomyces tunicatus (97) as an immunosuppressive agent (121–125), was approved for clinical use for organ transplant as Tacrolimus (98). Astel- las, Tokyo, Japan (126). Rapamycin (99), another macrocyclic lactone that is a very potent immunosuppressive agent, was approved as Sirolimus (100). Wyeth, Madison, NJ for clinical use for transplant rejection (127, 128). Rapamycin was isolated from Streptomyces hygroscopicus (129, 130). Mechanistically, all three of these compounds bind to their specific intracellular receptors, immunophilins, and the resulting complexes target the protein phosphatase, calcineurin (cyclosporin and FK506), and mammalian target of rapamycin (mTOR) to exert their immunosuppressive effects (131–133). Rapamycin and FK506 have played significant roles in studies of signal transduction and identification of various targets for other therapeutic applications, such as mTOR. Mycophenolic acid (101) originally was isolated from various species of Penicillium, and its antifungal activity has been known since 1932 (134). Mycophenolate was approved for acute rejection of kidney transplant (135, 136).

Cardiovascular Agents

The biggest impact made by natural products in the treatment of cardiovascular diseases is undoubtedly associated with the discovery of the first of the HMG CoA reductase inhibitors by enzyme-based screening of microbial extracts that led to the isolation (from Aspergillus terreus) and characterization of mevinolin (lovastatin 102), which binds to CD33 antigens expressed on the surface of leukemia blasts. Mylotarg (103), which is a homologue of compactin (300) that was discovered earlier (137, 138). These compounds possess a lipophilic hexahydrodecalin, a 2-methylbutanoate side chain, and a β-hydroxy-γ-lactone connected to the decalin unit with a two-carbon linker. These compounds are potent inhibitors of HMG CoA reductase, the rate-limiting enzyme of cholesterol biosynthesis, and inhibit the synthesis of cholesterol in the liver. Lovastatin (Mevacor®) Merck & Co., Inc., Whitehouse Station, NJ; was the first compound approved for lowering cholesterol in humans and became the cornerstone of all cholesterol-lowering agents generically called “statins.” The modification of 2-methylbutanoate to 2,2-dimethylbutanoate led to the semisynthetic derivative, simvastatin (Zocor®) Merck & Co., Inc., Whitehouse Station, NJ; 101), the second and more
effective agent approved for human use (139). Hydroxylation of compactin by biotransformation led to pravastatin (Pravachol™, BMS, NY) (139). The key pharmacophore of the statins is the \( \beta \)-hydroxy-\( \beta \)-lactone or open acid. As the importance and value of cholesterol lowering to human pathophysiology became clearer, the search for additional cholesterol-lowering agents became more prominent and led to the discovery and development of several other clinical agents. All these compounds retained nature’s gift of the pharmacophore, \( \beta \)-hydroxy-\( \beta \)-lactone (or open acid), with replacement of the decalin unit of the natural products with a variety of aromatic lipophilic groups that resulted in fluvastatin (105) (141), Cerivastatin (105, withdrawn from the clinic) (142), Rosuvastatin (105), and Pitavastatin (107). The statins have had tremendous impact in improvement of overall human health and quality of life because of the lowering of low-density lipoprotein (LDL) particles, which leads to a reduction in the incidence of coronary heart disease; arguably, they are the most successful class of medicines.

Ephedrine (108), isolated from the Chinese plant Ephedra sinica, was approved as one of the first bronchodilators and cardiovascular agents. This discovery led to a variety of such antihypertensive agents including \( \beta \)-blockers (4). An angiotensin-converting enzyme (ACE) converts angiotensin I to angiotensin II, and its inhibition has led to several very successful, clinically useful antihypertensive agents. Although these inhibitors are of synthetic origin, the original lead was modeled after a nonapeptide, teprotide (109). This peptide was isolated from snake (viper, Bothrops jararaca) venom by Ondetti et al. It had antihypertensive activity in the clinic by parenteral administration (138, 144, 145) but was devoid of oral activity. Ondetti and coworkers worked diligently, and, recognizing that ACE was a metallo-enzyme, they visualized the binding of a smaller snake-venom peptide SQ 20475 (110) with ACE; they modeled an acyl-proline with a sulfhydryl substitution at the zinc binding site, which led to the design and synthesis of captopril (111) as an orally active highly effective antihypertensive clinical agent. Additional application of the rational design by Patchett and coworkers led to the synthesis of enalapril (112) and other clinically relevant oral ACE inhibitors (138).

**Antiparasitic Agents**

Avermectins (113, 114) are a series of macrocyclic lactones that are broad-spectrum, highly potent, glutamate-gated, chloride channel-modulator antiparasitic agents produced by Streptomyces avermitilis (146, 147). Ivermectin, 23,24-dihydroavermectin \( B_1a/B_1b \) (115), was the first product approved in the mid-1980s for treatment of intestinal parasites in domesticated and farm animals, and it remains the standard of care (148, 149). The remarkable activity of ivermectin against Onchocerca volvulus, the causative parasite of onchocerciasis (river blindness), led to clinical development and the approval of Medizin™ Merck & Co., Inc., Whitehouse Station, NJ, for the treatment of such diseases. These parasitic diseases have debilitated millions of people in many countries in Africa and South America. Because Medizin™ is a very effective treatment, Merck is providing this drug free of cost to all people in need as a part of the “Medizin Donation Program,” which
Pharmaceuticals: Natural Products and Natural Product Models

Mycophenolic acid (98)

Tacrolimus (FK506) (96)

Rapamycin (97)

Cyclosporin A (95)

Lovastatin (99)

Fluvastatin (103)

Pravastatin (102)

Simvastatin (101)

Mevacor (100)

Nodulisporic acid A (118)

Latsensporic acid A (138)

Fluvastatin (103)

Rapamycin (97)

Nodulisporic acid A (138)

Spinosyns were discovered from the fermentation broth of Saccharopolyspora spinosa by screening for mortality of blowfly larvae, and a mixture of spinosyns A (136) and D (137) was approved and used successfully as a crop protection and antiparasitic animal health agent. (151) Nodulisporic acids are an indole diterpenoid class discovered from various species of Nodulisporium as orally active antiflea and antitick agents for dogs and cats (152, 153). The most active of the series is nodulisporic acid A (138), which selectively modulates the activity of insect-specific glutamate-gated chloride channels (153).

Pharmaceutical Models

The roles played by natural products as models for design and development of pharmaceutical agents are too many to cover in this overview. A few examples are illustrated during the discussions of specific disease areas above. For example, a...
Pharmaceuticals: Natural Products and Natural Product Models

Avemectin B1a, \( R = \text{CH(CH}_3\text{)}\text{CH}_2\text{CH}_3 \), \( \Delta \text{23,24} \) (113)
Avemectin B1b, \( R = \text{CH(CH}_3\text{)}_2 \), \( \Delta \text{23,24} \) (114)
Ivermectin, \( R = \text{CH(CH}_3\text{)}\text{CH}_2\text{CH}_3 + \text{CH(CH}_3\text{)}_2 \) (115)

Nodulisporic Acid A (118)
Spinosyn A, \( R = \text{H} \) (116)
Spinosyn D, \( R = \text{Me} \) (117)

marine sponge-derived nucleoside was the precursor for various nucleoside-based antiviral agents, pepstatin for renin and HIV protease inhibitors, snake venom peptide for ACE inhibitors, lovastatin andcompactin for all statins, and ephedrine for many painkillers and β-blockers. Below are a few critical examples that have played a big role in defining leads for some therapeutic areas but have not resulted in a drug yet.

Asperlicin (119) was isolated from Aspergillus alliaceus as a weak cholecystokinin A receptor (CCK-A) antagonist by using CCK receptor binding screening assays (154). It is a competitive antagonist of CCK-A (but not CCK-B) but did not have sufficient potency or oral activity to qualify as a drug candidate. In a remarkable strategy, medicinal chemists simplified the molecule to a benzodiazepine core of asperlicin, which led to the synthesis of potent, safe, and orally active analogs (120 and 121) with selectivity for either CCK-A (120) or CCK-B (122) receptors. They entered human clinical trials but were abandoned because of lack of efficacy (155, 156). The benzodiazepine scaffold was coined as “privileged structures” by Evans et al (155). This discovery is a beautiful demonstration of how a natural product became a model for the CCK program and played a pivotal role in defining the entire field (138).

The second example is apicidin (122), which is a cyclic tetrapeptide isolated from a fungus Fusarium pallidoroseum by using an empiric antiprotozoal screen (157, 158). It showed potent inhibition of apicomplexan protozoa including the malarial parasite Plasmodium falciparum and coccidiosis parasite Eimeria spp. It was effective in vivo against reducing malaria parasite infection in a mouse model (157) and exhibited strong activity against tumor cell lines (159, 160). Cyclic tetrapeptides with a terminal epoxy-ketone were known to be effective cytotoxic agents before the discovery of apicidin, but the pharmacophore was associated with the epoxy-ketone group (e.g., Trapoxin B, 123) with covalent modification as a mode of action. Apicidin does not contain the epoxy-ketone but showed potent antitumor activity (161). The mode of action of apicidin was shown to be the inhibition of histone deacetylase (HDAC) (157). The amino-oxo-decanoic acid (L-Aoda) mimics the acetylated lysine residue and positions itself at the zinc-binding site of HDAC (162). Chemical modification of apicidin with retention of the ethyl or methyl ketone led to the synthesis of small dipeptides (e.g., 124) that retained the HDAC and tumor cell line inhibitory activities with significant reduction of inhibition of normal cells (161).
 discovery of the fatty acid synthesis inhibitor antibiotic platensimycin (125) (163, 164) and platencin (126) (165, 166). The former shows exquisite selectivity for FabF, whereas the latter compound is a balanced inhibitor of both condensing enzymes, FabF and FabH.

References


84. Bollag DM, McQueney PA, Zhu J, Hensens O, Koupal L, Liesch
81. Propper DJ, McDonald AC, Man A, Thavasu P, Balkwill F,
79. Omura S, Sasaki Y, Iwai Y, Takeshima H. Staurosporine, a
78. Omura S. The expanded horizon for microbial metabolites–a
77. Sasaki Y, Seto M, Komatsu K, Omura S. Staurosporine, a
76. White RJ, Durr FE. Development of mitoxantrone. Invest. New
75. Hecht SM. Bleomycin: new perspectives on the mechanism of
74. Skorobogaty A, White RJ, Phillips DR, Reiss JA. The 5'-CA
70. Chaplin DJ, Pettit GR, Hill SA. Anti-vascular approaches to
69. Nabha SM, Mohammad RM, Dandashi MH, Coupaye-Gerard
68. Kirwan IG, Loadman PM, Swaine DJ, Anthoney DA, Pettit
55. Hecht SM. Bleomycin: new perspectives on the mechanism of
54. Skorobogaty A, Pettit GR, Al-Katib AM. Binding of dolastatin 10 to tubulin
52. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.
47. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.
44. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.
42. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.
40. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.
34. Baryza JL, Brenner SE, Craske ML, Meyer T, Wender PA.

Pharmaceuticals: Natural Products and Natural Product Models


Polyketide Biosynthesis, Fungi

Thomas J. Simpson and Russell J. Cox, School of Chemistry,
University of Bristol, Bristol, United Kingdom
doi: 10.1002/9780470048672.wecb458

Fungi produce a wide variety of biologically active compounds. Among these compounds, the polyketides form a large and structurally diverse group. These compounds are synthesized by highly programmed, large iterative multifunctional proteins, which are called the polyketide synthases. This review describes the structure and biosynthesis of polyketide fungal metabolites and highlights recent work on the links between gene sequence, protein architecture, and biosynthetic programming for fungal polyketide synthases.

Polyketides have long been recognized as one of the most important classes of secondary metabolites (1). They occur in plants, bacteria, and marine organisms as well as in fungi. Fungal polyketides vary from the simplest monocyclic aromatic compounds, for example, orsellinic acid and 6-methylsalicylic acid (6-MSA) to polycyclic aromatics such as citrinin, alternariol, islandicin, deoxyherqueinone, and norsolorinic acid. Although initially associated with the formation of aromatic compounds, many polyketides are nonaromatic (e.g., the macroclide deco restrictine and the long-chain polyfunctional molecules exemplified by T-toxin and the decalin, lovastatin, and compactin). Many other metabolites consist of an aromatic ring attached to a more highly reduced moiety (e.g., zearalenone, dehydrocurvularin, and monocerin). A different diversity results from extensive oxidative metabolism of preformed polyketide structures (e.g., penicillic acid and patulin), which are formed from cleavage and rearrangement of 6-MSA and orsellinic acid, respectively, and indeed ring cleavage is a very common feature with the potent hepatotoxic B1 and orsellinic acid, respectively, and indeed ring cleavage is a very common feature with the potent hepatotoxic aflatoxin B1, being derived by extensive ring cleavages and rearrangements of norsolorinic acid. Other metabolites contain a polyketide-derived moiety as part of a larger molecule whose biosynthesis is other than polyketide. A classic example is mycophenolic acid, in which the branched carboxylic side chain is derived via a cleaved farnesyl moiety. Other compounds of mixed terpenoid-polyketide origin include the mycotoxin viridicatumtoxin. A gain the origins of the polyketide-derived moiety may be disguised as a result of extensive metabolism as observed in the meroterpenoid metabolites azidin, paraherquonin, which are all derived via 3,5-dimethylorsellinic acid. The xenovulenes are an interesting group where it has been shown that the cyclopentanone, benzimid, and troponol moieties all have a common biosynthetic origin via ring expansion and ring contraction of 3-methylcyclopentadiene. Other groups contain Krebs’ cycle intermediates, for example, the tetrone acid, carlsic acid, and the squalestatins (see below) or amino-acid derived moieties (e.g., fusarins).

Biological Properties

A further important feature of fungal polyketides is their vast range of biological activities both beneficial and harmful. Thus, griseofulvin was one of the first effective antifungal agents. Penicillin was discovered shortly after the penicillin, is also a powerful antibiotic, but unfortunately it proved too toxic for clinical use. Mycophenolic acid has been “rediscovered” as an immunosuppressive agent. The statins (e.g., lovastatin, which was discovered shortly after the penicillin, is also a powerful antibiotic, but unfortunately it proved too toxic for clinical use. Mycophenolic acid has been “rediscovered” as an immunosuppressive agent. The statins (e.g., lovastatin), although not themselves used in the field formed the basis for the development of the widely used methoxyacrylate group of antifungal agents and, of course the statins, as represented by lovastatin, are among the most widely prescribed drugs for control of cholesterol levels and associated heart disease. As a cursory inspection of their structures would suggest, the squalestatins (e.g., lovastatin), although not themselves used in the field formed the basis for the development of the widely used methoxyacrylate group of antifungal agents and, of course the statins, as represented by lovastatin, are among the most widely prescribed drugs for control of cholesterol levels and associated heart disease. As a cursory inspection of their structures would suggest, the squalestatins (e.g., lovastatin) are effective inhibitors of squalestatin synthase though their early promise as cholesterol-lowering clinical candidates declined because of inherent toxicity. They contain two separate polyketide chains linked to oxaloacetate. Overall, fungal metabolites and their pharmaceutical and agrochemical derivatives have total sales of many tens of billions of pounds annually. In addition to these beneficial effects, the large group of mycotoxins represented first and foremost by aflatoxin B1, 17 and others such as the fuminosins, zearalenone, citrinin, ochratoxins (e.g., T-toxin), are the cause of many problems in both animal and human health, and spoilage of both growing crops and stored foodstuffs through contamination by mycotoxin producing fungi is a cause of major economic losses worldwide.
Polyketide Biosynthesis, Fungi

**Biosynthesis**

Although diverse in structure, the class is defined by the common biosynthetic origin of the carbon atoms: These atoms are derived from the CoA thioesters of small carboxylic acids, such as acetate and malonate. As long ago as 1953, Birch realized that polyketide biosynthesis is related to fatty acid biosynthesis, and some of the earliest applications of radioisotopes to natural product biosynthesis were to fungal polyketide metabolites, where the ease of fermentation and isolation of metabolites in pure form, and relatively efficient uptake of simple labeled precursors facilitated the work. In more recent years, fungal metabolites in general, and polyketides in particular, were the focus of the rapidly expanding applications of stable isotope labeling being greatly facilitated by the contemporaneous development of Fourier Transform methods and their application to 13C NMR. The application of doubly 13C-labeled precursors with analysis of regiospecificity of labeling being greatly facilitated by the contemporary development of Fourier Transform methods and their application to 13C NMR. The application of doubly 13C-labeled precursors led to the concept of bond labeling, which allowed investigators the mode of cyclization of linear polyketide precursors into polycyclic molecules, bond fragmentation, and rearrangements processes to be detected for the first time through analysis of the resulting 13C-13C coupling patterns. This method was then rapidly followed by applications of isotope induced shifts in 13C NMR, which allowed indirect detection of -H and 13C labels and the use of direct 1H NMR. These new methods allowed stereochemistry and regiochemistry of labeling to be detected and in particular permissible levels of oxidation and reduction in otherwise undetectable biosynthetic intermediates to be determined. Along with similar work with bacterial polyketides, this method laid the basis for the ideas of the processive mode of polyketide biosynthesis to be established. It was a major step because it changed fundamentally the idea that polyketide chains were assembled in their entirety and then subjected to necessary reductive modifications to the simple, in retrospect, idea that these changes occur concomitant with chain elongation rather than post-elongation. Although it is still not uncommon for these classic “Birch” fully oxygenated polyketide intermediates to be invoked it is now evident that in most polyketide metabolites, these have no reality. The concept of processive polyketide assembly and modification brought polyketide biosynthesis even closer to the process of fatty acid biosynthesis in which full reductive processing in each cycle of chain condensation and elongation is the norm. The rapid developments in understanding of the molecular genetics of polyketide biosynthesis particularly in bacteria in the 1990s were fully consistent with the processive mode. Understanding of the genetics of fungal polyketide biosynthesis still lags behind that of bacterial polyketides and the remainder of this article will provide a brief overview of current understanding.

**Polyketide assembly**

The basic assembly cycle for both polyketide and fatty acid biosynthesis is shown in Fig. 1 in which a starter unit, normally acetate is transferred to the ketosynthase (KS) or condensing enzyme which catalyzes a deacylative condensation with malonate, bound after after malonyl transferase (MT) catalyzed malonylation to the acyl carrier protein (ACP). During fatty acid biosynthesis, the resulting β-ketoester is subjected to additional chemical processing while attached to the terminal thiol of the ACP: first, it is reduced by a β-ketoacyl reductase (KR) to a secondary alcohol which then undergoes a dehydrogenase (DH)-catalyzed dehydration to form an α,β-unsaturated thioester, and finally enoyl reduction (ER) yields a fully saturated thioester. Fungal PKSs employ all these reactions, but additionally the chain can be methylated, using a methyl group from S-adenosylmethionine (SAM). This reaction probably occurs after KS, which gives an α-methyl-β-ketoester. During the biosynthesis of palmitic acid (C16), there are seven cycles of these reactions. The final reaction of FAS is hydrolysis of the thioester by a dedicated thioesterase (TE). Apart from the capacity for C-methylation, fungal PKSs in common with other PKSs have the ability to use the condensation and reductive cycle in a highly controlled manner to produce polyketide intermediates in which no reductive modification has occurred to give a classic poly-β-ketoester, or more generally “reduced” or “processive” polyketide intermediates in which a complete spectrum of reduction and/or C-methylation has occurred in each condensation cycle as indicated in Fig. 1 for the squaerylactide tetraketide synthase (see below). Another variation is that a range of alternate starter units can be used (e.g., hexanoate in the case of norsolorinic acid or benzoate in the case of the streptomycins). Enzymology

The understanding of the relationship between FAS and PKS enzymes, the application of molecular genetics, and more latterly genomics, has greatly facilitated the discovery and understanding of polyketide synthases from diverse sources. The homology in catalytic function between FAS and PKS enzymes is preserved in their respective gene sequences. It is now clear that fungal PKSs belong to the class of Type 1 iterative synthases represented by mammalian FAS. Type 1 FAS proteins are large multifunctional proteins in which single (or occasionally two) peptides contain the sequences for KS, ACP, AT, KR, DH, ER, and TE activities—these catalytic functions are carried out by particular functional domains. Similarly, the genes for Type 1 FAS proteins are correspondingly large single open reading frames, and Type 1 PKSs consist of very large multifunctional proteins with individual functional domains. Thus, PKSs use much the same array of chemical reactions as FASs—but the key differences is that of programming: FASs have to control chain length (i.e., the number of extensions), but PKSs can additionally control starter unit selection and the extent of reduction during each condensation cycle. Fungal PKSs can program the extent of chain methylation and the off-loading mechanism. The issue of programming is key to understanding and exploiting PKSs. In the case of the bacterial modular polyketide synthases, each condensation cycle is catalyzed by a discrete module containing all the catalytic domains required. In this case, the program is explicit in the order and composition of the modules. However, for the iterative Type 1 fungal polyketide synthases, the programme is cryptic—encoded in the PKS itself.
Polyketide Biosynthesis, Fungi

Figure 1

Generic polyketide assembly pathway reactions catalyzed by iterative fungal polyketide synthases. The assembly sequence for the squalesatin tetraketide intermediate 37 is shown for illustration.

Fungal Polyketide Synthases

Fungi make some of the simplest and some of the most complex polyketides known (3). It is useful to consider a hierarchy of complexity when considering the structures of fungal polyketides, because the complexity in chemical structure is generated by enzymes and ultimately genes. The simplest structures are those such as orsellinic acid 1. Addition of an extra acetate gives a pentaketide such as 1,3,6,8-tetrahydroxynaphthalene, which is a compound widely distributed in fungi and which is involved in melanization—a key component of appressorium formation and invasion of plant cells by plant pathogens such as Magnaporthe grisea, Colletotrichum lagenarium, and others.

Additional complexity is represented by compounds such as 6-MSA acid 2, in which a single, programmed reduction reaction occurs during biosynthesis. More reduction is then observed in compounds such as T-toxin 19 produced by Cochliobolus heterostrophus and lovastatin 10 produced by Aspergillus terreus. In these compounds, many more carbon atoms are used and many more reduction and dehydration reactions occur. Additionally, in many cases (e.g. fusarin C 26), pendant methyl groups have been added from the S-methyl of methionine, catalyzed by a C-methyl transferase domain.

Linking PKS genes and compounds in fungi

The huge structural variety of fungal polyketides is caused by differences in programming of their PKS proteins—apparent increases in structural complexity are caused by increasing use and control of reductive, dehydrative, and methylating steps by the PKS. This increase must be because of differences in PKS protein sequence and structure. This fact has been exploited in the development of rapid methods for the cloning of fungal PKS genes associated with the biosynthesis of particular fungal polyketide types.

Bingle et al. (3) realized that these subtle protein sequence differences should be reflected in DNA sequence, and polymerase chain reaction (PCR) primers could be designed to amplify fragments of fungal PKS genes selectively from fungal genomic DNA (or cDNA). In early work in this area, they hypothesized that fungal polyketides could be grouped into two classes: nonreduced (NR) compounds such as orsellinic acid 1, norsolorinic acid 7, and 1,3,6,8-tetrahydroxynaphthalene, and partially reduced (PR) compounds, such as 6-MSA 2. At the time, very few fungal PKS genes were known, and based on very limited sets of sequences, they designed degenerate PCR primers that were complimentary to conserved DNA sequences in the KS domains in fungal PKS responsible for the biosynthesis of NR and PR compounds. Later, the same analysis was extended to the KS domains of highly reduced (HR) compounds, such as lovastatin 30 when DNA sequence data became available for the lovastatin nonaketide and diketide synthases (LNKS and LDKS, respectively) (4). The availability of these sequences also allowed the development of selective PCR primers for CMeT domains.

This sequence analysis has been significantly extended as genomic approaches have been applied to fungi recently (5). Full genome sequences have now been obtained for more than a dozen fungi. In each organism, many PKS genes have been discovered. For example, Aspergillus niger contains 34 PKS genes, so several hundred fungal PKS genes are known. Sequence comparison of all these new PKS genes, however, shows that the three classes of fungal PKS genes predicted by Bingle et al. (3) are the same three classes observed in the most recent sequence comparisons (5). Despite the fact that so many fungal PKS genes have been discovered, however, relatively few genes have been definitively linked to the biosynthesis of specific metabolites. Because the NR, PR, and HR nomenclature is useful for describing both the chemical products and their cognate genes, the state of knowledge of fungal PKS is reviewed in this way.

Fungal NR-PKS

The tetraketide orsellinic acid 1 is the simplest tetraketide, which requires no reductions during its biosynthesis. One of the first discovered fungal PKS, orsellinic acid synthase (OSAS)
was isolated from Penicillium marneffei in 1968. Despite the early work with the protein, however, the OSAS-encoding gene has not yet been discovered, and nothing is known of the catalytic domains or their organization. However, genes involved in the biosynthesis of several other nonreduced polyketides are now known, and a general pattern of domain organization has emerged. In all known cases, these genes encode Type I iterative PKS proteins. At the N-terminus, a domain is present that seems to mediate the loading of a starter unit (Fig. 2a). It seems that the starter unit can derive from either a dedicated FAS, another PKS, or an acyl CoA. The starter unit loading domain (SAT) is followed by typical KS and AT domains responsible for chain extension and malonate loading. Beyond the AT is a conserved domain known as the product template (PT) domain. This domain is followed by an ACP. Some NR-PKS seem to terminate after the ACP, but many feature a diverse range of different domains that include cyclases, methyl transferases, and reductases. Thus, it seems that these synthases are arranged with an N-terminal loading component; a central chain extension component that consists of KS, AT, PT and ACP domains; and a C-terminal processing component.

NR-PKS loading component

Feeding experiments with isotopically labeled precursors have shown that many NR fungal polyketides are formed by the use of “advanced” starter units. In the classic case of norsolorinic acid biosynthesis, it has long been known that hexanoate forms the starter unit. Differential specific incorporation of acetate into the early and late positions in compounds such as citrinin have shown that many NR fungal polyketides are formed by the use of advanced starter units. For example it is now known (8) that two PKS genes are involved in the biosynthesis of zearealenone (Fig. 2b). In one case of this is a NR-PKS (see below) and probably provides a highly reduced hexaketide as a starter unit. The second zearealenone PKS is a NR-PKS possessing an N-terminal SAT domain, which likely loads the hexaketide ready for three further extensions. Most NR-PKS seem to possess potential SAT domains whether they require an acetate starter unit or not. For example, polyketide synthases involved in the biosynthesis of YWA1 (WA) and tetrahydroxynaphthalene (THNS). In the case of THNS from C. lagenarium, in vitro experiments have implicated that the purified protein uses malonyl CoA as the starter unit. The THNS SAT domain may therefore be involved with loading and decarboxylation of malonate to use as a starter unit, much as the bacterial Type 2 KS, component does.

NR-PKS chain extension component

The extension components of NR polyketide synthases consist of KS, AT, PT, and ACP domains. Sequence analysis of PT domains (450–550 residues) from a range of NR-PKS in which the chemical products are known suggests that it is conceivable that the PT domain is involved in chain-length determination. Comparison of the PT domains from the Acremonium strictum strain PKS1, citrinin PKS, zearealenone PKS-B, NSAS, sterigmatocystin PKS, dothistromin PKS, THNS from C. lagenarium and Wagneria dermatitis, and WAS suggests that these domains group into clades that correspond with chain length (9). The citrinin, ASPK51, and zearealenone-B groups correspond to tetrahydroxynaphthalene synthases; the NSAS, sterigmatocystin, and dothistromin PKS are all octaketide synthases; the THNS are hexaketide synthases; and the WAS synthase (WAS) forms an outgroup, which is a heptaketide.

The domains found after the chain extension components at the C-terminus of fungal NR-PKS are highly variable. These components include putative Claisen-cyclase/thioesterases (CLC/TE), C-methyl transferases (C-MtT), reductases (R), and additional ACP domains. Recent work indicates that these processing components act after chain assembly to modify either a poly-keto or a cyclized intermediate.

CLC/TE domains

The first fungal NR PKS gene to be cloned was Aspergillus nidulans wa (encoding WAS). Limited domain analysis was carried out to determine the presence of KAS, AT, and ACP domains. Later, it was shown that wa also possesses a TE domain, which forms one of the most common processing components of NR PKS. The TE domain can either operate as a standard thioesterase or be involved in a cyclization-release mechanism. Fujii et al. (10) expressed A. nidulans wa (encoding...
Polyketide Biosynthesis, Fungi

**Figure 2.** (a) General architecture of NR PKS genes; proposed mechanism of WAS CLC domain, (b) the Monascus purpureus citrinin PksCT and A. strictum MOS synthases; and (d) domain architecture of A. terreus and P. patulum MSAS, and proposed mechanism of 6-MSA 2 and TAL 44.

WAS in Aspergillus oryzae. Initially, the expression strain produced the isocoumarin 38, which indicated that WAS is a heptaketide synthase (Fig. 2b). It was then realized that the expression construct used in this experiment had a deletion that resulted in the expressed WAS missing the final 67 amino acids of the C-terminal TE domain. When the complete was gene was expressed, however, the heptaketide naphthopyrone YWA139 was produced (10). A series of experiments that involve step-wise shortening of the C-terminus of WAS showed that deletion of as few as 32 amino acids resulted in production of the isocoumarin 38. Site-directed mutagenesis of a conserved serine and histidine in the C-terminal domain also resulted in a switch from naphthopyrone production to citreoisocoumarin production.

Isotopic feeding experiments using 13C-labeled acetate indicated the folding pattern shown in Fig. 2b for the naphthopyrone 39. Thus, both the isocoumarin 38 and the naphthopyrone 39 must result from the cyclization of the common intermediate 40. This finding suggests that the WAS chain extension component produces a heptaketide and catalyzes the cyclization and aromatization of the first ring. The C-terminal domain must therefore catalyze a second (Claisen) cyclization reaction to...
form. The involvement of conserved serine and histidine residues suggests involvement of the CLC-bound intermediate 43 shown in Fig. 2b. Thus, the TE domain has been renamed as C-MeT domain. No obvious catalytic machinery exists for off-loading the product. The DNA sequence of the KS domain is distin-
guished from those of the other domains in the PKS and seems not to require a TE/CLC domain. Some HR PKSs possess an ER domain, but those that do not lack a fully functional ER domain. The PKS often terminates with an ACP domain. No obvious catalytic machinery exists for off-loading the product. The DNA sequence of the KS domain is distin-
guished from those of the other domains in the PKS and seems not to require a TE/CLC domain. Some HR PKSs possess an ER domain, but those that do not lack a fully functional ER domain.

C-MeT domains

Few NR PKSs are known to possess C-methylation domains, although many known fungal nonreduced polyketides are C-methy-
lated, such as 3,5-dimethylorsellinic acid. A small group of NR PKSs have been identified in genome sequences (5), which feature a C-MeT domain located after the ACP (Fig. 2c). The first correlation between a gene sequence and a compound came from Monascus 3, in which the PKS involved in citrinin biosynthesis has been isolated (11). Here, the C-MeT domain must be programmed because it acts twice during polyketide biosynthesis when a probable methylated diketide starter unit is extended. It is not yet clear whether the C-MeT domain acts during extension, after chain extension but before aromatization, or after aromatization. 1,3-Di-hydroxynonamethoxan is known to tautomerase easily to keto forms, and it is conceivably that it could act as the nucleophile for the reaction with SAM.

R domains

Reductases are currently rare as part of the processing compo-
nent of NR PKS. Evidence for the role of these reductase domains has been obtained by the isolation of the PKS respon-
sible for the formation of the tetraketide component found in xenovulene A (xenovulene A). The PKS gene (MOS) was found to have SAT, KS, AT, C-MeT, and R domains (Fig. 2d), and heterolo-
gous expression of the gene in Aspergillus oryzae resulted in high yields of 3-methylorcinolactone 45. Although not described in the literature, sequence analysis of the citrinin PKS sequence discussed above shows that it also possesses a C-terminal thi- olester reductase domain.

Similar domains are known from NRPS systems in which reductase domains are sometimes used as chain release mecha-
nisms, which release an aldehyde or primary alcohol. In the case of MOS and citrinin biosynthesis the reductive release mecha-
nism makes good sense as this provides the products with C-1 at the correct oxidation state (Fig. 2c).

Fungal PR-PKS

The domain structure of PR PKSs is much closer to mammalian FAS, with an N-terminal KS followed by AT, and DH domains (Fig. 2b). A so-called “core” domain follows the DH, which is followed by a KR and the PKS terminates with an ACP do-
main. The domain structure differs considerably from the NR PKS—no SAT domain or PT domain exists, and the PKS ter-
minishes after the ACP and seems not to require a TE/CLC domain. No obvious catalytic machinery exists for off-loading the product. The DNA sequence of the KS domain is distin-
guished from those of the other domains in the PKS and seems not to require a TE/CLC domain. Some HR PKSs possess an ER domain, but those that do not lack a fully functional ER domain. The PKS often terminates with an ACP domain. No domain similar to the PT domain of the NR PKS or the core domain of the PR PKS is known.
PKS seems to exist, and there is no N-terminal SAT domain as found in the NR PKS. As with the NRs and PR PKSs, many HR PKS genes are known from the numerous fungal genome sequences, but as yet few gene sequences have been linked to the production of known compounds. However, of the few cases in which both gene and chemical product are known, some progress has been made in understanding function and programming.

The lovastatin polyketide synthases

Lovastatin (also known as mevinolin) is produced by A. terreus. The related compound compactin (mevastatin) is produced by Penicillium citrinum and is identical to lovastatin apart from the C-12 methyl group absent in compactin. Isotopic feeding experiments have shown that two polyketide chains are required: a nonaketide and a methylated diketide. The requirement for two polyketide synthases is evident in the gene clusters associated with biosynthesis ofLovastatin and Compactin where two PKS genes are found (4), lovB and lovF, which encode LNKS and LDKS in the case of lovastatin (Fig. 3B). LNKS formally should synthesize a fully elaborated nonaketide such as 47, which could undergo a biological Diels-Alder reaction to form dihydromonacolin L 48—which is the observed first PKS-free intermediate. Note that it is possible that the Diels-Alder reaction actually occurs at the more activated hexaketide stage. When lovB was expressed in the heterologous fungal host A. nidulans (4) however, the polyunsaturated compounds...
Alternaria solani produces a wide range of bioactive compounds derived from polyketides fused to amino acids. Examples include fusarin C, equisetin 54, and tenellin 55. Fusarin C 26 consists of a tetramethylated heptaketide fused to homoserine and is produced by strains of the plant pathogens Fusarium moniliforme and Fusarium venenatum. Genetic DNA libraries from these organisms were used to isolate a gene cluster centered around a 12-Kb ORF encoding a PKS fused to a nonribosomal peptide synthetase (NRPS) module 19. The PKS region is homogeneous to LNSK: KAS, AT, and DH domains are followed by CMeT, a defective ER, KR, and ACP domains. Like LNSK, SQTKS possesses a functional ER domain, but SQTKS carries out three modifications reactions occur only after the final extension (Fig. 1). HR PKS from Alternaria solani

Alternaria solani is a plant pathogen and the causative agent of early blight in solanum species. It produces numerous polyketides, such as solanopyrone A 56 and alternaric acid 57, and it is thus an ideal target species for speculative PKS gene-fishing expeditions. Fuji et al. (18) have conducted just such investigations, using PCR primers based on conserved PKS sequences as probes with genomic DNA libraries (18). An early investigation yielded two hits—one a HR PKS gene named alt5, and another, a NR PKS named psla. The alt5 gene encodes a typical HR PKS, which is known as PKS A5, with the usual array of catalytic domains. Inspection of the ER sequence suggested that it should be functional like those from LDKS and SQTKS. Expression of alt5 showed this to be correct—a single compound was synthesized in good yield (7–15mgg⁻¹), which proved to be the octamethylated deacetate pyrone 53, named alternaric acid. This compound is the most complex polyketide detected so far, and it is programmed to elongate the ACP chain length control, keto-reduction, methylation, and enoyl reduction. HR PKS-NRPS

Fungi produce a wide range of bioactive compounds derived from polyketides fused to amino acids. Examples include fusarin C, equisetin, and tenellin. Fusarin C 26 consists of a tetra- or pentamethylated heptaketide linked to homoserine and is produced by strains of the plant pathogens Fusarium moniliforme and Fusarium venenatum. Genetic DNA libraries from these organisms were used to isolate a gene cluster centered around a 12-Kb ORF encoding a PKS fused to a nonribosomal peptide synthetase (NRPS) module 19. The PKS region is homologous to LNSK: KAS, AT, and DH domains are followed by CMeT, a defective ER, KR, and ACP domains. Like LNSK, the ACP is upstream of an NRPS condensation (C) domain, but in this case the NRPS module is complete, which features downstream adenylation (A), thiolation (T), and C-terminal thiolester reduce (R) domains (Fig. 1). Directed knockout of the PKS-NRPS gene proved it to be involved in the biosynthesis of 26 and was thus named fusarin synthetase (FUSS). The functional ER domain and the fact that no lovC homolog seems to exist in the cluster, is consistent with the polyunsaturated nature of the polyketide moiety.

It is probable that FUSS assembles a tetramethylated heptaketide 50 attached to the ACP (Fig. 3c) — the structure of which is...
similar to the heptadiketide pyrone \textsuperscript{45} produced by LNKS in the absence of the LovC protein. In parallel, the C domain of the NRPS module seems to select, activate, and attach homoserine \textsuperscript{57} to the thiolation domain. The C domain then uses the amide of homoserine to form an amide with the ACP-bound polyketide, which forms a covalently bound intermediate peptide \textsuperscript{58}. The final reaction catalyzed by FUS may be the reductive release of the thiolester, which forms peptide aldehydes \textsuperscript{59}. Finally, Knoevenagel cyclization would give the putative preufusarin \textsuperscript{60}. Other genes in the FUS cluster are presumably responsible for the required additional transformation of \textsuperscript{60} to fusarin C, \textsuperscript{26} itself: epoxidation; oxidation of a pendant methyl to a carboxylate and esterification; and hydroxylation a to nitrogen.

A highly homologous PKS-NRPS gene has been shown to be involved in the biosynthesis of equisetin \textsuperscript{54} in Fusarium heterosporum \textsuperscript{(20)}. EFS possesses the same catalytic domains as FUS, but examination of the structure of \textsuperscript{54} indicates that the pyrollidinone carbon derived from the carboxylate of the amino acid (serine in this case) is not reduced, which indicates either a reoxidation mechanism, or the fact that the R domain does not produce an aldehyde intermediate in this case.

Conclusion

Recognition that the fungal PKS can be categorized into three subsets has allowed more detailed consideration of the programming elements. The NR PKS are arranged into loading, extension, and processing components, and to some extent this hypothesis has been verified by expression and study of individual catalytic domains and by the construction of hybrid NR PKS genes. However, the programming elements of the FR and HR PKS remain obscure. The first experiments to probe programming in HR PKS have involved domain swaps, but few conclusions have yet been drawn. Little information is known about the three-dimensional structure of fungal PKS—whereas sequence and domain organization similarity with mammalian FAS mean that broad descriptions of the architecture of the catalytic domains can be modeled, detailed hypotheses that involve individual domains or peptide motifs cannot yet be linked with programming.

References


Further Reading

For general review of polyketides:
For comprehensive listings of fungal polyketide molecules:
Polyketide Biosynthesis, Fungi


See Also

Polyketide Biosynthesis, Enediyne Polyketides
Polyketide B Biosynthesis, Aromatic Polyketides
Polyketide B Biosynthesis, Modular Polyketide Synthases
Polyketide B Biosynthesis, Polyethers
Polyketides as Drugs
Polyketide Biosynthesis, Modular Polyketide Synthases
Tonia J. Buchholz, Jeffrey D. Kittendorf and David H. Sherman, University of Michigan, Ann Arbor, Michigan
doi: 10.1002/9780470048672.wecb459

Polyketides constitute a large class of microbial and plant-derived secondary metabolites that displays a vast array of structural diversity. These organic molecules vary in molecular weight, functional group modification, and include linear, polycyclic, and macrocyclic structural forms. Currently, polyketide natural products find clinical use as antibiotics, antiparasitic agents, antifungals, anticancer drugs, and immunosuppressants. Given these impressive and wide-ranging pharmacologic activities, an ever-increasing demand is placed on natural products research to uncover novel polyketide metabolites for the benefit of human and animal health. Modular polyketide synthases are nature’s platform for the expansion of chemical diversity. This review provides new perspectives on important biosynthetic mechanisms that contribute to this variety. This includes control of double-bond configuration and regiochemistry, introduction of β-branching during polyketide chain assembly, and other processes that contribute to introduction of unique chemical functionality into these fascinating systems.

Despite the promise of modern synthetic technologies to enhance pharmaceutical discovery significantly, natural products continue to be the greatest source of all new drug leads (1). Currently, many examples of natural product-derived pharmaceuticals are employed to benefit human health (2, 3); polyketides constitute a large class of microbial and plant-derived secondary metabolites that displays a vast array of structural diversity. These organic molecules vary in molecular weight and functional group modification; they include linear, polycyclic, and macrocyclic structural forms. Currently, polyketide natural products find clinical use as antibiotics, antiparasitic agents, antifungals, anticancer drugs, and immunosuppressants (Fig. 1a). Given these impressive and wide-ranging pharmacological activities, an ever-increasing demand is placed on natural products research to uncover novel polyketide metabolites for the benefit of human health.

Although the clinical use of polyketide-inspired pharmaceuticals has been appreciated for decades, polyketide-derived metabolites have been recognized recently for their role in bacterial virulence. For example, the pathogenesis of Mycobacterium ulcerans, the causative agent of the devastating skin disease known as Buruli ulcer, is the result of the secretion of polyketide-derived toxins known as the mycolactones (Fig. 1b) (4). These polyketide toxins are responsible largely for the necrotic lesions that are characteristic of this debilitating condition. As such, the disruption of mycolactone biosynthesis may lead to an effective chemotherapy for Buruli ulcer. Recent findings also suggest that the virulence of another mycobacterial species, Mycobacterium tuberculosis, could be partially dependent on polyketide biosynthesis. The cell surface sulfolipid-1 (SL-1) is among several virulence-associated molecules produced by M. tuberculosis. SL-1 consists of a sulfated disaccharide core (trehalose-2-sulfate) that displays four lipidic substituents; all but one substituent seems to be polyketide-derived (Fig. 1b) (5). Finally, the toxic agent in rice seedling blight, which is a highly destructive fungal disease that inflicts severe agricultural losses worldwide, has been identified recently as the polyketide metabolite, rhizoxin (Fig. 1b) (6). Interestingly, rhizoxin is not produced directly by the fungus (Rhizopus), but rather by the endosymbiotic bacteria Burkholderia that thrives within the fungus. Together, these three examples suggest that inhibition of polyketide biosynthesis may lead to effective chemotherapy for controlling certain human and plant bacterial diseases.
Polyketide Biosynthesis, Modular Polyketide Synthases

Prototypical Polyketide Biosynthesis

The biosynthesis of many important polyketide compounds occurs via a stepwise, assembly-line type mechanism that is catalyzed by type I modular polyketide synthases (PKSs). These modular PKSs are composed of several large, multifunctional enzymes that are responsible for catalyzing the initiation, elongation, and processing steps that ultimately give rise to the characteristic macrolactone scaffold (Fig. 2) (7-11). Structural studies have been critical in developing a sophisticated understanding of the overall architecture and mechanism of type I PKSs and their homologs in recent years (12-18). A review from the perspective of the 6-deoxyerythronolide B synthase (a well-studied type I PKS) was published recently by Khosla et al. (19).

It is well established that the sequential arrangement of modules within a PKS system serves effectively as a biosynthetic program, which is responsible for dictating the final size and structure of the polyketide core. Typically, initiation of polyketide biosynthesis begins by the acyltransferase (AT) catalyzed linkage of a coenzyme A (CoA) priming unit (e.g., methylmalonyl-CoA, malonyl-CoA, propionyl-CoA) to the acyl carrier protein (ACP) of the loading module. Once initiated, downstream elongation modules carry out repetitive extensions.
of the starter unit. In most PKS systems, each elongation mod-
ule contains at minimum an AT domain, an ACP domain, and a ketosynthase (KS) domain (Fig. 2a). The AT domain is re-
ponsible for loading the appropriate CoA extender unit onto the ACP domain (i.e., malonyl-CoA, methylmalonyl-CoA, etc.). The KS domain then catalyzes a decarboxylative condensa-
tion of the extender unit with the growing polyketide chain obtained from the preceding module to generate an ACP-bound β-ketoacyl product. In addition to the three core domains, each elongation module may contain up to three additional domains (ketoreductase [KR], dehydratase [DH], enoyl reductase [ER]) that are responsible for the reductive processing of the β-keto functionality prior to the next extension step (Fig. 2a). These reductive steps contribute greatly to the overall structural di-
versity that is observed among polyketide natural products. The presence of a KR domain alone generates a β-hydroxy function-
ality, the presence of both a KR and a DH domain generates an alkene, whereas the combination of KR, DH, and ER re-
results in complete reduction to the alkane. Finally, termination 
of polyketide biosynthesis is catalyzed by a thioesterase (TE) domain located at the carboxy terminus of the final elongation module. The activity of this domain results in the cleavage of
the acyl chain from the adjacent ACP; typically, intramolecular cyclization results in the formation and release of a macrocyclic ring. Tailoring enzymes, such as hydroxylases and glycosyl transferases, often serve to further modify the polyketide to yield the final bioactive compound.

The modular organization of type I PKSs has made them particularly attractive targets for rational bioengineering. Combinatorial biosynthetic efforts centered on prototypical modular PKSs have been the topic of many recent outstanding review articles (20–23). Currently, several strategies are being pursued that attempt to leverage PKS systems for the generation of structurally diverse polyketides. For example, it has been demonstrated that alterations of individual catalytic domains (i.e., inactivation, substitution, addition, deletion) within a PKS module can result in predicted structural alterations of the final PKS product. Likewise, the addition, deletion, or exchange of intact modules can also impart structural variety into polyketide metabolites. Using these and other approaches, hundreds of novel polyketide structures have been generated, which established the tremendous potential of these applications. However, these successes seem to be more the exception than the rule, as many efforts result in trace levels, or they fail to provide the desired metabolite. This finding suggests that much remains to be learned regarding the molecular intricacies of these complex biosynthetic machines. This review provides new perspectives on important mechanisms that contribute to structural diversity in modular PKSs. These mechanisms include control of double-bond configuration and regiochemistry, introduction of β-branching during polyketide chain assembly, and other processes that contribute to introduction of unique chemical functionality into these fascinating systems.

Polyketide Double Bonds

Trans double bonds

The presence of unsaturated carbon–carbon bonds within most polyketide compounds exemplifies the overall structural diversity that is a hallmark of this class of important natural products. Typically, the installation of double bonds into nascent polyketide chains relies on the two-step processing at the β-keto group by the successive activity of KR and DH domains that are embedded within a given PKS elongation module. After KS catalyzed chain elongation that extends the growing chain by two carbon atoms, the KR domain, when present, directs the NADPH-dependent reduction of the β-ketone to yield a 3-hydroxyacyl intermediate. Subsequently, an embedded DH domain within the elongation module catalyzes dehydration of the 3-hydroxyacyl intermediate, normally which results in the incorporation of an (E)-trans unsaturated bond into the growing polyketide chain (Fig. 3a). It should be noted that the KR catalyzed reduction of a β-ketoacyl intermediate has stereochemical consequences because a new chiral center is introduced into the growing oligoketide. A side from serving to enhance the structural diversity of the final polyketide product even more, the stereochemical outcome of this reaction can have profound effects on any subsequent processing or elongation reactions. As such, an understanding of how KR domains exert stereochemical control of their hydroxylated product is a critical aspect to deciphering the mechanism of DH-mediated double bond formation. Through bioinformatic and biochemical analyses, an appreciation of ketoreductase-influenced stereochemistry has emerged (24, 25). Thus, Caffrey (25) has proposed that KR domains can be divided into two classes, depending on the final configuration of the β-hydroxyl moiety. The so-called “A” class generates an L-3-hydroxy product, whereas the “B” class produces the D-3-hydroxy polyketide intermediate. A thorough little difference exists between these two putative classes at the amino acid sequence level, the presence of a conserved aspartate residue within an LDD motif correlates well with “B” class. This motif is absent in the defined “A” class of KR domains. An additional diagnostic feature of the “A” class of KR domains is the presence of a conserved tryptophan residue. Recently, Keatinge-Clay (18) has proposed a refinement of the KR class descriptions as originally suggested by Caffrey (25), effectively increasing the number of possible KR types from two to six (18). This new classification takes into consideration whether a given KR domain (either reductively competent or incompetent) is located in an epimerization competent module. Although this new classification offers a more complete description of PKS KR domains, for simplicity we will continue to use Caffrey’s KR nomenclature throughout our discussion of double-bond formation.

While examples of both D- and L-hydroxy group configurations can be found within polyketide natural products, recent evidence suggests that DH domains require a stereospecific 3-hydroxyacyl intermediate. Bioinformatic analyses performed on 71 KR domains for which the stereochemical outcome of the reduction is cryptic because of subsequent dehydration revealed that all belong to the “B” class of KR domains (25). As such, it appears that the generally preferred substrate for DH domains is a D-3-hydroxyacyl chain. However, direct experimental evidence has been difficult to obtain because the 3-hydroxyacyl intermediate is transient in modules that contain a DH domain. Recently, bioinformatic studies of the DH domain found in module 2 of the pikromycin PKS system (Fig. 2b) (26) have supported this hypothesis; inactivation of the DH domain resulted in the exclusive generation of the D-3-hydroxyacyl thioester intermediate from a diketide substrate (27). A side from this study, no other reports probe the substrate preference or catalytic mechanism of DH domains within PKS systems; therefore, much of what is known has been elucidated from studies of fatty acid biosynthesis (28). Previous studies on the dehydration step that is catalyzed by the yeast fatty acid synthase confirmed the syn elimination of water from a D-(3R)-hydroxyacylthioester substrate (29). This result is consistent with the stereospecificity of the PKS DH domain and may suggest that trans unsaturated bonds, typically found in polyketides, are likewise formed via syn water elimination.

cis double bonds

Although rare, several PKS biosynthetic systems can install cis double bonds into the final polyketide product (Fig. 4). Several possible mechanisms could account for the infrequent
Figure 3  (a) Traditional view of reductive processing at the β-ketone position in the growing polyketide chain. Presence of a ketoreductase domain leads to formation of an alcohol. An active dehydratase domain can further process the alcohol moiety to an alkene. Complete saturation to the alkane is accomplished by an enoyl reductase domain. (b) Proposed terminal double bond formation for curacin biosynthesis. (c) Proposed terminal double bond formation in tauromycin biosynthesis.

occurrence of this double bond configuration. One explanation is that an isomerization event occurs that converts a trans double bond into a cis double bond. This isomerization activity could be specified by the PKS elongation module, much like previously identified epimerization activities that are known to exist in some PKS and NRPS modules. Alternatively, the combined activity of KR-DH domains within certain modules could directly establish the cis double bond. Finally, it is possible that after reduction of the β-keto functionality, a trans acting DH could catalyze dehydration to form a (Z)-cis double bond.
bond. This trans activity might derive from a discrete enzyme encoded within the biosynthetic gene cluster or from an adjacent module within the PKS pathway. Additionally, examples exist in which some general rules may not hold true. Chivosazol, which is a potential antitumor agent, is one such example (30). The epothilones 6 (and Fig. 1a), which are produced by Sorangium cellulosum, are mixed NRPS-polyketide derived natural products that possess potent antitumor activity. Interestingly, some compounds feature a cis double bond between carbon atoms 12 and 13 that should be generated by PKS elongation module 4; however, sequence analysis of the epothilone biosynthetic gene cluster indicates that module 4 does not contain a DH domain requisite for the formation of the unsaturated bond (31). Thus, Tang et al. (31) hypothesized that the DH activity might occur from the subsequent module or by the
action of a post-PKS modifying enzyme. Biochemical experiments later demonstrated that the DH domain of module 5 catalyzed the cis double bond formation (32). To account for this atypical activity, it is proposed that the 3-hydroxythioester intermediate undergoes an ACP₄-To-ACP₅ transfer (32). A decarboxylation, the thioester intermediate would then be transferred to the KS₅ domain for subsequent elongation. Similarly, the an-
titumor phoslactomycin compounds (33) feature three cis double bonds, two of which seem to be installed by a KR-DH pair (see below); however, the elongation module (Pim₇) that should be responsible for generating the unsaturated bond between carbon atoms 2 and 3 does not appear to encode the required DH activ-
ity (33). Thus, it is likely that the source of this atypical activity comes from either a different module within the PKS system or a separate enzyme that could act either before or after TE medi-
ated termination of polyketide biosynthesis. A partial analysis is required to discriminate between these two possibilities.

Unlike the examples described above, most polyketide cis double bonds are installed through a successive KR-DH pair found embedded within the elongation module. In these cases, the stereochemistry of the 3-hydroxyacyl intermediate appears to be the discriminating factor between (Z)-cis or (E)-trans unsaturation. For example, the antitumor disorazolic compounds (34) display up to three cis double bonds per monomer (note: final compound is a condensed dimer). Sequence analyses of KR domains preceding DH domains that would be responsible for cis double bond formation suggest that they all belong to the "A" class (34). Thus, they are predicted to generate a L-3-hydroxyacyl intermediate. It is expected that the subsequent DH domain preferentially recognizes the L-3-hydroxyl group to facilitate the generation of the cis double bond. Interestingly, the module responsible for incorporating the cis unsaturated bond between carbon atoms 11 and 12 is split between two polypeptide chains (36). It cannot be ruled out that this modular dissection may play a role in formation of this particular double bond, as the cleavage point occurs between the DH and KR domain. Furthermore, it is intriguing that the major product, disorazole A₁, is composed of two nonidentical monomers that differ in saturation between carbon atoms 5 and 6. It has been suggested that the synthesis of the two different monomers is caused by poor activity of the DH domain (34); however, final proof requires additional experimental verification.

In addition to disorazole, several other known examples of polyketide natural products exist that have cis double bonds installed by an embedded KR-DH-DH domain pair. The potent antitumor compound curacin A (35) contains many interesting structural features. Among them is the presence of a cis double bond between carbon atoms 3 and 4. Sequence alignment of the KR domain encoded by curC (encoding the module responsible for generation of the cis double bond) suggests that it belongs to the "A" class of KR domains (35). Thus, this particular KR is predicted to generate a L-3-hydroxyacyl intermediate that is subsequently dehydrated to the cis double bond. Likewise, the KR domains that set up the cis double bonds found in the linear mixed NRPS-polyketide natural product bacillibrevin 1 are predicted to produce L-3-hydroxyacyl intermediates (36). Interestingly, two of the elongation modules that incorporate cis olefins are split between two polypeptides (36). As in disora-
zole biosynthesis, it is possible that these modular dissections contribute to the configuration of the unsaturated bond that is introduced by these modules.

Finally, we consider the conjugated cis olefins that span car-
bon atoms 12-13 of the phoslactomycin (33). An analysis of the K.R. sequences of elongation modules 1 and 2 could not clearly predict whether these reductive domains belonged to the "A" or "B" class. Therefore, Alhamadsheh et al. (37) genetically inactivated both the loading module and elongation module of Pim₁ and conducted feeding experiments with dithiolane analogs contain-
ing both cis and trans olefins. Results from this work indicated clearly that only the cis olefin containing dithiolane is accepted as a substrate for elongation module 2, suggesting that the product of module 1 must contain the cis double bond. Furthermore, this work demonstrated nicely that the phoslactomycin biosynthetic pathway cannot process trans dithiolane intermediates into ma-
ture products, which rules out the possibility of an isomerization domain in downstream modules.

Terminal double bonds

Termination of polyketide biosynthesis typically involves the TE mediated cleavage of the ACP₃-bound thioester, followed by cyclization to generate a macrocyclic lactone. Alternatively, the TE catalyzes the simple hydrolysis of the thioester to generate a linear free acid product. Here, we consider two of the relatively few known examples of polyketide natural products that are neither a macrocyclic nor a free acid, but instead terminate with a double bond.

Aside from containing a cis double bond noted above, the antitumor polyketide compound curacin A (3) also features a terminal olefin. Previously reported feeding studies suggested that the formation of the terminal double bond develops from successive decarboxylation and dehydration events (35). The biosynthetic gene cluster responsible for curacin A biosyn-
thesis has been identified and initially characterized, which enables the putative assignment of domains within the pre-
dicted elongation modules (35). Like most other known polyke-
tide biosynthetic pathways, the final elongation module of the curacin pathway, CurM, contains a terminal thioesterase do-
main that presumably plays a role in formation of the terminal olefin; however, biochemical evidence for such a role is lack-
ing. Interestingly, domain analysis of CurM also suggested the presence of a sulfotransferase (ST) domain immediately pre-
ceding the TE domain. Typically, ST domains are responsible for transferring a sulfuryl group from a donor molecule (such as 3'-phosphoadenosine-5'-phosphosulfate, PAPS) to a variety of acceptor carbohydrates, proteins and other low-molecular weight metabolites (38). Although STs have been characterized...
from both eubacterial and eukaryotic organisms, the presence of an ST domain within a PKS system is unprecedented.

Current efforts in our laboratory include elucidating the roles of the ST and TE domains in curacin A biosynthesis, and in particular, their potential functions in terminal olefin formation. One possible mechanism that is currently under consideration is shown in [Image 3b]. On reduction of the β-keto functionality of the ACP-bound thioester intermediate, the ST domain transfers a sulfuryl group from the donor molecule PAPS to the 3-hydroxyl group of the thioester chain. Consistent with this hypothesis, bioinformatic analysis suggests that the putative curacin ST domain contains the signature PAPS binding pocket (unpublished data); however, no experimental evidence suggests that this ST domain can catalyze the sulfuryl transfer or the formation of the terminal olefin product. Assuming our hypothesis is correct and that the ST domain functions as proposed, transfer of the sulfurylated intermediate to the TE domain would initiate hydrolytic termination of curacin A biosynthesis to produce the linear free acid. At this point, one of several chemical steps can be envisioned. Following hydrolysis, the TE may catalyze decarboxylation, after which the formation of the double bond would occur in a concerted process by displacement of the sulfate leaving group. Alternatively, a separately encoded enzyme might be responsible for decarboxylating the free acid generated by the TE domain. It is also conceivable that on TE catalyzed hydrolysis, the decarboxylation reaction occurs spontaneously due to the presence of the sulfate leaving group at carbon 3.

Similar to curacin, the polyketide metabolite tautomycetin 34 also possesses a terminal olefin. This polyketide metabolite has potential medicinal value because of its novel immunosuppressive activities (39). The tautomycetin biosynthetic gene cluster has been sequenced recently, which enables domain composition analysis of the terminal elongation module (40). Unlike the terminal module involved in curacin biosynthesis, the final tautomycetin elongation module does not contain the unusual ST domain. This finding may suggest that formation of the terminal olefin of tautomycetin 34 occurs by a different chemical mechanism. The final elongation module does contain a TE domain, which presumably terminates tautomycetin biosynthesis through generation of the free acid (Fig. 3c). As described for curacin, it is possible that this TE domain can also catalyze the subsequent decarboxylation event; however, in lieu of an activated leaving group at carbon 3, it is reasonable to expect that dehydration to remove the hydroxyl at carbon 3 would also be a catalyzed event. Alternatively, the terminal olefin could be installed during the post-PKS maturation of the polyketide to the final tautomycetin product. DNA sequence analysis of open reading frames that are downstream of the tautomycetin PKS gene cluster reveals two potential candidates, tncj and tncm, which may be involved in double bond formation (40). Bioinformatic analyses suggest that tncj might encode for a putative decarboxylase and that the gene product of tncm is a potential dehydratase.

**Atypical PKS Domains**

The wide distribution of PKSs in the microbial world and the extreme chemical diversity of their products do in fact result from a variety of the well-known catalytic domains described above for the canonical PKS systems. Taking a more theoretical view of polyketide diversity, González-Legier et al. (41) have suggested that even if the starter and extender units are fixed, over 100,000 linear heptaketide structures are possible using only the 5 common reductive outcomes at the β-carbon position (ketone, (R- or S-) alcohol, trans double bond, or alkane). Recently, it has become apparent that even this does not represent the upper limit for polyketide diversification. To create chemical functionalities beyond those mentioned above, nature has recruited some enzymes from sources other than fatty acid synthesis (the mevalonate pathway in primary metabolism is one example) not typically thought of as type I PKS domains. Next, we explore the ways PKS-containing systems have modified these domains for the catalysis of some unique chemistries observed in natural products.

**Methyl groups at the α- and β-carbons**

As described above for polyketide biosynthesis, the presence or absence of a methyl group on a carbon results from the growing polyketide chain is most often governed by the selection of the extender unit (malonyl-CoA versus methylmalonyl-CoA). However, in PKS systems that use trans acyltransferases (AT-less type I PKSs) (8, 42), the module by module control over extender unit selection is not possible. In most cases, malonyl-CoA is used as the extender unit, and a methyl group can be added to selected positions through the action of an embedded methyl transferase domain (MT) or downstream TE enzyme. For example, the C-6 methyl of leinamycin 8 is thought to be installed by the MT embedded in LnmJ (43), and the C-10 methyl of curacin A 3 is generated via the MT domain in CurJ (35), and the gem dimethyl groups on C-8 and C-18 of broycin 2 most likely are the consequence of the MT domains in BryD and BryC (44).

In contrast to the α-carbon methylations, the incorporation of methyl or methylene groups (or functional groups derived from such groups) at the β-position represents the assimilation of a full cassette of enzymes into the typical PKS machinery. Recently, a subset of type I modular PKSs (and hybrid NRPS/PKS megasynthases) have been identified that contain multiple enzymes acting in trans during the traditional linear assembly-line process to accomplish β-branching. Termed HMGS-CoA synthase (HMGS) cassettes, these enzyme systems provide a unique method of expanding the repertoire of the traditional reductive domains (KR, DH, ER). These enzymes work in conjunction with the PKS machinery to create unique functionalities observed at the branch points that include the pendant methyl groups of bacillaleus A 1 (36, 45, 46), mupirocin 9 (47), and viriginiamycin M 15 (48), which are the methoxymethyl and ethyl groups of myxavirucin A 30 (49, 50); the exo-methylene groups of difficidin B 4 (45), onamidic acid A 11 (51), and pedelin D 12 (51–53); the cyclopentyl ring of curacin A 3 (35, 54); the vinyl chloride of jasmineamide B 7 (55); the unique...
Polyketide Biosynthesis, Modular Polyketide Synthases

(a) HMG-CoA synthase (HMGS) reaction from primary metabolism. (b) An HMGS cassette can convert the β-ketone to an alkene (β,γ or γ,δ double bond) with a pendant methyl (or ethyl) group.

Figure 5. HMGS cassette reaction scheme. (a) HMG-CoA synthase (HMGS) reaction from primary metabolism. (b) An HMGS cassette can convert the β-ketone to an alkene (β,γ or γ,δ double bond) with a pendant methyl (or ethyl) group.

1,3-dioxo-1,2-dithiolane moiety of leinamycin B (43); and the exocyclic olefins in bryostatin 2 (Fig. 4) (44).

HMGS synthase cassettes

In primary metabolism, HMG-CoA synthase (HMGS) is responsible for the condensation of C-2 of acetyl-CoA onto the β-ketone of acetoacetyl-CoA to form 3-hydroxyl-3-methylglutaryl-CoA and free CoASH (Fig. 5a) (56). Several secondary metabolite pathways have been identified over the past five years that perform an analogous reaction, although they seem to use ACP-tethered acyl groups (Fig. 5b) as opposed to acyl-CoA substrates. After generation of the HMG-ACP analog on the growing polyketide chain, the product is usually dehydrated and decarboxylated to yield the branched intermediate. Found in 11 pathways to date (36, 45–55), included in the cassette are a discrete ACP, a decarboxylative KS (active site cysteine is replaced with a serine), an HMGS, and one or two enoyl CoA hydratase-like (ECH) domains. (Table 1, Figs. 5, 6).

HMGS cassette biochemistry

Three HMGS-containing cassettes (those in the curacin A, bacil laene, and myxovirescin pathways) have been validated biochemically in the past two years and will serve as the basis for our analysis of the individual components in this complex (46, 54, 57, 58). The mechanistic and structural details for HMG-CoA synthase in primary metabolism have been elucidated for both bacterial and eukaryotic HMGSs (59–63). Although polyketide HMGSs share only 20–30% sequence identity with their primary metabolism homologs (in both prokaryotes and eukaryotes), multiple sequence alignment reveals that the key catalytic residues (Glu/Cys/His) are conserved. As shown in Fig. 5b, the first step in the formation of the HMG-intermediate is the generation of acetyl-ACP. This step is accomplished through the loading via an AT of malonyl-CoA [or perhaps methylmalonyl-CoA in the case of TaE (58)]. Then, the decarboxylative KS converts the malonyl-ACP into acetyl-ACP, after which the tethered acyl group is condensed onto the β-ketone of the polyketide intermediate. Finally, formation of the HMG-analog is completed on addition of water.

Processing of the HMG-intermediate can vary considerably, but typically proceeds via dehydration and decarboxylation catalyzed by two enoyl-CoA hydratase-like domains (Fig. 5b). Based on sequence similarity, the members of the crotonase fold family observed in these HMGS cassettes can be subdivided into two groups, termed ECH1 and ECH2 (54). The successive dehydration and decarboxylation steps are catalyzed by the ECH1 and ECH2 enzymes/domains, respectively. Evidence for the specific function of the curacin ECH1 and ECH2 enzyme pair from the curacin pathway has been demonstrated using a coupled enzyme assay and ESI-FT-ICR MS (54). Using purified ECH1 (CurE) and ECH2 (the N-terminal domain of CurF) overexpressed in E. coli, (52) HMG-ACP was converted first to...
the above proteins, ACP, acetoacetyl-ACP, and malonyl-CoA in combination with PksL, PksF, PksG, PksH, and PksI. Using the model acceptor they assigned functional roles to AcpK, PksC, the tandem ACPs. Using radioactive biochemical assays together with mass spectrometry, reported the function of several discrete enzymes. Using radiative biochemical assays together with mass spectrometry, they assigned functional roles to AcpK, PksC, the tandem ACPs in PksL, PksF, PksG, PksH, and PksI. Using the model acceptor ACP, acetoacetyl-ACP, and malonyl-CoA in combination with the above proteins, a L2,3-isopropenyl-S-carrier protein was generated (46). Most recently, a similar in vivo investigation was conducted using the homologous enzymes from the myxovirescin pathway (58). The HMGS cassette reaction sequence proposed above held fast for the myxovirescin pathway, although the generation of the propionyl- or methylmalonyl-S-ACP could not be demonstrated. The authors have suggested that perhaps additional enzymes are yet to be identified to fill these roles to complete the $\beta$-ethylation at C16 in 3D. Two variations on the HMGS cassette theme already have been identified. In the biosynthesis of bryostatin 1 and leinamycin 8, one or both of the ECH-mediated steps is likely omitted based on the final natural product structures. The details of these deviations have not yet been established.

**Table 1** HMGS containing biosynthetic pathways and their producing organisms

<table>
<thead>
<tr>
<th>Natural product</th>
<th>Producing organism</th>
<th>Discrete ACP</th>
<th>KS (Cys→Ser)</th>
<th>HMGS</th>
<th>ECH1 (Dehydration)</th>
<th>ECH2 (Decarboxylation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacillaene</td>
<td>Bacillus subtilis</td>
<td>AcpK/BeF</td>
<td>PksF</td>
<td>PksJ/BaeG</td>
<td>PksK/BaeH</td>
<td>PksI/BaeH</td>
</tr>
<tr>
<td>Bryostatin</td>
<td>Candidatus</td>
<td>BryQ</td>
<td>PksF</td>
<td>PksJ/BaeG</td>
<td>PksK/BaeH</td>
<td>PksI/BaeH</td>
</tr>
<tr>
<td>Curacin</td>
<td>Lyngbya majuscule</td>
<td>CurB</td>
<td>CurC</td>
<td>CurD</td>
<td>CurE</td>
<td>CurF N-terminal domain</td>
</tr>
<tr>
<td>Diflucidin</td>
<td>Bacillus</td>
<td>DifC</td>
<td>DifN</td>
<td>DifO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jamaicamide</td>
<td>Lyngbya majuscule</td>
<td>JamF</td>
<td>JamG</td>
<td>JamH</td>
<td>JamI</td>
<td>JamJ N-terminal domain</td>
</tr>
<tr>
<td>Leinamycin</td>
<td>Streptomyces</td>
<td>LnmL</td>
<td>LnmM</td>
<td>LnmF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mupirocin</td>
<td>Pseudomonas</td>
<td>MupG</td>
<td>MupH</td>
<td>MupJ</td>
<td>MupK</td>
<td></td>
</tr>
<tr>
<td>Myxovirescin A (antibiotic TA)</td>
<td>Myxococcus sanchus</td>
<td>Tab &amp; TaE</td>
<td>TaK</td>
<td>TaF</td>
<td>TaF</td>
<td>TaF</td>
</tr>
<tr>
<td>Onnamide</td>
<td>Symbiont bacterium</td>
<td>OnnA</td>
<td>OnnB</td>
<td>OnnB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pederin</td>
<td>Symbiont bacterium</td>
<td>PedN</td>
<td>PedM</td>
<td>PedP</td>
<td>PedL</td>
<td>PedL (embedded)</td>
</tr>
<tr>
<td>Virginiamycin M</td>
<td>Streptomyces</td>
<td>VirB</td>
<td>VirC</td>
<td>VirD</td>
<td>VirE</td>
<td></td>
</tr>
</tbody>
</table>

3-methylglutaryl-ACP then to 3-methylcrotonyl-ACP, which is the gained intermediate for subsequent formation of the cyclopropyl ring. Insights into the mechanism of the CurF ECH2 decarboxylation have been gained based on the recent crystal structure of the curacin ECH2 domain (64). Additional in vitro evidence for the function of these enzymes has been generated using proteins from the PksX pathway of Bacillus subtilis (46) and Myxococcus xanthus (58). Prior to the identification of bacillaene as the product of the PksX pathway, Cáderone et al. (46) and Dorrestein et al. (57) reported the function of several discrete enzymes. Using radioactive biochemical assays together with mass spectrometry, they assigned functional roles to AcpK, PksC, the tandem ACPs in PksL, PksF, PksG, PksH, and PksI. Using the model acceptor ACP, acetoacetyl-ACP, and malonyl-CoA in combination with the above proteins, a L2,3-isopropenyl-S-carrier protein was generated (46). Most recently, a similar in vivo investigation was conducted using the homologous enzymes from the myxovirescin pathway (58). The HMGS cassette reaction sequence proposed above held fast for the myxovirescin pathway, although the generation of the propionyl- or methylmalonyl-S-ACP could not be demonstrated. The authors have suggested that perhaps additional enzymes are yet to be identified to fill these roles to complete the $\beta$-ethylation at C16 in 3D. Two variations on the HMGS cassette theme already have been identified. In the biosynthesis of bryostatin 1 and leinamycin 8, one or both of the ECH-mediated steps is likely omitted based on the final natural product structures. The details of these deviations have not yet been established.

Recently, in vivo evidence for the function of these HMGS cassettes has come from the Müller lab (49, 50, 65). To date, all HMGS cassette proteins for myxovirescin A (Tab/TaC, TaE/TaF, TaK, TaX, TaY) has been individually deleted, and the impact on the products of the engineered *Myxococcus xanthus* strains has been analyzed. In addition, analysis of products from TaV (the trans-acting AT), TaG (a cytochrome P450) that is thought to hydroxylate the HMGS-installed $\beta$-methyl group at C12 of Myxovirescin A was abolished (or greatly reduced) in all above deletion strains. A appearance of novel myxovirescin analogs ($\beta$-methyl vs $\beta$-methyl transferase necessary for completing the transformation to the final methoxy-methyl functionality) strains have provided insights into this complex pathway. Production of myxovirescin A was abolished (or greatly reduced) in all above deletion strains. A appearance of novel myxovirescin analogs ($\beta$-methyl vs $\beta$-ethyl) at C16 in the TaE & TaF strains appears to be a result of TaB or TaC supplementation, which provides direct evidence for TaE/TaF in the formation of the ethyl branch point. However, independent biochemical verification of TaF function has been difficult to obtain (58).

**HMGS cassette architecture**

Analysis of the placement of the known HMGS cassettes identified to date into their biosynthetic clusters reveals a variety...
Figure 6: HMGS containing biosynthetic pathways. Portions of the PKS and PKS/NRPS pathways where the HMGS and related enzymes are located.

of possible architectures (Fig. 6). For example, the ECH₂ decarboxylase exists as a discrete enzyme downstream of the ECH₁ dehydratase (mupirocin and others), as an N-terminal domain of a large PKS (curacin and jamaicamide), and as an embedded domain (pederin and onnamide). Although most clusters published to date are mixed PKS/NRPS systems with tandem ATs and tandem ACPs at the site of HMGS modification, exceptions exist for each example (difficidin is PKS only, curacin and jamaicamide contain embedded ATs, and bryostatin and myxovirescin do not contain tandem ACPs at the site of HMGS modification).

As HMGS enzyme cassettes have been identified and functionally characterized only recently, many mechanistic details as well as the key protein-protein interactions needed to orchestrate communication among the polyolide components remain unclear. Details on how the individual proteins are brought to the correct place in the pathway to perform their functions are still unknown for most pathways. In the case of the PksX/bacillaene pathway, some intriguing microscopy performed on B. subtilis suggests that the bacillaene proteins are clustered into a huge mega-enzyme factory inside the bacterial cell (66). Whether this organization extends (or is limited) to the other members of HMGS cassette containing pathways remains to be observed. Additionally, in some pathways, key enzymes have yet to be identified. Two lingering questions include: 1) Which AT domain loads the discrete ACP in the embedded AT systems specified by the curacin and jamaicamide pathways? and 2) where are the missing domains located in the incomplete cassettes? Despite these remaining issues, the stage is now set for these unique suites of enzymes to be included and applied in the growing metabolic engineering/combinatorial biosynthesis toolbox.

The goal of this review has been to highlight a series of novel systems for creating chemical diversity in polyketide natural product biosynthesis. This review includes the mechanistic basis for introduction of diverse branching functionality involves the HMGS-containing enzymes that are being identified in a growing number of PKS and mixed NRPS-PKS pathways. The rapidly increasing knowledge and mechanistic understanding of these complex metabolic systems will provide growing opportunities to engineer chemical diversity using rational approaches.

Acknowledgments

Research on modular type I polyketide synthases in the Sherman laboratory is supported generously by grants from the National Institutes of Health (GM 076477, CA 108874, TW 007404), the Hans and Ella McCollum Vahlteich Research Fund at the University of Michigan College of Pharmacy, and the H.W. Vahlteich Professorship in Medicinal Chemistry, J.D.K., is supported by an NRSA postdoctoral fellowship (GM 075641) from the NIH.

References


51. Simunovic V, Muller R. 3-Hydroxy-3-methylglutaryl-CoA-like synthases direct the formation of methyl and ethyl side groups in the biosynthesis of the antibiotic mycovirulenc A. ChemBioChem. 2007;8:497–500.


See Also

Nature: A Model System for Chemists
Antibiotics, Biosynthesis of
Natural Products in Microbes, Chemical Diversity of
Pharmaceuticals, Natural Products and Natural Product Models for
This review covers the biosynthesis of terrestrial and marine polyethers and discusses their biologic properties and the molecular genetics and enzymology of the proteins responsible for their formation. The biosynthesis of monensin, nanchangmycin, nonactin, and the marine polyether ladders are discussed in detail. Novel enzymes found only in type I polyketide polyether gene clusters that are responsible for the epoxidation and cyclization of polyene biosynthetic intermediates are described. The macrotetrolide biosynthetic gene cluster, which is an ACP-less type II polyketide synthase that functions noniteratively is reviewed.

**Biological Background**

Interest in these compounds (Fig. 1) derives from their ability to transport ions across biologic membranes, and some terrestrial polyethers have been used widely in veterinary medicine. Marine polyethers are responsible for numerous cases of human food poisoning and toxic algal tides, which cause massive fish kills.

**Terrestrial polyethers**

At physiological pH, the ionophores are ionized with the fat-soluble part of the molecule residing in the lipid bilayer of the membrane and the ionized moiety in the aqueous milieu (4). Binding of the metal ion takes place at the membrane surface. As successive ether oxygen atoms from the ionophore bind to the metal, it loses its solvated water molecules, thereby forming a neutral zwitterionic metal-ionophore complex. Transport across the membrane can now take place, and at the opposite surface, the process is reversed to leave the metal cation and the anionic ionophore on the other side of the membrane. The ionophore must bind another cation (usually a proton) to return to its original starting point.

The anionic ionophores are highly selective for particular metal cations, and both kinetic and thermodynamic terms determine which cation is selected. Note that the thermodynamic stability of the metal-ionophore complex does not always determine the transport rate. For example, nigericin forms a much more stable complex with potassium but transports sodium much more quickly. In the presence of both sodium and potassium, however, the extra stability of the nigericin–potassium complex results in the preferential transport of potassium over sodium. Selectivities of ionophores are shown in Table 1 (5, 6). The carboxyl group may (e.g., nigericin) or may not (e.g., monensin) be involved in cation liganding via an ionic bond. Monensin has six liganding oxygen atoms but none for ionic bonds, and it is a weaker complexing agent than nigericin. Lasalocid can form a singly charged complex with doubly charged cations such as Ca$^{2+}$. The broad specificity of lasalocid results from the metal's sitting on top rather than within the oxygen system of the ionophore. For nonactin, the ester carbonyl oxygen atoms and the ether oxygen atoms are all liganding and form the apices of the ionophore.

<table>
<thead>
<tr>
<th>Ionophore</th>
<th>MW</th>
<th>Selectivity sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigericin 1</td>
<td>724</td>
<td>K &gt; Rb &gt; Na &gt; Cs &gt; Li</td>
</tr>
<tr>
<td>Monensin A</td>
<td>670</td>
<td>Na ≈ K &gt; Rb &gt; Li &gt; Cs</td>
</tr>
<tr>
<td>Lasalocid 4</td>
<td>590</td>
<td>Cs &gt; Rb ≈ K &gt; Na &gt; Li; Ba &gt; Sr &gt; Ca &gt; Mg</td>
</tr>
<tr>
<td>Nonactin 5</td>
<td>736</td>
<td>NH$_4$ &gt; K &gt; Rb &gt; Cs &gt; Na</td>
</tr>
<tr>
<td>Tetractin 10</td>
<td>602</td>
<td>Ca &gt; Mg &gt; Na</td>
</tr>
</tbody>
</table>
Polyketide Biosynthesis, Polyethers

Commercial interest in the polyketide polyethers stems from their antibiotic activity. In many polyethers, the class is potent coccidiostats and can control coccidial infections in poultry when added at approximately 100 ppm to the feed (7). Coccidiosis is caused from an infection in the digestive tract by parasitic protozoa, most commonly in poultry from the genus _Eimeria_. Monensin was the first polyether antibiotic to be patented and was released onto the market in 1971. It is still widely used today and is marketed under the tradename of Coban® (Elanco, Toronto, Ontario, Canada). After monensin treatment, coccidia have been observed to literally explode at daily intake levels should be set in line with Australian limits of human consumption. Lasalocid residues have been detected in food for which 1.5% of these are fatal (11). These toxins are produced by marine microalgae, predominantly dinoflagellates, and are associated with red tides and lead to massive fish kills and molluscan contamination. Most dinoflagellate toxins are polyketides and include many polyethers with distinctive ring junctions resembling the rungs of a ladder. The dinoflagellate toxins are hazardous if ingested or ingested. They are odorless, tasteless, and not destroyed by cooking or autoclaving. These toxins are extremely toxic in minute quantities, and no acceptable exposure limits have been established. Okadaic acid 11 is a potent and specific inhibitor of protein phosphatases produced by the dinoflagellates _Prorocentrum lima_, _Dinophysis fortii_, and _Dinophysis acuminate_. It accumulates in bivalves and is one of the main toxins responsible for diarrhetic shellfish poisoning (DSP) (11). It has a highly unusual biosynthesis that has generated a lot of speculation because of the presence of "isolated" acetate chain methyl carbon atoms. Okadaic acid is a potent inhibitor of protein phosphatase A (11).

The brevetoxins (e.g., brevetoxin A 13 and brevetoxin B 12) are a family of lipid soluble neurotoxins produced by _Karenia brevis_ (formally known as _Gymnodinium breve_), which are struc-turally similar to yessotoxin. They are thought to exert their biological action through the depolarization of the sodium channels of the excitabile membranes (11). Binding of brevetoxin to the voltage-gated sodium channels changes its function by shifting the activation voltage for channel opening to a more negative value and inhibiting the inactivation of opened channels (resulting in persistent activation). If ingested, they cause neurotoxic shellfish poisoning. Brevetoxin derivatives have been patented as a treatment for cystic fibrosis, mucociliary dysfunction, and pulmonary diseases. Yessotoxins 14 are produced by dinoflagellates of the genera _Protoceratium_ and _Gonyaulax_. They cause selective disruption of the E-cadherin-catenin system in epithelial cells (11). In common with the other "ladder" shaped marine polyethers, yessotoxin interacts with transmembrane helix domains of the gambiaic acids 15, which are isolated from _Gambierdiscus toxicus_, are potent antifungal compounds that are particularly effective against filamentous fungi but inactive against yeasts. They are also cytotoxic, but they do not exhibit the neurotoxicity that is associated with other large marine polyethers such as yessotoxins, brevetoxins, and maitotoxins (11).

With a molecular weight of 3422 Da, maitotoxin 16 (see Fig. 4a later on in this article) is the largest nonproteinaceous natural product isolated (3). It is a marine polyether ladder produced by the dinoflagellate _G. toxicus_. Maitotoxin is a powerful activator of voltage-insensitive Ca²⁺ channels and exerts its biological effect through the increase of intracellular Ca²⁺ concentration (11). It can be used as a tool for studies on cellular events associated with Ca²⁺ flux. It stimulates synthesis and secretion of the nerve growth factor. The minimum lethal dose of the toxin in mice is 0.17 µg/kg (intraperitoneally).

The ciguatoxins (CTXs) are responsible for the symptoms of ciguatera fish poisoning caused by ingesting certain tropical and semitropical fish from the Indo-Pacific Oceans and Caribbean Sea. Ciguatoxin poisoning is the most frequent food-borne illness related to fish consumption although it is rarely fatal.
Symptoms include neurological, cardiovascular, and gastrointestinal disorders, and ciguatoxins can be detected in all body fluids (12). Most people recover slowly over time, which can take from weeks to years. P-CTX1 (17) is the most toxic of all ciguatoxins with an LD50 of 0.25 µg/kg intraperitoneally. The toxins derive from the Gambierdiscus spp., which produce less-potent compounds that are biotransformed in the livers of the fin-fish to the more toxic ciguatoxins. The compounds activate the voltage-gated sodium channels that result in an increase of intracellular sodium. In high doses, ciguatoxins block the voltage-gated potassium channels that lead to membrane depolarization and that contribute to a lowering of action potential threshold (12).
Chemistry

Classical feeding experiments with both stable and radioactive isotopic labels (7) enabled the biosynthetic origin of the polyethers to be elucidated and for a general stereochemical model to be proposed (1). More recent work on this class of compounds has focused on a genetic approach, and unusual and interesting genes specific to polyether biosynthesis have been isolated from these clusters.

Monensin

Monensins A and B are polyether ionophores produced by Streptomyces cinnamonensis that differ only in the sidechain at C16 (ethyl/methyl). Monensin acts as a specific ionophore to dissipate ionic gradients across cell membranes and is used widely in veterinary medicine and as a food additive in animal husbandry (7). Antimalarial activity has also been reported (8, 9). Monensin is the best studied of the polyether ionophore antibiotics, and it was the first to have its gene cluster sequenced (2).

Early feeding studies established that monensin A is biosynthesized from a classical polyketide pathway and is derived from five acetate, seven propionate, and one butyrate (for monensin B, an additional propionate unit replaces the butyrate) (7). Four of the nine oxygen atoms are derived from molecular oxygen, with the remaining five deriving from the corresponding carboxylic acid precursors (Fig. 2a). Based on these initial

![Chemical structures](image)

Figure 2  (a) Classic biosynthesis of monensin A (7); (b) novel monensins (16); (c) monensinyl N-acetylcysteamine thioester (14); (d) triene lactones produced by monC1 mutant (15); (e) biosynthesis of monensin: The timing of the methylation and hydroxylation steps and release from the PKS are not known for certain (2, 13–16).
# Polyketide Biosynthesis, Polyethers

## Table 2

<table>
<thead>
<tr>
<th>Polypeptide (size aa)</th>
<th>Proposed function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MonA</strong> (3025)</td>
<td></td>
</tr>
<tr>
<td>Loading module Module 1</td>
<td>KSQ, AT(A), ACP</td>
</tr>
<tr>
<td></td>
<td>KS, AT(P), DH, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAII</strong> (2226)</td>
<td></td>
</tr>
<tr>
<td>Module 2</td>
<td>KS, AT(P), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAIII</strong> (4332)</td>
<td></td>
</tr>
<tr>
<td>Module 3</td>
<td>KS, AT(A), DH, KR, ACP</td>
</tr>
<tr>
<td></td>
<td>KS, AT(P), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAIV</strong> (4336)</td>
<td></td>
</tr>
<tr>
<td>Module 5</td>
<td>KS, AT(B), DH, KR, ACP</td>
</tr>
<tr>
<td></td>
<td>KS, AT(A), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAV</strong> (4330)</td>
<td></td>
</tr>
<tr>
<td>Module 6</td>
<td>KS, AT(P), DH, KR, ACP</td>
</tr>
<tr>
<td></td>
<td>KS, AT(A), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAVI</strong> (1700)</td>
<td></td>
</tr>
<tr>
<td>Module 7</td>
<td>KS, AT(A), KR, ACP</td>
</tr>
<tr>
<td><strong>MonAVII</strong> (1641)</td>
<td></td>
</tr>
<tr>
<td>Module 10</td>
<td>KS, AT(P), KR, ACP</td>
</tr>
<tr>
<td><strong>MonAIX</strong> (3753)</td>
<td></td>
</tr>
<tr>
<td>Module 11</td>
<td>KS, AT(P), KR, ACP</td>
</tr>
<tr>
<td></td>
<td>KS, AT(P), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>MonAX</strong> (268)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PKS, Type II TE</td>
</tr>
<tr>
<td><strong>MonB</strong> (267)</td>
<td></td>
</tr>
<tr>
<td>Module 12</td>
<td>Epoxide hydrolase/cyclase(^*)</td>
</tr>
<tr>
<td><strong>MonC</strong> (496)</td>
<td></td>
</tr>
<tr>
<td>Module 14</td>
<td>Non-haem epoxidase</td>
</tr>
<tr>
<td><strong>MonD</strong> (299)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chain terminating TE(^*)</td>
</tr>
<tr>
<td><strong>MonE</strong> (458)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hydroxylase</td>
</tr>
<tr>
<td><strong>MonF</strong> (276)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methyltransferase</td>
</tr>
<tr>
<td><strong>MonG</strong>, <strong>MonRI</strong>, <strong>MonRII</strong></td>
<td>Transcriptional regulators</td>
</tr>
<tr>
<td><strong>MonH</strong>, <strong>MonRII</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Efflux protein, resistance protein</td>
</tr>
</tbody>
</table>

Reproduced from Reference 17 with permission from the Royal Society of Chemistry.

\(^*\)"unnecessary" domain (sequence is indistinguishable from active Mon DH domains).

\(^{†}\)The AT of module 5 should incorporate an ethylmalonate extender unit, and it contains a signature sequence very close to the propionate sequence.

\(^{‡}\)Predicted to be inactive.

\(^{||}\)Originally assigned as double bond isomerase.

\(^{§}\)Originally assigned as epoxide cyclase.

Experiments, it was proposed that the monensin PKS produced a linear E,E,E-triene precursor \(^{24}\) that was oxidized and cyclized to give the final structure \(^{1}\). Alternative proposals using Z,Z,Z- and E,Z,Z-trienes have been made \(^{2}\).

Publication of the gene cluster for monensin \(^{2, 13-17}\) showed that the PKS comprised 12 modules in eight contiguous open reading frames consistent with the production of the linear triene premonensin \(^{25}\). The loading module contains an N-terminal KSQ domain that functions as a malonylCoA decarboxylase to generate starter units in situ. The monensin PKS does not contain an integrated C-terminal thioesterase domain at the end of module 12; instead an unusual 121 amino acid extension, rich in glycine, asparagines, and glutamine, was found \(^{13}\). **MonCII\(^{||}\)**, which was originally assigned as an epoxide cyclase, has been shown to hydrolyze monensinyl NAC thioester \(^{21}\) (Fig. 2c) as well as two other model substrates \(^{14}\). Deletion of **monCII** gave a mutant that produced none or only trace amounts of monensin, which is consistent with its role as a chain-terminating TE. Complementation of **monCII** on a plasmid restored monensin production. Moreover, cell-free extracts from the **monCII** mutant treated with KOH gave significant amounts of monensin that resulted from the hydrolysis of a monensinyl ester or thioester present in the cell-free extract. Two ORFs, **monAIX** and **monAX**, have been shown to function as Type II thioesterases. Deletion of **monAIX** and/or **monAX** resulted in a modest drop in the monensin titer, which is consistent with their editing Type II TE role \(^{14}\). Genes for late steps such as methylation (**monE**) and hydroxylation (**monD**) and several regulatory genes (**monR, R1, R1I**) were found in the cluster \(^{2, 13}\). **MonT** is proposed to be involved in monensin export and consequently ensures the
producing strain is self-resistant. The monC gene product is a flavin-dependent epoxidase. Deletion of monC results in complete loss of monensin production and the accumulation of the linear E, E, E-triene lactones \(\text{22, 23, Fig. 2d,}\) which differ from premonensins \(\text{25}\) only by the different cyclization pattern of the polyketide chain: \(\text{3,3-lactone} \text{22, 23, compared with a hemi-ketal} \text{25} (12).\) Strains in which the adjacent genes monB1 and/or monBII are deleted in addition to monC show the same results. No oxidized derivatives of \(\text{22 and 23}\) were found, which suggests that the epoxidase encoded by monC is necessary and sufficient for epoxidation of all three double bonds. These experiments are consistent with an E,E,E-configured triene as originally proposed by Cané et al. (1). Heterologous expression of \(\text{monC}\) in Streptomyces coelicolor gave a strain that could convert linaroil to linalool oxide with a 10-20-fold greater conversion activity than studies carried out previously with \(S.\) cinnamonensis (13).

MonB1 and monBII are highly homologous to each other and significantly similar to a \(\alpha,\alpha,\alpha\)-ketoesteroid isomerase of \(S.\) monomos testosteroni. Originally, it was proposed that monB1 and monBII interconverted \(\text{E, E, E-}\)double bonds to \(\text{Z}\) via an extended enolate ion during polyketide biosynthesis in two modules (resulting in an \(\text{E, Z, Z-}\)triene intermediate) (2). Their assignment has been revised in light of new experimental evidence that has shown that they are involved in the epoxide ring opening and concomitant polymer ring formation (16). Deletion of monB1 and/or monBII gave strains that produced no monensin. Instead, \(\text{C-3,0-demethylmonensins 18, C-9-epi-monomensins 19, and C-26-deoxy-epi-monomensins 20 (Fig. 2a),}\) were produced in addition to numerous minor components. The major products from these mutant strains were the \(\text{C-3,0-dimethyl}\) analogs. Hence, in the absence of the \(\text{C-3-methoxy group, nucleophilic attack by the C-5 hydroxy group from the si face of the C-9 carbonyl giving the natural epimer of the spiroketal takes place (Fig. 2a). In contrast, the presence of the methoxy group at C-3 leads to the C-5 hydroxy attacking the C-9 carbonyl from the re face leading to the unnatural spiroketal epimer 39. The C-26-deoxy-epi-monomensins result from them being poor substrates for MonA. Treatment with acid converted the epi-monomensins into the more thermodynamically stable corresponding monomers. Additional minor metabolites were detected in the cell culture but were produced in insufficient quantities to allow for structure determination. LC-MS analysis revealed that their molecular weights were identical to monensins A/B and C-3,0-demethylmonensins A/B, which suggests that they were intermediates with the correct oxidation state but cyclized incompletely. Treatment of these minor metabolites with acid led to their conversion to monensins A/B and C-3,0-demethylmonensins A/B. Taken together, these results show that the MonB enzymes are epoxide hydrolyase/cyclase enzymes that accelerate, but do not change, the stereochemical course of polyether ring formation (16). Since the pattern of products in MonB1 and MonBII mutants seem to be identical, it seems that their actions are somehow coordinated and could in principle exist as a heterodimer in \(S.\) cinnamonensis. Although the arrangement of the proteins is not known, it is clear that ring closure in monensin biosynthesis occurs step-wise in a precise order under enzymatic control from one or both MonB enzymes. The timing of the hydroxylation step at C-26 has also been shown to occur before either methylation or polyether ring formation (16). The biosynthetic pathway for monensin is shown in Fig. 2e.

**Nanchangmycin**

Streptomyces nanchangensis produces the polyether antibiotic nanchangmycin 29, the 16-membered macrolide meilingmycin, and at least two other antibiotics of unknown structure (18).

The nanchangmycin gene cluster comprises 30 ORFs in a region of \(\text{132 kb}\) of DNA (Table 3 (19, 20). The type I PKS, encoding 14 modules in 11 ORFs, is arranged in two distinct groups (\text{nanA1-A6, nanA7-A11}) with the sugar biosynthesis genes located in between. A KSQ domain is located in the loading domain in \(\text{NanA1}\) and the decarboxylation of malonate provides the acetate start unit. The type II ACP (\text{NanA10}), which lacks an ACP function. Either side of \text{nanA10} three unique polymer genes are located: \text{nanI}, \text{nanO}, and \text{nanE}, which are homologous to \text{monB1/bi, monC1, and monCII.}\) The epoxidase encoded by \text{nanO} oxidizes the triene precursor \(\text{30, which is then cyclized by NanE (Fig. 2b). Originally,}\) the chain-releasing (CR) domain located in module 14 was thought to carry out the release of the polyether chain from the nanchangmycin PKS (19). Recent work has shown that although the CR domain is required in vivo for the efficient production of nanchangmycin, it does not catalyze the release of the polyether chain from the PKS (20). \text{NanE,}\) in fact responsible for the release of the fully processed nanchangmycin. It is homologous to both \text{monC11} and \text{NigCII (which carries out an analogous role in nigericin biosynthesis). A gene encoding a cytochrome P450 (nanF) is located adjacent to the second PKS cluster and is thought to catalyze the oxidation of the C30 methyl group (19). It is proposed that the biosynthesis of nanchangmycin (Fig. 3b (19, 20)) begins with loading of malonyl CoA onto \text{NanAI.}\) Subsequent decarboxylation and 13 rounds of extension would result in a tetradecaketide intermediate 30 that is attached to the type II independent ACP \text{NanA10}. Two epoxidations of the triene intermediate by \text{NanO} and ring opening of the di-epoxide 31 by \text{NanI} take place while the PKS intermediate is still attached to the independent ACP \text{NanA10}. The polymer chain 32 is then transferred to the ACP of module 14 for the final round of chain extension. Release from the PKS (\text{NanE}) followed by hydrolase (at C-30 by the putative monooxygenase \text{NanP}) would give the completed aglycone polyether 33. Biosynthesis of the 4-0-methyl-L-rhodinose from D-glucose-1-phosphate (\text{NarG/1-G4, NanM}) and attachment to the aglycone by \text{NanG/5 provides the completed nanchangmycin.}\)

The novel nanchangmycin aglycone 29 (Fig. 3a containing a keto group at C-19 was produced from a mutant containing an in-frame deletion of the module \(\text{6 K R (19).}\)

**Nonactin**

The macrotetrolides 5-9 are a family of macrocyclic polyethers produced by \(S.\) griseus that exhibit a broad spectrum of biological activities (antibacterial, antitumor, antifungal, and immunosuppressive) (5, 21). They can act as ionophores and...
are effective against Gram-positive bacteria, mycobacteria, and fungi. The parent compound, nonactin 5, is an achiral molecule assembled from four molecules of enantiomeric non-actic acid \[ \text{\textit{34}} \] (Fig. 3c) in a (+)-(−)-(−)−ester linkage. The homologues \[ \text{\textit{6–9}} \] are derived from the substitution of ethyl groups for methyl groups on the macrocyclic backbone.

Classical biosynthetic studies with \(^{14}\text{C}\)-labeled compounds and later with stable isotopes established that nonactic acid is derived from two acetates (or malonates), succinate and propionate \([\text{\textit{Fig. 3d}}]\) (21). Feeding studies with \[^{36}\text{C}\] has shown that the first committed step of macrotetrolide biosynthesis is the coupling of the succinate unit with an acetate (or malonate) to give \(-\alpha\)-ketoadipate \[ \text{\textit{40}} \] (22). The late steps of non-actic acid biosynthesis \([\text{\textit{Fig. 3g}}]\) were shown to involve the cyclization of \textit{42a}\ into (−)-nonactate and \textit{42b}\ into (+)-nonactate, which demonstrates that a pair of enantiocomplementary pathways are operating in nonactate biosynthesis \(\text{\textit{32}}\). (+)-Non-actic acid is incorporated efficiently into nonactin. Both (+) and (−)-nonactic acids as well as their dimers have been isolated from cultures of \textit{S. griseus}.

Genetic characterization of the nonactin gene cluster has revealed that an ACP-less Type II PKS is used to assemble the macrotetrolides \(\text{\textit{22}}\). The non-PKS is highly unusual, catalyzing both C–C and C–O bond-forming reactions, functioning noniteratively and acting on acyl CoA substrates \(\text{\textit{32}}\). The late steps of nonactate biosynthesis \([\text{\textit{Fig. 3g}}]\) were shown to involve the cyclization of \textit{42a}\ into (−)-nonactate and \textit{42b}\ into (+)-nonactate, which

### Table 3

<table>
<thead>
<tr>
<th>Polypeptide (size aa)</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NanA1</strong> (2902)</td>
<td>Loading Module K5Q, AT(A), ACP Module 1 K5, AT(P), DH, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA2</strong> (2223)</td>
<td>Module 1 K5, AT(P), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA3</strong> (4032)</td>
<td>Module 1 K5, AT(A), DH, KR, ACP Module 4 K5, AT(P), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA4</strong> (3956)</td>
<td>Module 5 K5, AT(P), DH, KR, ACP Module 6 K5, AT(A), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA5</strong> (3079)</td>
<td>Module 7 K5, AT(P), DH, KR, ACP Module 8 K5, AT(A), DH, ER, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA6</strong> (1665)</td>
<td>Module 9 K5, AT(A), KR, ACP</td>
</tr>
<tr>
<td><strong>NanA7</strong> (1646)</td>
<td>Module 10 K5, AT(P), KR, ACP</td>
</tr>
<tr>
<td><strong>NanA8</strong> (3405)</td>
<td>Module 11 K5, AT(P), KR, ACP Module 11 K5, AT(P), DH, KR, ACP</td>
</tr>
<tr>
<td><strong>NanA9</strong> (802)</td>
<td>Module 12 K5, AT(P), DH, KR, ACP</td>
</tr>
<tr>
<td>NanA13 (2387)</td>
<td>Module 13 K5, AT(P)</td>
</tr>
<tr>
<td><strong>NanA10</strong> (104)</td>
<td>Module 14 K5, AT(P), DH, ER, KR, ACP, CR</td>
</tr>
<tr>
<td>NanE (296)</td>
<td>Type I TE</td>
</tr>
<tr>
<td>NanG1-G5</td>
<td>4-0 -methyl-L-rhodinose biosynthesis</td>
</tr>
<tr>
<td>NanI (313)</td>
<td>Epoxide hydrolase/cyclase</td>
</tr>
<tr>
<td>NanM (305)</td>
<td>Methyl transferase</td>
</tr>
<tr>
<td>NanO (478)</td>
<td>Epoxidase</td>
</tr>
<tr>
<td>NanP (423)</td>
<td>Cytochrome P450</td>
</tr>
<tr>
<td>NanR1-R4</td>
<td>Regulation</td>
</tr>
<tr>
<td>NanT1-T5</td>
<td>Transporters, chemoreceptor, regulator</td>
</tr>
</tbody>
</table>

Reproduced from Reference 17 with permission from the Royal Society of Chemistry.

*Predicted to be inactive.
†Originally assigned as an epoxide hydrolase.
‡Originally assigned as a ketosteroid isomerase.

The parent compound, nonactin 5, is an achiral molecule assembled from four molecules of enantiomeric non-actic acid \[ \text{\textit{34}} \] (Fig. 3c) in a (+)-(−)-(−)−ester linkage. The homologues \[ \text{\textit{6–9}} \] are derived from the substitution of ethyl groups for methyl groups on the macrocyclic backbone.
Figure 3: (a) Nanchangmycin and novel nanchangmycin aglycone. (b) Nanchangmycin biosynthesis (19, 20). (Adapted from Reference 17 with permission from the Royal Society of Chemistry.) (c) Structure of building blocks for the macrotetrolides. (d) Biosynthetic origin of carbon skeleton of nonactic acid (21). (e) (2,3-13C2)-3-ketoadipate-N-caprylcysteamine thioester (22). (f) Homochiral nonactate dimer (23). (g) Nonactin biosynthesis (21). (Adapted from Reference 17 with permission from the Royal Society of Chemistry.)
these 23 nongenes in S. lividans resulted in macrotetrolide production, thereby confirming that only these genes are required for macrotetrolide biosynthesis; sequestration of an ACP from elsewhere on the S. griseus genome does not occur.

The proposed biosynthetic pathway (Fig. 3g) requires three condensation and four reduction steps that would require at least three K5 and four KR genes. It has not been established whether the five K5 and four KR proteins function independently, but it has been demonstrated that all nine enzymes are required for macrotetrolide biosynthesis. Shen and Kwon (21) have suggested that the macrotetrolide PKS comprises noniteratively functioning subunits, which is unprecedented in all Type II PKSs known to date. NonJ and NonK have been assigned roles in dimerization (see below), which leaves NonPQU to carry out the three C–C bond-forming reactions. A mutant lacking the NonPQU genes was able to convert non-actinic acid into nonactin, whereas deletion of NonK resulted in no macrotetrolide formation. Hence, NonPQU represents a novel Type II minimal PKS that acts noniteratively, does not have an ACP, and uses acyl CoA substrates directly for polyketide biosynthesis.

NonJ has been shown to catalyze the first dimerization step to give (+)-nonactyl(+)-nonactyl CoA (44), and NonK is responsible for the stereospecific cyclodimerization to afford nonactin (21). All C–C bond forming KSs are characterized by a Cys–His–His (Asn) catalytic triad (including Non-PQU), whereas NonK is characterized by a mutated catalytic triad: Cys–Gly–Tyr–His. Replacement of the conserved cysteine residue in NonJ or NonK with glycine gave mutants that could not catalyze the transformation of (+)-non-actinic acid into nonactin. Taken together, NonK cannot function as a decarbonylase and instead the entire pathway to the C–C bond-forming steps in nonactin biosynthesis using the same active site cysteine that other KSs use for C–C bond formation. In common with NonPQU, NonJ and NonK act noniteratively on CoA substrates.

The product of the minimal PKS, 4L, is the branch point of the pathway that diverges into a pair of enantiomeric-specific pathways, each of which involves two KRs and affords (+)- or (-)-42. The nonactate synthase, NonS, catalyzes the intramolecular Michael addition of (+)-nonactinic acid 43a from (-)-42a. A NonS mutant supplemented with (+)-nonactinic acid could produce nonactin, monactin, and dinactin, but not trinactin and tetranactin, which require one (–)-homomonactate moiety, respectively. This process suggests that NonS can cyclize (-)-42a into (+)-nonactin 43b and (+)-homomonactate 35 only and that another nonactate synthase is required for the cyclization of (+)-42b to give (+)-nonactin 43b and (+)-homomonactate. Based on its high sequence homology to CoA ligases, NonL was identified as a CoA ligase catalyzing the transformation of (+)-nonactinic acid into (+)-nonactyl CoA. The sequence of reactions for the biosynthesis of nonactin and the other macrotetrolides in S. griseus is shown in Fig. 3g.

S. griseus is likely to be protected from nonactin by two mechanisms. The first mechanism seems to involve pumping excess antibiotic from the cell. Two genes, orf5 and orf6, which are clustered with the rest of the nonactin biosynthesis genes, show homology with A/B transporter genes. The second mechanism is the enzyme catalyzed hydrolysis of nonactin and homologues by NonR, which is homologous to serine protease and esterase enzymes and confers tetranactin resistance to S. lividans TK24. Overexpressed NonR has been shown to stereospecifically catalyze the hydrolysis of the macrotetrolide ring into homochiral nonactane-dimer 37 (Fig. 3f) in a two-step process (23). The macrotetrolides are converted initially into their secotetramer species and subsequently hydrolyzed to the dimer; in both cases, it is the bond between the alcohol of the (–)-nonactate and the carboxylate of (–)-nonactate that is cleaved. No trimer or monomers were detected.

**Marine polyether ladders**

All marine polyether ladders thus far characterized (e.g., maitotoxin 26 (Fig. 4a), the brevetoxins 12 and 13 and hemibrevetoxin B, the yessotoxins 34 (and the truncated atradiotoxin), the Pacific and Caribbean ciguatoxins, the gambieric acids 15 and gemberial (the gymnocins and brevena) can be grouped into 14 backbone structures (3).

Retrobiosynthetic analysis of these structures has led to the development of a model for the biosynthesis of these complex structures from the cyclization of a polyepoxide precursor (3). The model accounts for the conserved stereochemistry of the numerous ring junctions in these polyethers, which are syntrans (Fig. 4b). The required configuration of the rings can be derived from stereochimically identical all (R,R)- or (S,S)-trans epoxides, which are derived from the appropriate polyene (which may contain over 20 double bonds). Epoxidation of the polyene precursor takes place from the same face, and consequently, it is possible for a single monooxygenase with broad specificity to produce all trans epoxides. Ring closure requires an endo-selective opening of each epoxide.

Independently, Prasad and Shimizu (24) as well as Lee et al. (25) proposed that brevetoxin A is biosynthesized from the cyclization of a polyepoxide precursor in a series of 5,6 (R,R)-trans epoxide openings (Fig. 4c). Gallimore and Spencer (3) argue that the nine disulfated endo-tet closures required for this mechanism makes it mechanically unlikely and point out that an alternative cascade of 5S2 epoxide openings in the opposite direction from all (S,S)-trans epoxides yields the same structure.

Although it can be envisaged that a trans-polyene is converted by a monooxygenase to a polyepoxide intermediate, which is processed by an epoxide hydrolase (cf. NonR in monoenes biosynthesis) to the polyyclic ether, because of the reactivity of the polyepoxide intermediate (which may contain in excess of 20 reactive epoxide groups), this seems unlikely. An alternative mechanism is for the epoxidation and cyclization steps to be coupled in an iterative process whereby the production of an epoxide is followed by ring closure (Fig. 4d). Gallimore and Spencer (3) suggest that it may be possible for this process to be affected by the monooxygenase only: After epoxidation, the bound enzyme stabilizes the endo transition state relative to the exo as the hydroxyl nucleophile attacks. Once the ring has been closed, the enzyme would dissociate and move on to the next double bond. In this manner, a single enzyme could be responsible for the conversion of a polyene chain to a polyether ladder. In both cases, no restriction on...
Polyketide Biosynthesis, Polyethers

Figure 4  (a) Maitotoxin. (b) Common structure feature of marine polyether ring junctions (3). (c) Brevetoxin A biosynthesis as proposed by 1) Shimizu/Nakanishi (24, 25) and 2) Gallimore and Spencer from all (S,S)-trans epoxides (3). (d) Possible enzymatic routes to a fused polyether using 1) a mono-oxygenase and an epoxide hydrolase or 2) a mono-oxygenase only (3).

Of all the marine polyether ladder metabolites examined, only a single ring junction in the largest natural product, maitotoxin, could not be explained using the model (3). Three of the four ladders in maitotoxin conformed to the stereochemical model; however, ladder C requires an epoxide with the opposite stereochemistry to the other centers, which gives rise to the only
exceptional ring junction ("the K ring junction") in any of the known polyether ladders. To explain this anomaly, Gallimore and Spencer (3) refer back to the original stereochemical assignment by Satake et al., which is described as challenging, and they suggest that the assignment be reexamined.

Chemical Tools and Techniques

Early work on polyethers relied on the isolation and structure determination of the natural products followed by classical feeding experiments using both radioactive and stable isotopes to elucidate their biosynthesis (7). Today, advances in genetic techniques have allowed the gene sequences of the terrestrial polyethers to be studied although the number of published polyether clusters is small compared with macrolides and mixed NRPS/PKS systems. No marine polyether has had its gene cluster characterized, which reflects the difficulties with working with marine organisms and the lack of molecular tools available for these systems.

An interdisciplinary approach is required to study the polyethers involving chemists, biochemists, and molecular biologists. The first step is the isolation and structure determination of the polyether by natural product chemists. For the marine polyether ladders, this can be a huge undertaking in itself. Maitotoxin 16, for example, has four ladders, 32 rings, and a molecular weight of 3422 Da! Sophisticated NMR experiments have allowed the gene sequences of the terrestrial polyether clusters to be determined, which reflects the difficulties with working with marine organisms and the lack of molecular tools available for these systems.

Locating the polyether gene cluster can be done by using probes specific for polyketide genes to screen a cosmid library of total DNA. For both monensin 2 (2, 13) and nanchangmycin 28 (19), probes from the erythromycin PKS were used. Alternatively, the polyether’s resistance gene can be used to locate the gene cluster [e.g., monB (5)]. Sequencing and mapping of the cosmids in the library allows the gene cluster size and composition to be determined. More than one PKS sequence can be revealed, so care must be taken to ensure that the gene cluster of interest is being sequenced. For example, eight clusters were identified from the nanchangmycin producer S. nanchangensis, and the nanchangmycin gene cluster was identified through gene disruption (18). The organization and assignment of function of each ORF is carried out by sequence comparison. Cloning and overexpression of individual enzymes in the gene cluster is an important step in confirming the function and mechanism of a particular ORF. For example, sequence comparisons of both the monensin and the nanchangmycin clusters revealed no TE function required for the release of the polyketide from the PKS. However, cloning, overexpression, and characterization of M. Oilya and N. Esg (20) resulted in the reassignment from oxidative cyclization to Type I TE.

The preparation of mutants is an important tool to investigate the role of individual enzymes, and it is used to ascertain which genes in the cluster are essential for the biosynthesis of the polyether and whether modified/truncated products are formed. The effected gene can then be supplied on a plasmid to see whether activity can be restored. Chemists are required to analyze the fermentation mixtures for minor and truncated metabolites. In this manner, the triene lactones Z2 and Z3 were isolated from a 7L culture of a mutant strain of S. cinnamonensis lacking MonCI, and their structure was determined by NMR (15).

HPLC-MS is an essential technique for the separation and identification of these trace compounds, but unless sufficient quantities can be isolated for structure determination by NMR, chemists are also required to synthesize standards for comparison. Characterization of the gene cluster requires chemists to synthesize intermediates (often with stable isotopes at specific places) to be used as substrates for purified enzymes or to confirm the structure of products formed from modified genes.

References


Further Reading


See Also

Polyketide Biosynthesis, Polyethers
Polyketide Biosynthesis, Aromatic Polyketides
Polyketide Biosynthesis, Enediyne Polyketides
Polyketide Biosynthesis, Fungi
Polyketide Biosynthesis, Modular Polyketides
Polyketides as Drugs
Terpenoids in Plant Signaling, Chemical Ecology

Iris F. Kappers, Laboratory of Entomology & Plant Research International, Wageningen UR, Wageningen, The Netherlands
Marcel Dicke, Laboratory of Entomology, Wageningen UR, Wageningen, The Netherlands
Harro J. Bouwmeester, Laboratory for Plant Physiology and Plant Research International, Wageningen UR, Wageningen, The Netherlands
doi: 10.1002/9780470048672.wecb652

Terpenoids constitute the largest class of secondary metabolites in the plant kingdom. Because of their immense structural diversity and the resulting diversity in physiochemical properties, these molecules are particularly important for plant communication with other organisms. In this article, we will describe the ecological significance of terpenoids for plants, how terpenoid formation is regulated, and the tools we have to improve our understanding of the role of terpenoids in plant ecology and to create crop plants with improved resistance.

As plants are sessile, they cannot run away to avoid confrontation. Instead, they have evolved many different defense strategies, which include morphological (e.g., thorns, spines, and thick cuticle) and chemical defenses (e.g., repellents, toxic proteins, and toxic metabolites). Here, we will focus on the chemical defense of plants mediated by chemical compounds and one class of metabolites in particular: the terpenoids. They are of great importance to plants because of their multitude of functions in signaling and defense. Virtually all plant species have been shown to contain terpenoids and/or to release them from leaves, flowers, fruits, and roots into the environment to defend themselves—directly or indirectly—against herbivores and pathogens or to provide a reproductive advantage by attracting pollinators or seed dispersing animals.

As plants are sessile, they cannot run away to avoid confrontation. Instead, they have evolved many different defense strategies, which include morphological (e.g., thorns, spines, and thick cuticle) and chemical defenses (e.g., repellents, toxic proteins, and toxic metabolites). Here, we will focus on the chemical defense of plants mediated by chemical compounds and one class of metabolites in particular: the terpenoids. They are of great importance to plants because of their multitude of functions in signaling and defense. Virtually all plant species have been shown to contain terpenoids and/or to release them from leaves, flowers, fruits, and roots into the environment to defend themselves—directly or indirectly—against herbivores and pathogens or to provide a reproductive advantage by attracting pollinators or seed dispersing animals.

Terpenoids, which are also known as isoprenoids, constitute the most abundant and structurally diverse group of plant secondary metabolites, consisting of more than 40,000 different chemical structures. The isoprenoid biosynthetic pathway generates both primary and secondary metabolites that are of great importance to plant growth and survival. Among the primary metabolites produced by this pathway are phytohormones, such as gibberellic acid (GA), abscisic acid (ABA), and cytokinins; the carotenoids, such as chlorophylls and plastoquinones involved in photosynthesis; the ubiquinones required for respiration; and the sterols that influence membrane structure (see also Steroid and Triterpene Biosynthesis) (Fig. 1). Monoterpene (C10), sesquiterpenes (C15), diterpenoids (C20), and triterpenoids (C30) are considered to be secondary metabolites (Fig. 1). Many secondary metabolite terpenoids are of commercial interest because of their flavor, fragrance, or medicinal properties. Here, we will discuss the role of terpenoids in plant signaling.

Biological Background

Induction of biosynthetic pathways, mechanisms, and functions

Terpenoids are derived from the cytosolic mevalonate pathway or from the plastidial 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway (see also Terpenoid Biosynthesis). Both pathways lead to the formation of the C5 units isopentenyl diphosphate and its allylic isomer dimethylallyl diphosphate, which are the basic terpenoid biosynthesis building blocks (Fig. 1). Although increasing evidence suggests that exchange of intermediates occurs between these compartments, the cytoplasmic mevalonate pathway is generally considered to supply the precursors for the production of sesquiterpenes and triterpenes (including sterols) and to provide precursors for protein prenylation and for ubiquinone and heme-A production in mitochondria. In the plastids, the MEP pathway supplies the precursors for the production of isoprene, monoterpenes, diterpenes (e.g., GAs), and tetraterpenes (e.g., carotenoids).
After the formation of the acyclic precursors geranyl diphosphate, farnesyl diphosphate, and geranylgeranyl diphosphate, terpenoid scaffolds are generated through the action of terpene synthases (TPSs). Primary terpene skeletons formed by TPSs can be modified even more by the action of various other enzyme classes, such as the cytochrome P450 hydroxylases, dehydrogenases (alcohol and aldehyde oxidoreductases), reductases, glycosyl transferases, and methyl transferases (see also Chemistry of Cytochrome P450 Monooxygenases, Glycosyl Transferases and Methyl Transferases). localization of terpenoid biosynthesis

Terpenoids are implicated in several ecological and physiological functions and are often emitted from specific tissues at particular times related to their function. Many monoterpene and sesquiterpene synthase genes have been isolated and characterized from terpene-accumulating cells and tissues, such as leaf glandular trichomes, specific floral tissues, and fruits of agriculturally important plants. For plants that contain glandular trichomes, monoterpene production is considered to be localized exclusively in these organs. Resin ducts or oil glands can accumulate large amounts of terpenoids.

In floral tissues, volatile terpenes are often emitted at particular times to attract pollinators or to repel herbivores or microbial pathogens. Biosynthesis of the monoterpene β-ocimene and myrcene in snapdragon flowers, for example, is correlated with expression patterns of the corresponding genes in the flower petals that showed a (weak) diurnal oscillation under the control of a circadian clock (1). This finding indicates that these terpenes function as attractants for pollinating insects. In Arabidopsis, monoterpene and sesquiterpene synthases are not expressed in flower petals, but they are limited to the stigma, anthers, nectaries, and sepals (2), which suggest the importance of terpenoids for the defense of floral tissues against herbivores or microbial pathogens next to attraction of pollinators. Some terpene synthase genes exhibit expression in flowers as well as in fruits, whereas other genes are specifically expressed at particular stages of fruit development or ripening.

In contrast to the above-ground organs of plants, roots represent an unexplored area of terpene biosynthesis and function. To date, just a small number of terpene synthases have been identified in plant roots. In Arabidopsis, the terpene synthase genes that encode 1,8-cineole synthase and (Z)-γ-bisabolene synthase are expressed differentially in the stele of younger root growth zones and in the cortex and epidermis of older roots (reviewed...
in Reference 3). Little is known about the biological functions of volatile terpenes in roots at different developmental stages and in the interactions with root herbivores, microorganisms, and parasites. Terpene biosynthesis can be induced in roots under stress conditions, as shown in maize roots in response to attack by herbivores (4).

Terpenoids and Plant Signaling

Terpenoids are one of the major classes of compounds used by plants to communicate with their environment. This communication includes attraction of beneficial organisms (and unwanted attraction of herbivorous organisms) as well as defense against harmful organisms. With regard to the latter, plants employ direct as well as indirect defense mechanisms, against herbivores or fungal and bacterial pathogens. Direct defense includes physical structures, such as thorns and trichomes, and the accumulation of toxic metabolites or proteins to deter or even kill attackers. In many plant species, diterpenes and sesquiterpenes act as phytoalexins, which are low-molecular-weight compounds that are produced as part of the plant defense system against microorganisms. Compounds such as the bitter sesquiterpene cucurbitacin and the pungent sesquiterpenoid polygodial have been shown to be involved in direct defense against insects (Fig. 2).

Indirect defense implies that plants defend themselves against herbivores by enhancing the effectiveness of the natural enemies of these herbivores. These defense mechanisms can be constitutive like the formation of domatia, which serve as homes for ants and nites, or the production of foliar nectaries and nutritional structures that can be used by natural enemies of the herbivores. In addition, indirect defense mechanisms in plants can be induced. One of the most intriguing examples of this function is the emission of herbivore-induced plant volatiles, which attract the carnivorous natural enemies of herbivores. These herbivore-induced plant volatiles predominantly consist of terpenoids that mediate many interactions in a plant–insect community, both above and below ground (4, 5). The volatiles that plants produce in response to herbivore damage can affect various other interactions of the plant with community members (5, 6). Moreover, herbivore-induced plant volatiles can affect herbivore-plant and carnivore-herbivore interactions on neighboring plants through their effect on the neighbor’s phenotype (5, 7).

Terpenoids in direct defense

Direct-defense compounds can be either constitutively present in specific parts of the plant or be produced after induction by pathogens or herbivores. The latter compound will be less costly for the plant. For example, elicitor-induced accumulation of the antimicrobial sesquiterpenoid capsidiol correlated with the induction of 5-epi-aristolochene synthase, which is a branch-point sesquiterpene cyclase involved in the synthesis of sesquiterpene phytoalexins (8). In rice (Oryza sativa L.), 14 diterpenoid phytoalexins have been identified. All these compounds are accumulated in rice leaves after inoculation with the pathogenic blast fungus Magnaporthe grisea and exhibit antimicrobial properties (9). Another example is pyropolyp, which has generated considerable interest because of its potent insect antifeedant activity. In cotton (Gossypium spp.), gossypol and related sesquiterpene alkaloids, which are all derived from (-)-δ-cadinene, provide both constitutive and inducible protection against pests and diseases.

Figure 2. An overview of terpenoid-mediated interactions between plants with the surrounding environment. Floral scent to attract pollinators (1, e.g., β-myrcene); Protection of reproductive organs from fungal or bacterial infection (2, e.g., thujaplicin); Direct defense: repellency of herbivorous insects by volatile terpenoids (3, e.g., zingibene); Attraction of predators and parasitic wasps on insect or spider mite herbivory (4, e.g., 4,8-dimethyl-3(2H)-menthene); Priming or elucidation of defense in neighboring plants (5, e.g., α-cadinene). Defense compounds in beans against insect herbivores, fungi and bacteria (6, e.g., polygodial), fruit aroma to stimulate consumption and thereby seed dispersal (7, e.g., β-caryophyllene); Germination of parasitic plant seeds (8, e.g., strigol); Stimulation of growth and attachment of symbiotic mycorrhizal fungi (9, e.g., δ-desmosterol); Attraction of entomopathogenic nematodes after root feeding of beetle larvae (10, β-caryophyllene); Direct defense to protect against insect herbivores, fungal infection, or bacterial infection (11, rhamnol).
Terpenoids in Plant Signaling, Chemical Ecology

<table>
<thead>
<tr>
<th>Cucumber</th>
<th>Spider mite infested</th>
<th>Mechanically damaged</th>
<th>Control, undamaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potato</td>
<td>Spider mite infested</td>
<td>Thrips infested</td>
<td>Control, undamaged</td>
</tr>
</tbody>
</table>

**Figure 1** Induction of terpenoids in cucumber and potato leaves after herbivory by spider mites (*Trionyxus unarius*) or thrips (*Frankinia occidentalis*) (Kappers R, Buskensmeier HJ, Dieder M, Unpublished results).

(Reviewed in Reference 12). Within a plant species, the quality of the volatile blend may be affected by the developmental stage of the herbivore, and the volatiles produced by plants upon insect egg deposition may differ from the one induced by feeding (11). Typical volatiles released from a multitude of species after herbivory are the so-called green leaf volatiles such as C6-alcohols, -aldehydes, and -esters, which are derivatives of the shikimate pathway such as methyl salicylate, and terpenoids such as (E)-4,8-dimethyl-1,3(*E*,*E*)-7-nonatriene, and 4,8,12-trimethyl-1,3(*E*,*E*)-11-tetradeca-tetraene. The terpenoids are by far the most important contributors to the induced volatile blend. To date, little information exists about how natural enemies respond to individual components found in induced odor blends, even though it is known whether they can distinguish between complex odor mixtures, A. Also, it is still unclear whether natural enemies use only a few compounds present in an odor blend for prey identification, or whether they use information from all odor compounds. For a multitude of tritrophic systems, it was shown that predatory mites and parasitoid wasps were attracted by terpenoid components from induced volatile blends, for instance (E)-4,8-dimethyl-1,3(*E*,*E*)-7-nonatriene and 4,8,12-trimethyl-1,3(*E*,*E*)-11-tetradeca-tetraene (4, 6, 12).

**Terpenoids in Aboveground Plant–Plant Communication**

Volatiles released from herbivore-infested plants mediate plant-plant interactions and may induce the expression of defense genes and emission of volatiles in healthy leaves on the same plant or of neighboring unattacked plants, thus increasing their attractiveness to natural enemies and decreasing their susceptibility to the damaging herbivores (Fig. 2) (reviewed in Reference 13). This phenomenon is called priming, and it prepares neighboring plants to respond more rapidly and intensively against subsequent attack by herbivorous insects (13), for example by increasing extra-floral nectar secretion to attract predatory arthropods (14).

Several reports showed that herbivore- and elicitor-induced plant volatiles, in particular green leaf volatiles and terpenoids, influence gene expression and result in priming of defense responses of neighboring conspecific and nonconspecific plants that were not attacked. Terpenoids emitted from herbivore-infested *Nicotiana attenuata* enhance the expression of numerous genes of neighboring conspecifics (15). One example is (E)-4,8-dimethyl-1,3(*E*,*E*)-7-nonatriene that can act as plant-plant signal by upregulating signaling pathways of *jasmonic acid* and ethylene in neighboring plants (16).

**Terpenoids in rhizosphere communication**

In the rhizosphere, plants use terpenoids for communication with other organisms. The fact that only a few such relationships have been demonstrated probably more reflects the difficulty of studying chemical signaling in the soil than the actual contribution of rhizosphere signaling to plant functioning. Surprisingly, among those rhizosphere-signaling relationships that have been uncovered, several constitute the attraction of pathogenic organisms by highly specific signaling molecules. Examples are the hatching of cyst nematodes, which is triggered by terpenoids (e.g., soybean cyst nematodes by glycoeclepin and potato cyst nematodes by solanoeclepin) (Figs. 1 and 2). From an evolutionary point of view, one must conclude that these molecules must have another, as yet unknown, positive function for plants or otherwise they would have been selected against. This function is illustrated by the example of strigolactones, which are apocarotenoid-signaling molecules that are secreted by the roots of many plant species. In the 1960s, these compounds were identified as the germination stimulants that trigger germination of the seeds of the root parasite Orobanchaceae (*Striga* spp. and *Orobanche* spp.) (17). These obligate parasites can only survive if they grow on the roots of a host plant from which they take water, assimilates, and nutrients. Several *Striga* and *Orobanche* species can be a nuisance in agriculture, where they can destroy complete harvests. To prevent the tiny seeds from germinating at too large distance from a host root, parasitic plants have evolved a requirement for so-called germination stimulants, collectively called the strigolactones, which are compounds that are produced by the roots of their hosts (Figs. 1 and 2).

The reason for the existence of the strigolactones remained unknown until 2005 when it was discovered that they are an important host-finding factor for arbuscular mycorrhizal (AM) fungi (18). In the arbuscular mycorrhizal symbiosis, plants obtain water and mineral nutrients from their fungal partners, which allow them to survive under various stressful conditions. AM fungi are obligate symbionts that have facilitated the adaptation of primitive plant species to life on land and colonize the roots of most land plant species. Apparently, mycorrhizal symbiosis required the production of strigolactones throughout the plant kingdom and then, indirectly, allowed for the later evolution of the host detection mechanism of parasitic plants using the same compounds (17).

Tritrophic interactions, for which so much evidence exists aboveground, are gradually being uncovered in the rhizosphere.
Regulation of Terpenoid Formation

Biotic and abiotic factors affecting terpenoid formation

The production of secondary metabolites by plants has been shown by many authors to be influenced by environmental conditions. Therefore, it can be expected that this finding holds for terpenoids involved in signaling. Knowledge of these effects may be important for improvement of biological control, resistance against attackers, or attraction of beneficial organisms (e.g., by applying the most optimal conditions for efficient signaling molecule production). The factors that have been shown to affect secondary metabolite production such as light, temperature, and water availability have also been investigated for their effect on herbivore-induced volatile biosynthesis. High light intensity and water stress are generally reported to increase induced volatile production and/or predator attraction for example in lime bean, kidney bean, maize, and cotton (reviewed in Reference 19). Fertilization had a strong positive effect on emission of induced volatiles in maize, even when results were corrected for plant biomass (19). Hence, climatic conditions and nutrient availability can be important factors in determining the intensity and variability in the release of induced plant volatiles.

In rhizosphere signaling, environmental conditions affect the production of signaling molecules. A particularly clear case is the production of strigolactones, which are the host-presence signaling molecules for AM fungi (see above). AM fungi help plants to absorb nitrogen and phosphate and hence improve plant growth in areas of the world where the concentration or availability of particularly phosphate in the soil is limited (17). Interestingly, root exudates of red clover, tomato, and rice grown under phosphate limitation produce much more strigolactones. It can be argued that in this way, plants improve the chance to attract AM fungi as quickly as possible. After colonization by AM fungi, plants seem to produce less strigolactones, which would be consistent with improved phosphate availability.

Genetic variation

Variability in induced plant volatiles complicates the reliance of natural enemies on these cues. One way of dealing with variability is through associative learning, which may allow parasitoids to learn which cues are most likely to lead them to suitable hosts at a particular time in a particular area. Moreover, recent studies suggest that plant volatile blends alone carry specific information on the herbivores by which they are attacked. For example, predatory mites can distinguish between the blends of apple trees infested by two herbivores species (20). Du et al. (21) showed that different aphid species elicit different volatile blends in bean plants and that the aphid parasitoid, Aphidius ervi, can use these differences to distinguish plants infested by its host, Aphis pisoni from those infested by a nonhost, Aphis fabae.

Little is known about the genetic variability in such herbivore-induced plant signals and about how the emissions in cultivated plants compare with those of their wild relatives. For conventional plant breeding for improved biological control through enhanced volatile production and hence predator attraction to be successful, genetic variation in the ability to produce herbivore-induce volatile-attracting volatiles is a prerequisite. The little information available on the extent of the variability comes mostly from studies on cultivated plants. Although rice plants infested with the brown planthopper, Nilaparvata lugens, were more attractive than uninfested plants, Rupas et al. (22) showed that constitutively produced rice volatiles in 6 out of 15 uninfested rice cultivars also attractive to the predator Corcythorinus lividipennis when compared with clean air. In gerbera, several cultivars differed in composition and amount of volatiles produced in response to spider mite feeding (23). Y-tube olfactometer experiments revealed differences between the gerbera cultivars in the odor preference of predatory mites. The composition of the volatile blend seemed to be more important for this difference than the total amount of volatiles produced, and particularly the terpenoids (E)-β-caryophyllene and limonene were mentioned by the authors as possibly important candidates in determining the difference in attractiveness between cultivars. Between maize cultivars and between different Zea species, large differences were found in the composition of the volatile blend induced by the application of the oral secretion of Spodoptera littoralis to mechanically damaged leaves (24).

Several problems have been associated with the comparison of genotypes for their production of induced volatiles when other differences between the genotypes can not be controlled (23). For example, differences may exist in direct defense between genotypes that cause differences in developmental rate of herbivores that may lead to differences in volatile formation. To circumvent this problem, in addition to spider mite infestation, we used jasmonic acid treatment in a comparison between seven cucumber genotypes. Earlier research had shown that jasmonic acid treatment mimics the effect of spider mite infestation in several plant species. Different cucumber genotypes.
produced different volatile blends on jasmonic acid treatment or spider mite infestation that are reflected in differences in the attractiveness of these genotypes to predatory mites (Kappers, Bouwmeester, Dicke, unpublished results).

Role of Terpenoids in Other Aspects

Terpenoids in plant reproduction

To attract pollinators and seed-dispersing animals and to ensure reproductive and evolutionary success, many flowering species release blends of volatile compounds from their flowers and fruits in addition to visual and tactile cues (Fig. 2). The bio-genetic pathways of fruit and flower volatiles can be derived from enzymatically controlled lipid, terpene, amino acid, carbohydrate, and phenylpropane metabolism. Floral scent bouquets may contain from 1 to 100 different volatiles, but most species emit between 20 and 60 different compounds (25). The total amount of emitted floral volatiles varies from the low picogram range to more than 30 µg/hour (25). Although flowers could be identical in their color or shape, no two floral scents are exactly the same because a large diversity of volatile compounds and their relative abundance and interactions within the scent bouquet. In addition to attracting insects to flowers and guiding them to food resources within the flower, floral volatiles are essential in allowing insects to discriminate among plant species and even among individual flowers of a single species (1).

To date, little information exists about how insects respond to individual components found in floral scents, even though it is known that they can distinguish between complex floral scent mixtures. It is still unclear whether insect pollinators use only a few compounds present in a scent for floral identification or whether they use information from all scent compounds. Recently, it was shown that honeybees can use all floral volatiles to discriminate subtle differences in the scent of four snapdragon cultivars that emit the same volatile compounds but at different levels (26). Floral volatiles could play many roles instead of or in addition to pollinator attraction. For example, many terpenes, including β-myrcene, (E)-β-ocimene, linalool, and (E)-β-caryophyllene, react readily with ozone and other reactive oxygen species (27). Thus, floral volatiles could function to protect the reproductive organs from oxidative damage. A variety of monoterpens and sesquiterpenes is reported to have antimicrobial activity (28). Hence, floral terpenes could help defend floral organs, like the moist stigma, from bacterial or fungal infection.

Practical Applications and Future Prospects

Practical applications

The great significance of terpenoids in mediating the interactions of plants with other organisms has prompted many researchers to explore the possibilities of using this knowledge to improve resistance of plants against attacking organisms. Now that many induced volatile blends have been identified, artificial mixtures could be composed that are effective in attracting natural enemies and could be used in crops. Alternatively, crops could be sprayed with jasmonic acid to induce volatile production that should lead to the increased presence of natural enemies. Occasionally, an example exists in which this approach has been unsuccessful (29), and several authors have expressed the feeling that this approach should fail in the long term, as the presence of the volatile cue and a prey are uncoupled.

If the volatile cue and the presence of a prey are not uncoupled (i.e., attractive volatiles are only or mainly produced on herbivory), then an adequate response of the crop to herbivory is most important. We have reviewed several studies in which the effect of environmental conditions on volatile production and herbivore attraction has been demonstrated, and researchers should take these results into account when designing their experiments. However, for a practical application such as the optimization of biological control, these factors may be important. It would be of interest to observe whether environmental conditions that stimulate induced volatile formation actually improve biological control in a field situation. A rather as yet completely ignored factor in the optimization of biological control is the selection for genotypes with improved (faster, stronger) response. Our results on cucumber and the results of other plant species demonstrate that genetic variation for this response is available. Additional research could demonstrate the effectiveness and the best and easiest way to exploit this variation in breeding.

Metabolic engineering

Several research groups have made significant progress with the metabolic engineering (see also Metabolic Engineering) of particularly monoterpene and sesquiterpene biosyntheses in a range of plant species (reviewed in Reference 30). These studies have shown that a high production rate of terpenes, including modified products, can be obtained using metabolic engineering. The importance of terpenoids in the interaction of plants with other organisms implies that their modification by plant metabolic engineering will have major effects on their response to the environment. Petunia plants that express the Clarkia breweri linalool synthase showed a delayed and less severe natural infection by mildew than the nontransformed plants under standard greenhouse conditions. Fruit of tomato plants transformed with the same gene were much more resistant to postharvest pathogens than the nontransgenic controls (30). Some effects of transgenic, volatile producing plants on insects have been reported. Transgenic tobacco plants transformed with three lemon monoterpene synthases were visited much less by herbivorous insects (e.g., whiteflies) but more by fruit flies than wild-type tobacco plants in the same greenhouse compartment (30).
choice assays, Arabidopsis plants transformed with the strawberry linalool/nerolidol synthase, which emit greater linalool levels than the control plants, significantly repelled the aphid Myzus persicae (reviewed in Reference 30). Recently, these observations were extended with even more convincing results: Transgenic chrysanthemum (Chrysanthemum x grandiflorum) producing linalool repelled western flower thrips (Frankliniella occidentalis) (reviewed in Reference 30). Increased levels of the diterpenoid cembratienols in trichome exudates of the transgenic tobacco plants resulted in greater resistance to aphids, which also occurred in field tests (reviewed in Reference 30). In studies on tritrophic interactions, transgenic plants become an important tool. A transgenic plants transformed with the strawberry linalool/nerolidol synthase with mitochondrial targeting signal emitted (35S)-E-nerolidol and (E)-DMNT and were attractive to carnivorous predatory mites (Phillasselus persimilis), which are the natural enemies of spider mites (reviewed in Reference 30). Transgenic Arabidopsis plants engineered for the production of sesquiterpenes, which is normally emitted by maize, attracted females of the parasitoid Cotesia marginiventris that located their lepidopteran hosts (the parasitoids were first exposed to the volatiles in association with their hosts) (reviewed in Reference 30). The sesquiterpene (E)-β-farnesene acts as alarm pheromone for many species of aphids, which causes dispersion in response to attack by predators or parasitoids. Overexpression of an (E)-β-farnesene synthase cloned from Melita s piperita, in Arabidopsis thaliana, yielded plants that had high emission of pure (E)-β-farnesene (31). These plants elicited potenti potential effects on behavior of the aphid Myzus persicae (alarm and repellent responses) and its parasitoid Diaeretiella rapae (an arrestant response).

It is conceivable that changes in the (induced) volatile production in commercial crops could lead to the development of biological control packages in which biological control agents trained specifically for the modified crop are included. It will be exciting to see whether these approaches can lead to plants with altered (improved) predator behavior and to crops with improved biological control.

Acknowledgments

This work was supported in part by the Technology Foundation (STW) of the Netherlands Organisation for Scientific Research (NWO) to F.K. (WP 5479) and the Netherlands Organisation for Scientific Research (NWO) to M.D. and J.B.J. (Viet-grants).

References


Further Reading


See Also

Chemistry of Cytochrome P450 Monooxygenases, Glycosyl Transferases and Methyl Transferases
Metabolic Engineering
Steroid and Triterpene Biosynthesis
Terpenoid Biosynthesis
Terpenoids in Plants
Christopher I. Keeling, Michael Smith Laboratories, University of British Columbia, British Columbia, Canada
Jörg Bohlmann, Michael Smith Laboratories and Departments of Botany and Forest Sciences, University of British Columbia, British Columbia, Canada
doi: 10.1002/9780470048672.wecb596

Terpenoids are the largest class of all known natural products. Plants produce a variety of terpenoid compounds that number in the thousands. Some terpenoids are involved in plant growth and development directly (i.e., in primary metabolism), but most plant terpenoids are thought to function in interactions of plants with their biotic and abiotic environment and have traditionally been referred to as secondary metabolites. In addition to the isolation and identification of plant terpenoids, research has concentrated on the biosynthesis, the biological function, and the exploitation of plant terpenoids for human use as biomaterials and pharmaceuticals. Plant terpenoids are biosynthesized from C5 precursors by the action of prenyl transferases and terpenoid synthases. Often, terpenes are acted on by cytochromes P450 and other enzymes to increase their functionalization. Terpenoid biosynthesis in plants involves several subcellular compartments. The accumulation of terpenoids requires efficient transport systems and specialized anatomical structures. Using isoprene (a hemiterpene), menthol (a monoterpene), artemisinin (a sesquiterpene), and paclitaxel (better known under the registered trademark Taxol (Bristol Myers Squibb, New York)) and diterpene resin acids (diterpenes) as examples, we highlight some strategies, techniques, and results of plant terpenoid research with a strict focus on the low-molecular-weight (C5–C20) terpenoids of specialized plant metabolism.

Biological Background

Over 20,000 terpenoids have been identified (1), and more are being discovered continuously. Plant terpenoids are important in both primary and secondary (specialized) metabolism. Their importance in primary metabolism includes physiological, metabolic, and structural roles such as plant hormones, chloroplast pigments, roles in electron transport systems, and roles in the posttranslational modification of proteins. In secondary metabolism, the roles of plant terpenoids are incredibly diverse but are associated most often with defense and communication of sessile plants interacting with other organisms. Examples include terpenoid chemicals that form physical and chemical barriers, antibiotics, phytoalexins, repellents and antifeedants against insects and other herbivores, toxins, attractants for pollinators or fruit-dispersing animals, host/nonhost selection cues for herbivores, and mediators of plant-plant and mycorrhiza interactions (2, 3).

Some plants produce terpenoids in specialized cells or tissues such as the glandular trichomes on the surface of peppermint...
leaves, scent-releasing epidermal cells of certain flowers, or the epithelial cells that surround the resin ducts of conifers. These structures place high concentrations of terpenoids in areas most likely to be encountered by the interacting organism.

The biosynthetic pathways of plant terpenoids are highly regulated and highly spatially organized in subcellular compartments and sometimes in specialized cells. Terpenoid biosynthesis can be regulated by plant hormones, developmental programs, diurnal cycles, herbivory, or pathogen infection. Identifying what role specific terpenoids play in plants, how and where they are biosynthesized, and how their biosyntheses are regulated allows us to better understand their importance to the survival of the plant and thus make use of this knowledge in crop improvement or in the production of terpenoids for medical or industrial uses.

**Chemistry**

Much chemistry research in plant terpenoids has been to elucidate the structure, define the biosynthetic pathways, characterize the enzymes involved, and develop systems for the large-scale production of medicinal or industrially important terpenoids (4). Progress on the identification and the study of plant terpenoids is reviewed regularly in the journal *Natural Product Reports,* and the biosynthesis of terpenoids, including plant terpenoids, has been reviewed comprehensively (5). Plant terpenoids can be volatile or nonvolatile, lipophilic or hydrophilic, cyclic or acyclic, chiral or achiral, and they often have double-bond stereochemistry. The chemical diversity of terpenoid structures originates largely from the terpenoid synthases that stabilize different carbocation intermediates, allow rearrangements or water termination, and direct stereochemistry; the diversity also originates from the many different terpenoid-modifying enzymes.

Two major complementary approaches to studying plant terpenoids have been established. One approach involves the isolation and the structural identification of terpenoid chemicals of interest from plant tissues based on traditional natural product research followed by targeted search for the relevant enzymes and genes that control biosynthesis. The second approach explores the emerging plant genome sequences to discover complete sets of genes that encode terpenoid biosynthetic enzymes. The combination of these two approaches is the most powerful approach to a comprehensive understanding of plant terpenoid chemistry and its biosynthetic origins.

The diversity of plant terpenoids reflects the complexity and the diversity of the pathways that biosynthesize them. The recent sequencing of the genomes of four different plant species and large collections of expressed sequence tags (ESTs) from many other plants may indicate the diversity of pathways and chemicals we might expect in any one species. For example, the genes that encode putatively active terpenoid synthases (TPS) comprise at least 32 in the Arabidopsis (Arabidopsis thaliana) genome (6), at least 15 in the rice (Oryza sativa) genome (7), at least 47 in the poplar (Populus trichocarpa) genome (8), and at least 89 in the genome of a highly inbred grapevine (Vitis vinifera) Pinot Noir variety (9). The large gene family of TPS, which is important to generating structural diversity of terpenoid chemicals, apparently results from repeated gene duplication and subsequent neofunctionalization or subfunctionalization (10, 11). Most TPS produce more than one product from a single substrate, and these products are often modified by the action of additional enzymes such as cytochromes P450 and reductases. Thus, the number of distinct terpenoids found in any one plant species is predicted to be manifold higher than the number of TPS genes present in that species.

Genomics approaches, which can identify the candidate genes for terpenoid production, together with functional characterization of heterologously expressed enzymes and the identification of the resulting plant terpenoids, can enhance the discovery of the biochemical pathways substantially in plants as has been demonstrated in recent years with research in Arabidopsis (12), rice (13), and grapevine (14). Ideally, the functional genomics approach is combined with classical and modern approaches of isolation, identification, and metabolite profiling of terpenoids from plant tissue.

The combined genomics and chemical approaches to plant terpenoid research are not restricted to the few plant species for which more or less complete genome sequences are now available. The discovery of many of the genes and enzymes for the formation of terpenoids such as menthol and related monoterpenes in peppermint (Mentha x piperita) (15), artemisinin in Artemisia annua (16), taxol in the yew tree (Taxus) (17), and conifer diterpene resin acids in species of spruce (Picea) and pine (Pinus) (18) have been possible on the foundation of highly specialized efforts of EST and full-length cDNA sequencing combined with characterization of recombinant enzymes and analysis of the terpenoid metabolome of the target plant species.

**Common steps in plant terpenoid biosynthesis**

The universal precursors to terpenoids, the C5-compounds dimethylallyl pyrophosphate (DMAPP) and isopentenyl pyrophosphate (IPP), originate from two pathways in plants (Fig. 1). The mevalonate (MEV) pathway is well described in many eukaryotic organisms. This pathway is present in the cytosol/endoplasmic reticulum of plants. More recently, another pathway has been described, the 2C-methyl-D-erythritol-4-phosphate (MEP) pathway, which is found in the plastids of plants (19). The localization of the different pathways and the plastid-directing transit peptides found in hem-TPS, mono-TPS, and di-TPS, but not in sesqui-TPS, result in the production of terpenoids from at least two different precursors pools.

Hemiterpenoids are produced from the isoprenyl diphosphate (DMAPP). All other terpenoids are produced from DMAPP and IPP via longer-chain prenyl diphosphate intermediates formed by prenyl transferases. Prenyl transferases (20) catalyze the formation of geranyl diphosphate (GPP), farnesyl diphosphate (FPP), and geranylgeranyl diphosphate (GGPP) from one molecule of DMAPP and one, two, or three molecules of IPP, respectively (Fig. 2). Isoprenyl diphosphates are the substrates for all TPS, which lead to the hemiterpenoids, monoterpenoids, sesquiterpenoids, and diterpenoids, which will be highlighted with selected examples in the following sections.
Isoprene (C5, hemiterpene)

Isoprene (2-methyl 1,3-butadiene, Fig. 2) is the simplest terpenoid produced by many plants, and is produced abundantly by some tree species such as poplars. The mechanism of isoprene biosynthesis is a straightforward diphosphate ionization of the DMAPP precursor followed by deprotonation of the allylic cation (Fig. 2). Only a few isoprene synthases have been identified and characterized in plants, from Populus spp. (poplar) (21, 22) and Pueraria montana (kudzu vine) (23). These TPS contain a transit peptide that targets the plastids, and thus isoprene biosynthesis is derived from products of the MEP pathway.
isoprene is a major biogenic volatile organic compound (VOC; rivalling methane in global production) with emission from plants estimated in the order of more than 10^12 kg per year and therefore has been well studied for its role in atmospheric chemistry (24). Isoprene is also important in the context of global cycles of carbon fixation versus carbon emission from natural sources. Although isoprene is emitted in large amounts from poplars, which are actively being promoted as plantation species for biofuel (ethanol) production, its physiological function in plants is somewhat unclear. The protection from thermal and oxidative stress as well as release of excess carbon flux and photosynthetic energy are thought to be the main functions of isoprene in plants (21, 22).

Very recently, new molecular approaches have established a function of isoprene in thermotolerance through the use of under- and over-producing transgenic lines of poplar (25) or Arabidopsis (which normally produces no isoprene) using the poplar isoprene synthase (26). Other possible functions of isoprene remain to be tested using similar molecular approaches. The effect of down-regulation of isoprene emission in poplars remains to be tested in the field to explore whether it is viable to reduce emission of this biogenic VOC in plantation forests for biofuels production to maximize carbon fixation and minimize carbon emission. A closely related hemiterpene of plant origin isoprene remain to be tested using similar molecular approaches. The effect of down-regulation of isoprene emission in poplars is also under investigation in order to explore whether it is viable to reduce emission of this VOC in plantation forests for biofuels production to maximize carbon fixation and minimize carbon emission. A closely related hemiterpene of plant origin isoprene remain to be tested using similar molecular approaches. The effect of down-regulation of isoprene emission in poplars remains to be tested in the field to explore whether it is viable to reduce emission of this biogenic VOC in plantation forests for biofuels production to maximize carbon fixation and minimize carbon emission.

First, using geranyl diphosphate as substrate, (-)-limonene synthase generates (-)-limonene and minor amounts of myrcene, (-)-alpha-pinene, and (-)-beta-pinene. (-)-Limonene then undergoes a series of transformations ultimately to yield (-)-menthol. These modifications involve first the allylic hydroxylation to (-)-trans-isopiperitenol by the cytochrome P450 limonene-3-hydroxylase (CYP71D13). Allylic oxidation of this alcohol to (-)-isopiperitenone is then catalyzed by the NAD-dependent isopiperitenol dehydrogenase. Subsequently, NADPH-dependent (-)-isopiperitenone reductase catalyzes the formation of (+)-cis-isopulegone. (+)-cis-Isopulegone is enzymatically isomerized to the more stable alpha, beta-unsaturated ketone (+)-isomenthone—by (+)-cis-isopulegone isomerase. (-)-Menthone and (+)-isomenthone (in a 2:1 to 10:1 ratio) are then formed by the action of the NADPH-dependent (-)-isopulegone reductase. Finally, (-)-menthone reductase reduces (-)-menthone to (-)-menthol. All of these enzymes and their corresponding genes in Mentha have been isolated, functionally characterized, and their enzymology studied (15). For some enzymes, substrate specificities or product outcomes after directed mutations as well as enzyme localization have been investigated. The structure of (-)-limonene synthase has been determined recently (29). Based on the detailed knowledge of (-)-menthol and its biosynthetic pathway, it has become possible to improve the composition of the monoterpene-rich essential oil of Mentha through metabolic engineering (30, 31).

Artemisinin (C_{15}, sesquiterpenoid)

Artemisinin is used here as an example of a plant sesquiterpenoid with both traditional value as well as medicinal and social value in the twenty-first century. Research on artemisinin has also established new benchmarks for biochemical engineering and functional genomics of plant terpenoids. Artemisinin is a functionalized sesquiterpene with a unique peroxide linkage from the sweet wormwood (Artemisia annua). Chinese herbalists have used it since ancient times, and it is now used for its unique efficacy to treat multidrug-resistant strains of the malaria parasite Plasmodium falciparum. Its medicinal importance has prompted studies into its biosynthesis and its biochemical engineering so that cost-effective methods for producing it in large scale and in consistent quality may be realized.

Biosynthesis of artemisinin in planta begins with the formation of the sesquiterpene amorphophyll-4,11-diene in glandular trichomes of A. annua leaves (Fig. 4) by amorphophyll-4,11-diene synthase (32, 33). Amorphophyllene is oxidized to artemisinic acid in three steps by a multifunctional cytochrome P450 (CYP71AV1) (16). The remaining steps have not yet been established but are predicted to include nonsymmetrically catalyzed photooxidation reactions (34). The application of a semi-synthetic route from artemisinic acid to artemisinin, along with the availability of the characterized plant enzymes to produce artemisinic acid described above from A. annua, have permitted the complete synthesis of artemisinin via microbial host cells (16). Introducing these enzymes into Escherichia coli or Saccharomyces cerevisiae and engineering an unnatural, or fine-tuning the natural, mevalonate pathway in these microorganisms have resulted in significant production of artemisinic acid in fermentations (14, 35, 36).
Terpenoids in Plants

Figure 3 Pathway of (−)-menthol biosynthesis in Mentha. LS, (−)-limonene synthase; L3OH, (−)-limonene-3-hydroxylase; PD, (−)-trans-isopinocampheol dehydrogenase; PR, (−)-isopinocampheol reductase; IP, (−)-cis-isopulegone isomerase; P, (−)-pulegone reductase; MR, (−)-menthone reductase.

Figure 4 Pathway of artemisinin biosynthesis in Artemisia annua. ADS, amorphadiene synthase; CPR, cytochrome P450 reductase.

Taxol (C20, diterpenoid)

Taxol is another example of a medicinally important functionalized plant terpene, in this case a diterpenoid, with large pharmaceutical and economic value (Fig. 5). Taxol is a potent anticancer drug that was isolated and identified from the bark of the Pacific yew (Taxus brevifolia) more than 35 years ago (37). The name Taxol is now a registered trademark, but the literature commonly uses this name rather than the generic name paclitaxel. The total synthesis of Taxol is possible (38), but it is not economically feasible currently because of the challenges of stereochemistry, low yield, and high cost. The study of Taxol by Croteau et al. (17) is an exceptional example of how a terpenoid biosynthetic pathway was rationalized, and the synthesis and testing of various hypothetical precursors with cell-free extracts have yielded the discovery of many enzymes and genes in this complex pathway. Because Taxol is only found at very low levels in slow growing trees, and is one of hundreds of Taxol-like compounds produced in a metabolic grid, the use of an inducible Taxus cell culture system has accelerated this research.

Taxol is biosynthesized in 19 steps from GGPP that originates from precursors of the MEP pathway (Fig. 5). The biosynthesis of Taxol begins with the formation of the tricyclic diterpene skeleton of taxa-4(5),11(12)-diene (17, 39). All genes for the enzymes in this early pathway have been identified in Taxus cuspidata and a taxa-4(5),11(12)-diene synthase has been identified in several Taxus species. The mechanism of this di-TPS has been explored in detail (39). Taxa-4(5),11(12)-diene is then hydroxylated by several cytochrome P450 taxoid oxygenases to yield a putative intermediate decorated with seven alcohol or ester groups. Many cytochrome P450 enzymes that catalyze these transformations have been identified and characterized, but two enzymes remain uncharacterized (17). The biosynthesis of the ester functionalities have been studied, and several acyl and aryl transferases have been identified and characterized. Finally, studies on the steps in the aromatic side chain assembly and attachment have yielded several enzymes that include a phenylalanine aminotransferase, a C13-phenylpropanoyl-CoA transferase, and an N-benzoyl transferase. Although the pathway of biosynthesis has not been resolved fully, what is known can be applied to improve Taxol production, which continues to rely on Taxus plants or cell cultures.

Diterpene resin acids (C20, diterpenoids)

Diterpene resin acids are abundantly produced in conifers of the pine family (Pinaceae) and in other plant species (Fig. 6). They are produced in the epithelial cells that surround the resin ducts that are found constitutively or in resin ducts induced in the xylem upon wounding and are important for the physical and chemical plant defenses against herbivores and pathogens (18, 40). Industrially, diterpene resin acids are important chemicals for the naval stores industry, in printing inks, as potential antimicrobials and pharmaceuticals, and are byproducts of wood pulping processes.

Two major steps exist in the biosynthesis of diterpene resin acids: the formation of the diterpene and the stepwise oxidation of the diterpene to the corresponding acid. Most conifer diterpenes are tricyclic; they are biosynthesized by bifunctional di-TPS that first cyclize geranylgeranyl diphosphate to (−)-copaeryl diphosphate and then cyclize this intermediate even...
Terpenoids in Plants

Figure 5 Pathway of Taxol biosynthesis in Taxus spp.

bacatin III, R₁ = Ac 10-deacetyl-baccatin III, R₁ = H

Stereochemistry often is an integral component to both the chemical structure and the biological function of plant terpenoids. For volatile terpenoids, chiral GC stationary phases provide the enantiomeric separation for quantitative analysis, and, provided an authentic standard of known absolute stereochemistry is available, the availability of retention index information on at least two columns of different polarity increases the confidence in structural assignment. Such information is available for the more volatile terpenoids involved in flavors and fragrances (46), and insect semiochemicals (47). For nonvolatile terpenoids, LC-MS is often used, but there are no comparable databases of terpenoids for LC-MS as there are for GC-MS.

Structural determination

Hanson (43) has published an excellent review of the methods and the strategies for rigorous structural determination of terpenoids. Routine survey-style analyses of plant terpenoids often are not so rigorous in structural assignment. Databases of mass spectra [such as Wiley (New York) and NIST (Gaithersburg, MD) MS databases] or databases that combine mass spectra and retention indices (44, 45) facilitate assignment of some commonly encountered plant terpenoids. Of course, these databases are only helpful if they include the specific terpenoids that are being analyzed and the likely alternatives. Sometimes, in addition to having similar mass-spectral fragmentation patterns, two terpenoids may share very similar retention indices on the 5% phenyl methyl polysiloxane GC column used for these databases, which makes structural assignment without additional information impossible. The availability of retention index information on at least two columns of different polarity increases the confidence in structural assignment. Such information is available for the more volatile terpenoids involved in flavors and fragrances (46), and insect semiochemicals (47). For nonvolatile terpenoids, LC-MS is often used, but there are no comparable databases of terpenoids for LC-MS as there are for GC-MS.
Figure 6. Pathway of conifer diterpene resin acid biosynthesis. Bifunctional di-TPS convert geranylgeranyl diphosphate to various diterpenes, which are oxidized stepwise by multisubstrate and multifunctional cytochromes P450 to the corresponding diterpene acid. The oxidation of (-)-abietadiene to (-)-abietic acid is shown as an example.
configuration is available, elution comparison can establish absolute configuration of the unknown sample (49). Chiral phases for liquid chromatography can also be used to resolve enantiomers, both analytically and preparatively (50). Often, chiral synthesis is necessary to clearly establish the absolute configuration of an unknown terpenoid.

Molecular biology

The recent availability of plant genome sequences and the methods to clone homologous genes from different plant species using molecular biology approaches has provided the ability to identify the capacity of plants to produce additional terpenoids that may not have been detected in that specific plant species before. Rather than isolate, fractionate, and chemically identify a particular plant terpenoid profile, plant TPS or cytochromes P450 are cloned into a heterologous expression system, and the enzyme assay products of these recombinant enzymes with isoprenoid substrates are analyzed directly. This approach has proven useful in situations in which the terpenoids themselves may not be detectable in the plant under normal growing conditions, and the inducer or environmental conditions required for their production are not yet known (12).

Synthesis

Independent synthesis of the identified terpenoid is often required to confirm structural assignment (51) and to test the biological and/or pharmacological functions. In addition, biosynthetic studies often require the synthesis of putative precursors for functional characterization of enzymes. Terpenoid structures challenge chemists in many of the same ways that other natural products do. The stereochemistry, complex overall structure (polysaccharides, carboxylic skeletons, and often multiple functionalization) provides opportunities for synthetic organic chemists to develop new methodologies for synthesis. Recently, Maione and Baran (52) have reviewed some synthetic challenges in plants. Terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. M.G. Grelet. Genetics 2002 287:730-743.

Biochemical and metabolic engineering

Often, the original natural plant sources of medicinally or industrially important terpenoids cannot supply sufficient material for their demand. In some instances, such as Taxus for the production of Taxol (53), the original plant species is amenable to growth in plant cell culture for commercial production. For example, Taxol (53), the original plant species is amenable to growth in plant cell culture for commercial production. For example, Taxol is produced by site-directed mutagenesis of amino acids that are important in influencing product outcomes. These studies not only allow us to understand how the diversity of terpenoids evolved from gene duplication and neofunctionalization, but also allow us to modify multi-product TPS to favor one of the wild type products or to force the enzyme into producing unnatural products. Both techniques have use in the large-scale production of medicinally or industrially important terpenoids.

References

Terpenoids in Plants


Terpenoids in Plants

Design and Engineer Natural Products, Methods to Isoprenoids
Natural Products: an Overview
Natural Products as Anticancer Agents
Natural Products Discovery, Molecular Biological Approaches to Terpenoid Biosynthesis
The Biology and Biochemistry of Steroid Hormones

S. Stoney Simons, Jr., Steroid Hormones Section, NIDDK/CEB, NIH, Bethesda, Maryland

doi: 10.1002/9780470048672.wecb563

Five categories of steroid hormones exist in humans, including androgens, estrogens, glucocorticoids, mineralocorticoids, and progestins. These hormones affect virtually every tissue and organ in the human body and play major roles in the development, differentiation, and homeostasis of normal individuals. Antisteroids usually possess nonsteroidal structures but still block the actions of the steroid hormones and are important tools in endocrine therapies of pathologic conditions. Therefore, how the body regulates where, when, and how much a response to steroids occurs is of major importance. Here we survey what is known about the genomic responses to steroid hormones, each of which is mediated by a unique intracellular receptor protein that interacts with the cellular DNA to modify the rates of gene transcription. These receptors are members of a much larger superfamily of steroid/nuclear receptors, most of which bind either nonsteroidal ligands or no known ligand. Nongenomic (i.e., pathways without initial involvement of genomic DNA) and secondary responses (i.e., changes that require protein synthesis to alter gene transcription) are additional important effects of steroid hormones but are not discussed here. The emphasis is on the biochemistry of the five classes of steroid hormones, the techniques used to study steroid hormone action, and the basic mechanistic steps by which steroids alter gene expression.

Steroid hormones can increase and decrease the level and/or activity of a large number of proteins in eukaryotes. Steroid hormones were first discovered in humans, where they play essential roles in development, differentiation, homeostasis, and endocrine therapies. However, current interest in steroid hormones is increasing because they constitute excellent model systems for examining the control of gene expression. Many human pathologies result from the inappropriate expression of protein(s). Thus, to treat disease states, it is critical to understand the normal processes governing how, when, and how much of the information encoded in the DNA of cells is transcribed to mRNA's and eventually into proteins, which perform most of the functions of cells. Steroid hormones provide excellent model systems with which to address these clinically relevant questions.

Biochemistry of Steroid Hormones

Five classes of steroid hormones are produced in humans: androgens, estrogens, glucocorticoids, mineralocorticoids, and progestins. The structures of several steroid hormones were elucidated in the 1930s. The common feature among each compound is the fused four-ring structure (Fig. 1a) or steroid ring structure that is the defining characteristic of all steroids. The first synthetic steroid, the estrogen equilenin, was prepared in 1939 (1). The numerous synthetic schemes for preparing steroid hormones will not be addressed here. The reader is directed to the article on the Chemistry of Steroid Hormones elsewhere in this reference for this material.
Steroid hormones bind to receptor proteins

The discovery that steroid hormones bind to unique proteins with high affinity and selectivity commenced with the demonstration by Jensen and Jacobson that [3H]estradiol bound tightly to species found only in the target tissues of estrogens (2). These binders were discovered to be intracellular proteins. In general, the protein that bound a given steroid with the highest affinity was called the cognate receptor for that steroid—i.e., androgen receptor (AR), estrogen receptor (ER), glucocorticoid receptor (GR), mineralocorticoid receptor (MR), and progesterone receptor (PR). More recently, these steroid receptors have been found to be members of a larger superfamily of about 150 members in over 30 subfamilies across a range of species (3). Humans contain 48 receptors. Each receptor protein shares a highly homologous domain of about 66 amino acids that contains two “zinc finger” motifs, each of which includes four cysteine residues (4). The homology between the rest of the receptor proteins is much less. Many members of this superfamily do not bind any known ligand and are thus called orphan receptors. Other members bind ligands that are not steroids, such as triiodothyronine (T3), retinoic acid, and vitamin D, which is not a steroid although its precursor is a steroid. The receptors for these later ligands are found almost exclusively in the nucleus of cells, in either the presence or the absence of ligand, and are therefore collectively called nuclear receptors. These receptors are discussed in the article entitled the Chemistry of Nuclear Receptors. This article will cover only the receptors of the five steroid hormones of Fig. 1b.

Structure of steroid hormones

Agonist steroids

The structures of the major steroid hormones in humans are shown in Fig. 1b. Each of these steroids elicits the maximal response from their cognate receptor and are called agonist steroids. The years after the structural elucidation of these steroids witnessed an explosion of synthetic activity aimed at making “better” steroids. These synthetic steroids (e.g., Fig. 1c) have largely supplanted the naturally occurring steroids both in clinical applications, which will not be covered here, and in research. The structures of the synthetic steroid hormones can be different from those of the natural steroids (e.g., DES...
The Biology and Biochemistry of Steroid Hormones

Synthetic Agonist Steroids

(c) R1881 (Androgen)

Diethylstilbestrol [DES] (Estrogen)

Dexamethasone (Glucocorticoid)

Deacylcortivazol [DAC] (Glucocorticoid)

Promegestone [R5020] (Progesterin)

Figure 1 (continued)

and DAC in Fig. 1c (5). These synthetic ligands are all active at lower concentrations than the natural steroids of Fig. 1b, and thus, they exhibit a greater sensitivity, or potency but have equal efficacy. More precisely, full agonists for a given receptor induce (or repress) gene expression to the same level. However, the concentration of these synthetic agonists required for half of the maximal response (=EC50) is lower than that of the natural steroids. The synthetic ligands are more stable to metabolic degradation, and thus, they are longer acting, and generally retain more activity when taken orally. They also possess greater specificity than the natural steroids so that there is less interaction with noncognate receptors (5).

Antisteroids

Not all compounds that bind to a given receptor give full agonist activity, and some give no activity. These agents also antagonize the actions of the above agonist steroids and are therefore called antagonists or antisteroids. The amount of residual agonist activity of an antisteroid, or the partial agonist activity, is expressed as percent of maximal activity of a full agonist steroid. Thus, a molecule with 15% partial agonist activity will be a more effective antisteroid than a different compound with 50% partial agonist activity. Some naturally occurring steroids are antisteroids for other receptors, such as progesterone, which is an antisteroid for glucocorticoid and mineralocorticoid receptors. Most antisteroids are synthetic and have very different structures (Fig. 1d). RU486 (RU38,486 or mifepristone) is of interest as it is an efficient antisteroid for PRs, GRs, and ARs (6) but not MRs (7) or ERs. Unfortunately, it is currently not possible to predict either what kind of activity a new steroid will have or even to which receptor it will bind. This issue will be discussed in greater detail below.

Current scope of steroid hormones chemistry

Initially, the study of steroid hormones involved relatively unmanipulatable biologic systems and was the purview of biologists. Over the years, new techniques in molecular biology such as cloning, protein identification, and protein purification have made it possible to alter selected macromolecules in cells and
The Biology and Biochemistry of Steroid Hormones

Antiestrogens

Figure 1 (continued)

Biologic Background of Steroid Hormones

In several instances, the steroid hormones were named for their most prevalent biologic activity. Thus, the steroids found to cause mobilization of glucose were called glucocorticoids. Mineralocorticoids were those steroids that regulated the balance of the mineral's sodium and potassium.

Source of steroids

All steroids are synthesized in humans from cholesterol. The adrenal gland can synthesize all endogenous steroid hormones and is the major source of glucocorticoids and mineralocorticoids. Most androgens are secreted from the testes, whereas most estrogens and progestins derive from the ovaries. In postmenopausal women and men, a variety of sources, such as adipose tissue, are the major source of estrogens, although in much lower amounts. The environment contributes many compounds, called xenobiotics and endocrine disrupters, that can compete with the actions of endogenous steroids with often incompletely documented effects, whereas plants are the source of phytoestrogens that are ingested both intentionally and unintentionally.

Transport of steroids

Almost all steroids reach the target organs through the circulatory system, which is the definition of an endocrine system. A target organ, or cell, for a steroid is one that contains the cognate receptors. Given the hydrophobic properties and relatively low solubility in water of most steroids, their concentration in the circulatory system is increased by their complexation with serum-binding proteins such as corticosteroid binding globulin (CBG). Although these proteins were initially thought to be passive carriers of steroid, evidence now exists that they play...
The Biology and Biochemistry of Steroid Hormones

a more active role in steroid hormone action (12). In general, steroids enter cells by passive diffusion, although multi-drug resistance transporters can facilitate the depletion of intracellular steroids (13).

Basic features of steroid hormone action

The general mechanism of receptor-mediated action of steroid hormones is very similar for each of the five classes. Briefly, the steroid enters the cell and binds to an intracellular protein. The resulting receptor-steroid complex is converted in a poorly understood step called activation to a form that binds to specific, biologically active DNA sequences (called hormone response elements, or HREs) of the nuclear chromatin. These HRE-bound receptor-steroid complexes recruit various cofactors and then interact with the transcription complex containing RNA polymerase II to modify the rates of transcription of a nearby DNA sequence coding for an expressed protein (Fig. 2). This alteration of transcription rate is typically fast (15–30 min) (14).

The mechanism of steroid receptor-mediated changes in transcription has occupied center stage over the last 40 years. Serious attention is only just beginning to be paid to the role of steroid receptors in other, often more rapid, processes that occur within minutes (15, 16), such as the estradiol-mediated activation of eNOS that proceeds in about 5 min (15). These types of responses are frequently called “nongenomic” to indicate that the steroid receptor does not interact with genomic DNA. Steroid receptors have been reported in mitochondria (17) and cell membranes, although it is not yet clear whether all of these receptors are the same as the intracellular steroid receptors (16, 18–22 vs. 23). Some of the membrane-bound receptors for steroids are G protein-coupled receptors (24–27). A recent report suggests that membrane-bound steroid receptors can interact with, and augment the transcriptional activity of, the intracellular receptors (24). Finally, steroids can bind to nonreceptor molecules such as enzymes and transport proteins (see above), which may have yet undiscovered consequences.

The “nongenomic” and “nonreceptor” responses to steroids will not be covered here. Instead, we will concentrate on the mechanisms by which the classic steroid hormones alter gene transcription via their intracellular cognate receptor protein. We will discuss only the primary effects of steroid hormones, which are those rapid (15–30 min) events that lead to changes in gene transcription without any requirement for protein synthesis. It should be remembered that at a sufficiently precise level, the mechanism of action of each class of steroid hormone is different from that of the others.

Methods for Study of Steroid Hormone Action

Most readers are probably familiar with the use of radioactive steroids, Scatchard plots, affinity labeling, sodium molybdate,

Figure 2: General steps in steroid hormone action and their assays. The basic model depicts steroid (S) binding to its receptor molecule (R) to form receptor-steroid complexes (RS), which attach to biologically active DNA binding sites (HRE) to eventually produce changes in the levels of specific proteins. Experimental techniques to follow R at various stages in this pathway are indicated at the first point that each method can detect a signal. Most methods can also be used to detect receptors at any step downstream of the one for which it is first used.
immunofluorescence, gel shift assays, and Western and Northern blots to study steroid receptors. Over the last approximately 10 years, molecular biology has introduced many new techniques that are directed toward detecting receptors at various stages in their mechanism of action (Fig. 2). Two hybrid assays were developed to detect interacting proteins (screening done in yeast) and then to characterize these interactions (usually performed in mammalian cells) (28, 29). Cell-free pulldown assays with one partially purified protein offer more definitive evidence for a direct interaction under cell-free conditions because many potential adapter proteins have been removed. This removal can be accomplished either by attaching an immunogenic or high affinity tag on one protein (30) or, in the case of steroid receptors, using the DNA binding properties of receptors to immobilize the receptor and associated proteins (31). Coimmunoprecipitation (co-P) assays demonstrate that two or more molecules from intact cells are present in the same complex. Mass spectral identification of copurified proteins (by binding or co-P) offer a more rapid method of identifying a large number of potential associated proteins, but time usually limits the number of subsequently examined proteins to those that seem to be particularly interesting (32). Fluorescent-tagged receptors are excellent for observing the real-time location of receptors, with FLIP and FRAP being methods of determining the rate of protein movement (33). Unfortunately, these methods are not yet sensitive enough to see receptors at a single copy of responsive genes. Chromatin immunoprecipitation (ChIP) assays reveal the presence and kinetics of molecules binding to small regions of responsive genes, like the HREs (34). The advantage of ChIP assays is their ability to interrogate endogenous genes, even if the resolution is low (∼500 bp). Chromatin conformation capture (CCC) assays detect two separated DNA sequences that are brought together because of the binding of one or more proteins (35). The use of transiently transfected siRNAs in tissue culture cells offers a much simpler method for blocking the expression of selected genes, or at least for reducing the level of translation from the corresponding mRNA, than for preparing gene knockouts in whole animals (36). Microarrays, in which the level of less of thousands of mRNAs can be determined, offer an unprecedented ability to determine the effects of steroid hormones on essentially the entire genome of a tissue (37). ChIP-on-chip assays use the microarray technology to determine to which regions of the genome receptors (and other proteins) are bound in ChIP assays (37). Systems biology approaches, and the associated model building, provide a powerful method of identifying testable mechanistic hypotheses for which no experimental evidence previously existed (38).

**Molecular Biology/Chemistry of Steroid Hormone Action**

**Structure of steroid receptors**

All members of the steroid/nuclear receptor superfamily have the same overall structural organization (Fig. 3) (39). The 66 amino acid DNA binding domain (DBD) is in the middle and is the criterion for superfamily membership because of the high amount of conserved sequence (4). The amino-terminal domain containing the activation function 1 (AF1) displays the greatest variability in length (∼200 to 600 amino acids) and possesses <15% homology between receptors. A small hinge-region separates the DBD from the multifunctional ligand binding domain (LBD) (40), which includes the second activation function (AF2) and is 245 residues long for GR (41). High amounts of homology exist between the LBDs of several steroid receptors, which accounts for the fact that ligand binding is rarely totally specific (5). Given the large number of biologic activities that are mediated by the LBD (see below and Reference 40), it is not surprising that most mutations are loss of function or neutral. Gain of function mutations that augment receptor properties are rare (42, 43), although many more mutations alter the binding properties of receptors (44–46).

**Receptor isoforms and variants**

Variations in the translational start site, and in mRNA splicing, are two common ways in which receptor isoforms can be formed. Isoforms with identical LBDs and different amino termini occur because of different transcriptional or translational start sites, such as the long and short forms of PR (PR-B and PR-A, respectively), a variety of ERs (47), and multiple forms of GR (48). These isoforms bind ligand with relatively unchanged affinity but usually display different biologic activities. Splicing isoforms can have dramatically altered properties, such as GRβ, which no longer binds steroid. GRβ is found in humans (49) but not in mice (50) and contains a different and smaller sequence for the C-terminal 50 amino acids than the full-length GRα. The sole steroid receptor encoded by two different genes is ER. The classic ER is ERα. The closely related ERβ was discovered only in 1996 (51) and often has very different biologic activities (52). The properties of these interesting additional forms of the steroid receptors, including post-translationally modified receptors, are beyond the scope of this article and will not be covered.

**Intracellular localization of receptors**

In contrast to most receptors, most ligand-free steroid receptors are not membrane bound. They can be cytoplasmic or nuclear, depending on the receptor and the time scale. At a given instant, ERs are mostly nuclear, like the nuclear receptors. For the other receptors, the amount of nuclear localization is PR > MR > AR and GR. However, the dynamic picture is that receptors are “shuttling” back and forth across the nuclear membrane (53).
Proteins associated with unactivated receptors

All steroid-free receptors are associated with other proteins regardless of their whole cell localization. Detailed experiments with PRs and GRs have uncovered the workings of a complex of five proteins (hsps90, hsp70, Hop, hsp40, and p23), often called chaperone proteins, which participate in the assembly of newly synthesized receptors to the form capable of binding steroid (reviewed in Reference 55). Although hsp90 is required for the "maturation" of receptors, it does not seem to be required for the de novo folding of proteins (56). This "maturation" consists of both promoting cleft opening (to allow high affinity steroid binding) and to limit excessive cleft opening (to prevent receptors from being targeted for degradation) (57). The final assembled complex retains hsp90 and p23 along with one of four tetratricopeptide repeat (TPR) proteins: three immunophilins (FKBP51, FKBP52, and Cyp40) or protein phosphatase 5 (PP5). This heterogeneity of ligand-free receptors theoretically could lead to a diversity of biologic responses but that has not yet been documented. Intriguingly, it seems that chaperone proteins such as p23 and hsp90 may also facilitate the disassembly of receptor-containing transcriptional complexes (58, 59). The nuclear receptor RXR is a ubiquitous heterodimerizing partner of other nuclear receptors. The only receptor LBDs exist (see below for DBD structures), and they seem that chaperone proteins such as p23 and hsp90 may also facilitate the disassembly of receptor-containing transcriptional complexes (58, 59). The nuclear receptor RXR is a ubiquitous heterodimerizing partner of other nuclear receptors. The only receptor was defined before it was possible to examine receptor binding to the promoter regions of endogenous genes. This definition may need revision because it is now clear

Steroid-induced changes in receptor structure

The first biochemical evidence that steroid binding modifies receptor structure was that protease digestion produced different sized fragments (61). The ability of this simple technique to discriminate between agonist and antagonist ligands has met with limited success (62, 63 vs. 64, 65). Hydrogen/deuterium exchange mass spectroscopy has been used to distinguish between agonist and antagonist binding (66). Nuclear magnetic resonance permits a more thorough examination of receptor structure, and especially protein motion, but the current methods of signal acquisition limit the size of molecules studied to ~50 kDa. However, new techniques may be able to increase the size limitations to include intact steroid receptors (67).

X-ray structures of receptors

X-ray crystallography provides the most complete structural data of proteins. Unfortunately, no X-ray structure of an intact receptor is yet available. Numerous structure determinations of receptor LBDs exist (see below for DBD structures), and they all show the same basic feature of an "α-helical sandwich" composed of 12 α-helices and 2 β-sheets (Fig. 4) (68, 69), although the number of α-helices (10-13) and β-sheets (2-4) can vary across the entire family of steroid/nuclear receptors. Significant differences were noted between agonist- and antagonist-bound receptors because of induced-fit changes in protein structure, but they are not yet predictable (70) because small changes in ligand-binding position can yield much larger effects on receptor structure (71).

The first X-ray structures of ligand-free and ligand-bound receptors were not of the same receptor. Nevertheless, they suggested an attractive model in which the C-terminal helix (helix 12) was triggered (like a "mousetrap") to close over the ligand-binding pocket upon steroid binding (68). The generality of this model is unclear, though, because no repositioning of helix 12 was observed in ligand-free and ligand-bound forms of two nuclear receptors [PPARγ (72) and PXR (73)] and because residues C-terminal of helix 12 seem to be important for steroid binding to PRs (74), GRs (65), and ARs (75). An alternative path for ligand binding to and dissociation from thyroid receptors (a nuclear receptor) has been proposed to occur through an opening caused by a proline that creates a kink in helix 3 (76). Again, this may not be general as all steroid receptors are lacking the comparable proline residue (68). Thus, multiple binding mechanisms with attending conformational changes may exist.

Nuclear Binding of Receptors

Activation

Regardless of whether the initially formed receptor-steroid complex is cytoplasmic or nuclear, a still poorly understood process called activation (or transformation) converts the complex into a species with increased affinity for DNA and for nucleic. The term "activation" was defined before it was possible to examine receptor binding to the promoter regions of endogenous genes. This definition may need revision because it is now clear...
that ligand-free steroid receptors can bind to the HREs of regulated genes, albeit with little observable transcriptional activity (77, 78, 79). This finding should be contrasted with the promoter binding of ligand-free nuclear receptors, which usually decrease gene transcription. In all cases, however, agonist steroid binding initiates events that alter gene transcription. A activation of DNA binding is affected by heat, salt, dilution, ATP, RNAse, and high pH (80), and it is blocked by the salts of molybdate, vanadate, and tungstate (81). It is also accompanied by the loss and/or increased rate of dissociation of many associated non-receptor proteins (57). This step is not microscopically reversible, but unactivated GR and PR can be regenerated in an ATP-dependent, reticulocyte lysate system (reviewed in Reference 55).

Nuclear translocation and DNA binding

The only way into the nucleus is through the nuclear pores. The entry of all molecules larger than \( \approx 60 \text{kDa} \), including all steroid receptors, is mediated by two nuclear translocation sequences in the hinge region and LBD of the receptors and proceeds via a two-step process. The first step, binding to nuclear pores, is followed by active transport of the receptor through the pores (82). Once in the nucleus, the activated receptor–steroid complexes readily bind to naked DNA, both biologically active sequences (or HREs) and nonspecific DNA. Whether receptors bind to DNA as monomers or preformed dimers is still debated. Recent studies with progesterone receptors (83) and other DNA-binding proteins (84) suggest that preexisting dimers dissociate and bind to DNA cooperatively as monomers. Nonspecific DNA binding is thought to be important both for buffering the binding to HREs and for facilitating the search of the HREs within the cellular genome (85). Each receptor binds to an HRE either as a homodimer or as a heterodimer with a closely related receptor (e.g., PR-AR-P-R (86); ER\( \alpha \)-ER\( \beta \)(87); and GR\( \alpha \)/GR\( \beta \) heterodimers (88)). In contrast with the nuclear receptors, no functional heterodimer of RXR with a steroid receptor has been reported. The HRE has the features of an inverted palindrome of six nucleotides on each side of a 3-nucleotide spacer. Considerable variation in HRE sequence is tolerated by GRs (89) and PRs (90), and maximal gene induction can be achieved with suboptimal in vitro DNA binding sequences. Surprisingly, ARs, GRs, Mxs, and PRs, but not ERs, can all recognize many of the same HRE sequences. The HRE sequence recognized by ERs is much closer to that of the nuclear receptors (4). The binding of ARs, GRs, Mxs, and PRs to the same HRE would cause much of the specificity that is gained by having separate receptors for each steroid to be lost. Among the many mechanisms that may restore specificity are differential responses of cofactor association with receptors (90) and effects of flanking and spacer DNA (91, 92).

All HREs have the properties of an enhancer in that their activity to induce gene transcription is position and orientation independent. HREs are commonly found within the 2-kb upstream of the promoter of regulated genes but can be much further upstream (93, 94) and even downstream (95) of the start of transcription. Those genes that are repressed by steroid receptors usually do not contain the same HRE sequences (96) because the receptors are often bound to another protein that directly contacts other DNA sequences (97, 98) but see (99).

Gene repression can also occur by preventing other factors from binding to their regulatory sites in responsive genes (100, 101).

X-ray structures of DNA-bound receptors

Almost all X-ray structures to date are of the DBD complexed with DNA. They all display several common features, including a dimerization of the DNA-bound DBDs, which is consistent with the highly conserved amino acid sequence of the DBDs, including two "zinc fingers," each of which contains four cysteine residues that complex one Zn\(^{2+}\) ion (Fig. 5). The right-hand "knuckle" of the first zinc finger, or p-box, contains part of the recognition helix for DNA binding and contacts the major groove of the HRE double helical sequence. The recognition helix is dominated by three amino acids of the p-box: GS--V for AR, GR, MR, and PR and EG--A for ER. The identical sequence in AR, GR, MR, and PR explains why they all can bind to the same HREs. The sequence in ER is very similar to the EG--G of the nuclear receptors. Interestingly, despite the critical role of these residues in determining the specificity of DNA binding, only one amino acid contacts the DNA of the HRE (4, 102). The first X-ray structure of a steroid receptor DBD, including the C-terminal extension (CTE) of the DBD, is that for PR. This structure shows that, like the nuclear receptors (4), additional contacts are made by the CTE in the minor groove flanking the PRE sequence, thus extending the size of the PR binding site (92).

Receptor binding to chromatin

Most DNA in cells is not present as naked DNA, waiting for receptors to bind. Instead, a large assortment of proteins decorates the DNA to give chromosomal DNA. The most abundant proteins are the histones, which form nucleosomes and obscure many of the intrinsic binding sites of proteins. In this manner, most regions of the genome are rendered transcriptionally inactive. In many cases, chromatin architecture greatly reduces the basal level of gene expression and steroids increase the efficiency of transcription by altering chromatin structure. In other cases, often where the basal transcription levels are much higher, there is less need for chromatin remodeling, and the fold-increases in gene induction are correspondingly less (103). In several GR-regulated genes, the nucleosomes in the region of the HREs are phased so that the HRE sequence is facing away from the center of the nucleosome and is accessible to activated receptor–steroid complexes (104).

After a receptor binds to the HRE, a host of other factors are thought to be recruited (78). One group of proteins is involved in chromatin remodeling. These proteins fall into two subgroups: species like pCAF and CARM1, which have kinase, acetylase, or methylase activity and covalently modify histones (reviewed in Reference 105), and ATP-dependent complexes, such as SWI/SNF and NURD (reviewed in References 106 and 107; see also the article on Chromatin Remodeling). Elegant real-time, whole-cell fluorescence studies indicate that GR binding to chromosomal GREs is much more rapid; transient, and readily exchangeable than previously suspected (33). These results suggest that GR presence is not needed to perpetuate the initial activating signal. This process could occur by cofactors (e.g., CBP/p300 and CARM1) covalently modifying histories...
to facilitate transcription initiation, with subsequent rounds occurring in the absence of the initially bound receptor-steroid complex. A iteratively, the entire cycle of receptor binding and regulated gene transcription could be rapid (108).

Control of transcription—induction vs. repression

Regardless of whether the receptor is mediating gene induction or repression, the same two domains are involved. These domains, called activation function (AF) 1 and 2, are located in the N- and C-terminal regions of the receptor, respectively. The N-terminal AF1 domain is usually the most active and displays intrinsic transcriptional activity independent of the rest of the receptor protein. No canonical activation sequences, such as acidic blobs, glutamine-rich regions, and amphipathic helices (109), have been identified in the steroid receptors. In fact, the AF1 domain, and the N-terminal half of steroid receptors, is generally unstructured but is induced to fold into a more regular structure both when exposed to helix-stabilizing environments (110, 111) and upon interacting with other transcription cofactors (112). The AF2 domain commonly has much less intrinsic activity and is comprised of helices 3, 4, 5, and 12. These helices form a hydrophobic pocket that binds coactivators, such as the p160 coactivators SRC-1, TIF2/GRIP1, and AIB1/pCIP/ACTR/RAC3/TRAM1 (Fig. 4), and corepressors, such as NCoR and SMRT. Coactivator binding is mediated by α-helical structures with the sequence LxxL, where L is leucine and x is any amino acid. Corepressor binding to an overlapping, but not identical, region of the receptor involves a related sequence of LxxL/IxxxxL/L (123). This sequence, and region of corepressors, was initially defined from interactions with nuclear receptors but has been confirmed for the steroid receptors (114, 115). However, not all LxxL or LxxL/IxxxxL/L sequences in coactivators or corepressors (or other molecules) are sufficient for binding to receptors as adjacent residues also make contributions (116, 117). Given the partial overlap of the binding sites, it is not surprising that the association of coactivators and corepressors to a given receptor-steroid complex displays competitive inhibition in a manner that is controlled, at least in part, by the ratio of coactivators to corepressors (114, 118, 119). These receptor-associated cofactors are then thought to recruit a burgeoning array of additional factors, although probably not at the same time (78, 120, 121; see also the articles on Transcriptional Control and on Activators and Repressors of Transcription). The mechanism of action of most of these factors remains poorly understood.

Steroid hormones both increase and decrease the levels of gene expression to give induction and repression, respectively. However, it is not yet clear whether all HRE-bound receptor-agonist complexes are transcriptionally active (77). A major unanswered question is how superficially similar steps can cause opposite responses. As described, the HREs for gene induction and repression are different. Repression is often achieved via receptors that are tethered to a different DNA-bound protein (97, 98) but see (99) as opposed to the
receptor binding directly to DNA, as observed in induction. Tethering is not per se contraindicated for gene induction, as observed by the ability of GR sequences fused to the GAL4 DBD to bind via the GAL4-DBD to a GAL4 upstream activating sequence and still give gene induction that is augmented by added coactivators (31, 122). Also, the coactivators TIF2 and STAMP still augment the activity of GR-agonist complexes in both GR-mediated induction and repression (123). Hence, the ability of the same receptor-agonist steroid complex to cause increased or decreased gene expression implicates the importance of other processes such as DNA-induced conformational changes (124, 125) and the ability of nearby DNA-bound factors to influence the recruitment of and/or interaction with additional cofactors (93, 92, 99).

The opposite effect of agonist steroids in induction versus repression also causes some confusion regarding the role of coactivators and corepressors. In the absence of precise mechanistic data, we currently must rely on phenomenological descriptions. The original definition of a coactivator was a factor that increases the activity of an agonist steroid (126). Accordingly, a protein that is labeled a coactivator because it increases steroid-regulated induction would be expected to enhance steroid-mediated repression and afford less gene expression (123). Therefore, the apparently opposing effects of a coactivator during induction and repression simply reflect the opposite and currently unknown actions of the agonist steroid. Virtually all studies of gene induction involve averages from populations of cells, which led to the assumption that all cells respond equally to the same concentration of steroid. However, starting with the demonstration that two daughter cells can display drastically different levels of response to a common steroid concentration (127), it has become increasingly clear that transcription versus receptor binding to the gene was recently described in a real-time, single-cell study of the binding of green fluorescent, protein-labeled GRs to an integrated 200-copy tandem array of a reporter gene under the control of the mouse mammary tumor virus promoter (129).

Modulation of the parameters of receptor-mediated gene expression

Two distinctive parameters for gene induction (or repression) are the total activity (equivalent to V_{max} in enzyme kinetics) and the steroid concentration required for half-maximal activity, or EC_{50}. Most factors represented in the current model of steroid hormone action (e.g., Fig. 2) were identified on the basis of their ability to increase, or decrease, the total transcriptional activity (reviewed in References 113 and 130). The determinants of the EC_{50} have usually not been considered, partially because it has long been considered to be controlled predominantly by the affinity of steroid binding to receptor (131). However, this prediction that all genes regulated by a given steroid hormone will be induced (or repressed) with the same EC_{50}, in contrast, it is well known that EC_{50}s of multiple genes induced by the same receptor-steroid complex are not the same, even within the same cell (132, 133 and reviewed in Reference 134). Furthermore, the EC_{50} for gene repression is often much lower than for gene induction (reviewed in References 119 and 135). The circulating concentrations of steroids (e.g., cortisol ≈ 0.4 µM, estradiol ≈ 0.1 nM, or progesterone ≈ 5 nM) are in the region of half-maximal induction of many regulated genes. Consequently, genes with different EC_{50}s will display different extents of response to the single circulating concentration of each steroid that is present at any one time. In many instances, the levels of steroid change, such as for glucocorticoids during the normal 24-hr cycle and during stress or for estrogens and progestins during the female menstrual cycle and pregnancy. Under these conditions, the transcriptional levels of those genes with EC_{50}s near the average of the initial and final steroid concentration will be altered more than the levels of other genes.

Antiestrogens are widely used in endocrine therapies to block the action of endogenous steroids, such as androgen-dependent prostate cancer and estrogen-dependent breast cancer. A very important parameter under these conditions is the amount of residual agonist activity, or partial agonist activity, of the antagonist. The expectation from the general model of steroid action that the amount of partial agonist activity of an antagonist will be the same with all responsive genes has not been experimentally verified. Instead, as above with the EC_{50}, it was found that the amount of partial agonist activity usually varies with the gene (136, 137). Indeed, for reasons that are not understood, there seems to be an inverse correlation between these two parameters. Thus, for receptor-mediated induction of a given gene, the partial agonist activity of an antagonist invariably increases when the EC_{50} of an agonist decreases and vice versa (reviewed in References 119 and 135). It was then realized that these gene-specific differences in partial agonist activity were desirable and offered a theoretical means of blocking only those genes responsible for the undesired pathology while retaining nearly normal levels of other regulated genes. A prime example is the ability of the antiestrogen raloxifene to block estrogen actions in breast cancer but not in bone (138). In recognition of the importance of having the amount of partial agonist activity of a steroid vary in a gene- and cell-specific manner, the term “selective receptor modulator” (SRM) is increasingly used instead of antagonist (139, 140). Thus, antiestrogens are often referred to as SERMs (selective estrogen receptor modulators) and so forth.

Elucidating the mechanisms driving these changes in EC_{50}s for gene induction, and in partial agonist activity of antiestrogens or SRMs, may greatly expand the available therapeutic targets for treatment of a variety of human pathologies. Initial progress has been made with the findings that these changes can be reproduced by varying the concentration of steroid, receptor, coactivator, compresor, and other transcriptional cofactors such as CBP, pCAF, UbCB9, (reviewed in References 119 and 135) and a new protein STAMP (123). The physiologic relevance of these fluctuations is strengthened by the report that the levels of GR mRNA display circadian rhythms in several mouse tissues (141). The EC_{50} for glucocorticoid killing of thymocytes was lowered 10-fold in transgenic mice containing a 2-fold increase in GR gene dosage (142). Also, increasing the concentration of ER{alpha} 10-fold in human breast cancer cells increases the percentage of those genes that are induced by ER{alpha} from 22%
Much has been learned in the last 40 years since steroid hormone receptors were found to act via soluble intracellular proteins. This progress has been made under the guise of different labels, each of which is appropriately a discipline of chemical biology. First, better, longer lasting, and more specific steroids were prepared by chemists. Next, biochemists and biologists uncovered the basic steps by which steroid hormones affect the functioning of many cells and tissues of organisms. Biochemists and molecular endocrinologists discovered that the steroid receptors were a small subgroup of a much larger superfamily of related proteins that share many of the same features when inducing or repressing gene transcription. Molecular biologists have uncovered numerous additional factors that participate in steroid receptor regulation of gene transcription. Molecular physiologists are currently using powerful new techniques to determine the extent to which those mechanisms observed in isolated cell systems are employed in intact organisms for endogenous genes. System biologists are constructing mathematical models incorporating the burgeoning number of relevant factors in an effort to predict the effects of changing one or more components of the system. Eventually, clinicians will employ this wealth of information to correct selectively numerous pathologic conditions of the endocrine system. It is now clear that the specifics of steroid hormone action vary with the gene being regulated. The challenge ahead will be to gather sufficient gene-specific information to limit the effects of steroids to selected target genes and tissues, thereby increasing the desired therapeutic outcomes while reducing the number of unwanted responses.

Disassembly and deinduction: turning off steroid hormone action

Many steroid-regulated genes are only transiently induced or repressed, which is especially true for GR- and MR-responsive genes as the concentration of the activating hormone varies daily. Thus, cells need to be able to turn off steroid-controlled responses reasonably rapidly. Regrettably, very little research has been focused in this direction and even less is known. Interestingly, the chaperone proteins p23 and hsp90, which are required for the proper assembly of functional receptors (see above), may also promote the disassembly of the receptor transcriptional complex (58, 59 vs. 143). Over the last few years, evidence has emerged that the proteolysis of transcription factors and the 26S proteasome seem to be linked to active tran-
scription (77, 143, 144). More recent results, however, suggest that the role of the proteasome may be independent of proteol-
ysis (145). Furthermore, the proteasome inhibitor MG132 has been reported to both increase and decrease gene transcription (106). Additional work is clearly required to determine the gene-

Summary

Much has been learned in the last 40 years since steroid hor-
mones were found to act via soluble intracellular proteins. This progress has been made under the guise of different labels, each of which is appropriately a discipline of chemical biology. First, better, longer lasting, and more specific steroids were prepared by chemists. Next, biochemists and biologists uncovered the basic steps by which steroid hormones affect the functioning of many cells and tissues of organisms. Biochemists and molecular endocrinologists discovered that the steroid receptors were a small subgroup of a much larger superfamily of related proteins that share many of the same features when inducing or repressing gene transcription. Molecular biologists have uncovered numerous additional factors that participate in steroid receptor regulation of gene transcription. Molecular physiologists are currently using powerful new techniques to determine

Acknowledgments

This research was supported by the Intramural Research Pro-
gram of the NIH and NIDDK. The critical review of this article by John Funder (Prince Henry’s Institute of Medical Research, Australia), William Pratt (University of Michigan Medical School), Brad Thompson (University of Texas Medical Branch, Galveston), Douglas Forrest (NIH), members of the Steroid Hormones Section, NIDDK/NIH, and the referees is greatly appreciated. We regret that space constraints prevented mentioning many other excellent and informative studies.

References

5. O’ojar T, Dore J-C, Gilbert J, Raynaud J-P. Binding of steroid to the progesterin and glucocorticoid receptors analyzed by corre-
9. Gray LEJ, Wilson VS, Stoker T, Lambriith C, Furr J, Ron-
cussion 105–108.
10. Tabi M, Blumberg B. New modes of action for endocrine-


87. Hall JM, McDonnell DP. The estrogen receptor beta isoform (ER β b) of the human estrogen receptor modulates ER α-mediated transcriptional activity and is a key regulator of the cellular response to estrogens and antiestrogens. Endocrinology 1999;140:5566-5578.


Further Reading


The Biology and Biochemistry of Steroid Hormones


See Also

Steroids, Synthesis of
Nuclear Receptors, Chemistry of
Chromatin Remodeling
Transcriptional Control
Transcription, Activators and Repressors of
Catalytic Modes in Natural Ribozymes

Michael E. Harris, Center for RNA Molecular Biology, and Vernon E. Anderson, Department of Biochemistry, Case Western Reserve University School of Medicine, Cleveland, Ohio

doi: 10.1002/790407048672.wexc640

In nature, RNA transmits genetic information, and it catalyzes reactions that are universal in biology. The reactions of phosphorus and carbon esters that are catalyzed by natural ribozymes occur with specific geometries and are susceptible to both acid and base catalysis as well as catalysis that involves metal ions. High-resolution structures of a few ribozymes implicate the involvement of individual nucleotides in performing these catalytic strategies. Detailed biochemical tests of the mechanisms of these ribozymes provide evidence for the involvement of nucleobase functional groups in acid/base catalysis, yet debate continues concerning their precise roles in proton transfer. Complementary structural and biochemical support has been gained for the involvement of specific RNA functional groups in positioning metal ions for catalysis. Still, only a few systems have been probed in sufficient depth to infer mechanistic detail and to achieve a complete chemical description of how ribozyme active sites enact specific catalytic strategies remains a key goal.

RNA Catalysis Is Ubiquitous in Biology

RNA holds a unique position in living systems because it can both convey genetic information as well as act as an enzyme to accomplish catalysis (1–3). RNAs that promote catalysis are called ribozymes, and the examples discovered thus far in nature catalyze addition/displacement reactions that involve phosphoesters and carbon esters with rate enhancements approaching that of enzymes (4). Now appreciated as a fundamental aspect of terrestrial biology, the existence of ribozymes was a milestone recognized by the Nobel Prize in Chemistry to Cech and Altman in 1986 (5, 6). The discovery of RNA catalysis was important not only because it broadened our understanding of biologic catalysis, but also because it suggests a central role for RNA in the development of self-replicating systems that are believed to underlie the evolution of cells and ultimately all of biology (7).

The known ribozymes in biology may be divided into three different classes defined by differences in structural complexity as well as by the chemical reactions they catalyze (Fig. 1): Small ribozymes cleave an RNA phosphodiester backbone by catalyzing the intramolecular attack of a ribose 2'-OH on the adjacent 3'-5' phosphodiester, which generates characteristic 5'-hydroxyl and 2',3'-cyclic phosphate products. In biology, small ribozymes are involved in replication of RNA viruses and are found in the 3' untranslated regions of genes where they are involved in gene regulation (8–11). Large ribozymes also catalyze addition/displacement reactions of phosphodiester bonds, but activate oxygen nucleophiles that include 2' and 3' ribose hydroxyls, or water for intermolecular nucleophilic attack displacing the 3' O. A new 2'5' or 2'3' phosphodiester is generated or, if the nucleophile is in water (or rather its lyate ion hydroxide), the RNA chain is cleaved producing 5' monophosphate and 2', 3' cis-diol termini. Large ribozymes include two classes of self-splicing introns (termed Group I and Group II), which catalyze two successive, site-specific transesterification reactions, and P RNA, the catalytic subunit of the RNA processing endonuclease, ribonuclease P (12–14). The peptidyl transferase active site of ribosomal RNA, which is responsible for the synthesis of all proteins in the cell, comprises a third class (15). The RNA active site within the large ribosomal subunit RNA catalyzes attack of a primary amine on the carbon ester linkage between amino acid and 3'-OH of tRNA, which results in formation of a peptide bond and free 3'-OH.

By analogy with protein enzymes and precepts of transition state theory, ribozyme catalysis should involve an array of interactions between RNA functional groups and the reacting groups of the substrate that lower the free energy of the reaction and provide a driving force for chemical conversion. Detailed biochemical tests of the mechanisms of these ribozymes provide evidence for the involvement of nucleobase functional groups in acid/base catalysis, yet debate continues concerning their precise roles in proton transfer. Complementary structural and biochemical support has been gained for the involvement of specific RNA functional groups in positioning metal ions for catalysis. Still, only a few systems have been probed in sufficient depth to infer mechanistic detail and to achieve a complete chemical description of how ribozyme active sites enact specific catalytic strategies remains a key goal.
Catalytic Modes in Natural Ribozymes

Figure 1  General chemical mechanism of addition/displacement reactions at RNA phosphodiesters and amino acid esters catalyzed by ribozymes in biology. (a) Intramolecular hydrolysis of RNA via attack of an adjacent 2'-OH catalyzed by small ribozymes. The action of acid and base in leaving group protonation and nucleophile deprotonation are indicated in gray. Catalytic interactions with the nonbridging oxygens that result in protonation of a phosphorane intermediate or by stabilization of a negatively charged phosphorane-like transition state is also indicated. (b) Intermolecular phosphoryl transfer catalyzed by large ribozymes. In the reactions of large ribozymes, the nucleophile (R1) is a ribose 2'- or 3' hydroxyl or a water molecule. Interactions that result in acid/base or electrostatic catalysis are indicated as in part A. (c) Amidolysis of an amino acid ester that results in peptide bond formation catalyzed by the peptidyl transferase active site of the ribosome. The R1 and R2 groups represent the 2'-OH groups of the aminoacyl-tRNAs in the ribosomal A-site and P-site, respectively. Similar catalytic interactions involved in facilitating proton transfer and charge stabilization are indicated. Note that the net proton transfer from the nucleophile to the leaving group could be mediated potentially by a single functional group.

Reactions Catalyzed by Natural Ribozymes

Reactions of phosphodiesters

RNA phosphodiester bonds are cleaved under both basic and acidic conditions in reactions that involve intramolecular displacement of the 5'-G by attack of the adjacent 2'-O, giving 2',3'-cyclic phosphates (Fig. 1). Under acidic conditions, the 2',3'-cyclic phosphates can isomerize to form 2',5'-cyclic phosphates (See Reference 16 and references therein). A lack of nucleophiles on a tetrahedral phosphate enter gives a pentacoordinated species, the structure of which is a trigonal bipyramid that has two apical and three equatorial ligands. Nucleophiles may enter and leave the intermediate transition state only through apical positions, but the pentacoordinated phosphorane may be sufficiently stable to allow ligand reorganization by a pseudorotation. In acid-catalyzed hydrolysis, a phosphorane intermediate...
forms, which is sufficiently stable to pseudorotate while the base-catalyzed reaction is thought to be concerted. Rates of phosphodiester addition/displacement reactions in solution are sensitive to both nucleophile and leaving group pKₐ. Bronsted analyses support a concerted mechanism for the base-catalyzed reaction in which significant charge on both positions exists in the transition state. For attack of an adjacent ribose hydroxyl, a change in the degree of sensitivity occurs with leaving groups with pKₐ > 12, which supports a change in the mechanism that involves formation of a phosphorane intermediate as the leaving group becomes less reactive (17). Kinetic isotope effects on the nonbridging and leaving group oxygens, as well as solvent deuterium isotope effects, have been measured for the hydrolysis of phosphate diesters with nitrophenol and nitrobenzyl leaving groups (18, 19). The observation of measurable nucleophile and leaving group KIEs for base-catalyzed nitrophenol ester hydrolysis demonstrate a concerted mechanism and indicate equilibrium deprotonation of the nucleophile prior to nucleophilic attack. For the acid-catalyzed reaction of less reactive nitrobenzyl esters, leaving group KIE analysis are consistent with a pre-equilibrium proton transfer to the ester oxygen atom, followed by rate-limiting P–O bond fission (20). Thus, enzyme active site interactions will facilitate proton transfer to and from the leaving group and nucleophile and will influence their reactivity provide powerful catalytic strategies.

In addition to acids and bases, metal ions and their complexes also catalyze phosphoryl transfer reactions in solution, which provides a model for possible mechanisms in enzyme catalysis (e.g., References 21–24). Metal ion coordination lowers the pKₐ of the interacting alcohol, which reduces its nucleophilicity but increases correspondingly the concentration of the lyate ion at neutral pH. Because both phosphodiester reactions are sensitive to the pKₐ of the nucleophile and leaving group, metal coordination at these positions can accelerate the reaction. Sensitivities of phosphoryl transfer reactions to ionic strength indicate that significant electrostatic repulsion occurs between anionic nucleophiles and the negatively charged phosphoryl center. Thus, metal ion interactions that offset this repulsion or interact more favorably with the negatively charged transition state relative to the ground state can provide catalysis as well. Additionally, as outlined above, these reactions occur with specific ionic geometries for the nucleophile and leaving group. Because they can coordinate multiple electronegative ligands, divalent metal ions can also promote catalysis by simultaneous interactions with the nonbridging phosphate oxygens and the nucleophile to offset the free energy costs of decrease in entropy in the transition state. Such additional catalysis beyond effects on pKₐ and ionic strength have been referred to as induced intramolecularity (23).

Transpeptidation–aminolysis of a carbon ester

The transpeptidation reaction catalyzed by the ribosome involves nucleophilic attack of an amine on a carbon ester with subsequent displacement of the alcohol (Fig 3c). Isotope exchange from water into the substrate ester during aqueous ester hydrolysis provided evidence for exchange between the carbonyl oxygen and water, which implicates the formation of a tetrahedral intermediate (e.g., Reference 25). A minolyses of esters in solution can be general base- or general acid-catalyzed; however, a break in the pH-rate profile at lower pH exists, which suggests a change in the rate-limiting step. Extensive analysis of pH and substituent effects show that with most esters, nucleophilic attack to form the intermediate is rapid and reversible, and the rate-determining step at high pH is proton transfer within the intermediate or its breakdown. For very fast reactions, evidence exists for a change to rate-determining formation of the intermediate (e.g., References 26 and 27).

Nucleophilic attack on carbon esters also shows dependence on both nucleophile and leaving group pKₐ, and structure reactivity studies support a stepwise mechanism in which a tetrahedral intermediate forms and is more stable for esters with poor leaving groups like the 3′-ribose OH of aminkyl-RNA esters used by the ribosome (28). The transition state for formation of the intermediate is believed to be neutral with a zwitterionic character. The transfer of a proton from nucleophile to carbonyl oxygen or the leaving group oxygen is proposed to be the rate-determining step. Isotope effects on the nucleophile, esterified carbon, and ester oxygen provide strong support for a stepwise mechanism for attack of oxygen and nitrogen nucleophiles on nitrophenyl esters (29). Thus, like phosphodiester reactions, specific reaction geometry is dictated and facilitation of this geometry, as well as of proton transfer from the nucleophile to carbonyl oxygen, or the leaving group oxygen can provide catalysis.

The Potential for Catalysis by Ribozymes

The potential for RNA to act as a catalyst is dictated by its structure as a linear polymer of the four common ribonucleotides. Like DNA, RNA can form double stranded, antiparallel helices via traditional Watson-Crick base pairing. However, the backbone of nucleic acids is highly flexible and RNA can form complex tertiary structures that often involve non-Watson-Crick base pairing to create active site crevices for catalysis. The phosphodiester backbone is charged negatively and interacts electrostatically as well as direct coordination with solution divalent cations. Ribose, purines, and pyrimidine bases contain both H-bond donors and acceptors that help stabilize higher-order structures and provide for substrate positioning, as well as participate in active site interactions. The potential for general acid/base catalysis is constrained by the available functional groups that can participate in proton transfer. The pKₐs of RNA nucleobases are 3.5 and 4.2 for A and C, and 9.2 for G and U and the ribose 2′OH has a pKₐ of ~12 (Fig 2a). As outlined below, metal ions can be bound tightly by RNA and such hydrated Mg(2⁺) ions have pKₐ of ~10. Thus, at neutral pH, A and C would be good acids, but at a low concentration relative to the unprotonated form. Althoug G and U could serve as bases, they exist predominantly in their protonated forms. Conversely, A and C could act as bases
Catalytic Modes in Natural Ribozymes

Figure 2

Ionizable functional groups in RNA. The base and acid forms of the nucleobase and ribose pKa values that are nearest to neutral are shown. The pKa of a water that forms an inner sphere coordination interaction with Mg(2+) is also shown. Because they are predominantly unprotonated, but ther low pKa, they provide less driving force for proton transfer at neutral pH. Similarly, G and U have protons to donate, but have high pKa. However, large pKa shifts to increase the concentration of the active protonation state of amino acid functional groups often are observed in protein enzymes (30), and no a priori reason seems to exist why RNA would not be prone to similar pKa perturbations. Indeed, nucleobases with altered pKa have been observed in model RNAs (31), and coupling structure formation has been highlighted as a means to provide the driving force for pKa shifting in RNA (32, 33). Furthermore, the sensitivity to pKa will depend on the Bronsted value for that reaction, which can lessen the impact of nonoptimal pKa as on reaction rate (34). Similarly, general acid/base catalysis will always occur when a large change occurs in the pKa of the reacting group, and when the pKa of the catalyst is intermediate between the initial and the final pKa values of the substrate group (35). For example, the pKa of the 2'-OH nucleophile undergoes a very large change in pH a, from a value of 12 for the substrate 2'-OH to a value of 0 for protonation of the product ester. Thus, nucleobase functional groups with pKa as considered nonoptimal (pKa 3–4 or 9.2) are between these values and can nonetheless provide significant general acid/base catalysis.

Like protein enzymes, an understanding of the involvement of ionized functional groups can be gained by analyzing the effects of changes in pH on rate. Much of the detailed analysis of RNA catalysis has involved analysis of mechanistic detail from such dependencies. Typically, functional groups that are active in their protonated state (acids) as well as in their unprotonated states (bases) are invoked. However, the same pH-rate profile can be observed for different mechanisms (Fig. 3). For example, a reaction that is catalyzed by both acid and base displays a bell-shaped rate dependence on pH with two apparent pKa as for the acid and base. However, switching the pKa as of these species does not change the observed data. This kinetic ambiguity (36) has a very large impact on mechanistic interpretations of structure-function relationships in catalytic RNAs (34, 36). Indeed, the presence of multiple titratable groups and/or the presence of specific acid/base catalysis adds even more ambiguity with respect to identification of mechanisms that can develop observed pH-rate behavior. Nonetheless, kinetic data

Figure 3

Idealized pH dependence of a ribozyme reaction. Ideal pH-species plots and pH-kin plots profiles according to a kinetic model for general acid/base catalysis. The solid lines depict a mechanism in which the species with the lower pKa (pKa,1) acts as the general base (shown by blue lines), and the species with the higher pKa (pKa,2) acts as the general acid (shown by red lines). The black line indicates the observed pH dependence of the reaction rate. The dotted lines simulate a mechanism in which the catalytic role of the species with pKa,1 and pKa,2 have been switched. Adapted from References 34 and 35.
provide a direct functional read-out and necessarily must be accom-
modated by models of ribozyme mechanism. An ongoing challenge, as described below, is in defining specific experi-
mental tests to assign catalytic roles to individual active site inter-
actions.

Metal ions can promote phosphoryl and carbonyl reactions in 
solution in several ways and are a common feature of protein 
enzyme active sites. Like acid/base catalysis, information on the 
importance of metal ions to catalysis often has been gained by 
analyzing the dependence of activity on metal ion concentra-
tion and identity. However, the dependence of RNA catalysis on 
Mg(2+) concentrations can be complex. Importantly, metal 
ions make large contributions to RNA folding and promote 
activity because of stabilization of the active RNA conformation 
(37, 38). Thus, the observed concentration dependence of the 
ribozyme reaction may (or may not) represent a direct role in 
catalysis (Fig. 4). For example, metal ions may bind to the 
folded RNA conformation at the active site with high affinity. At concentrations of ions that are saturated for folding, 
the observed metal ion dependence will reflect ion binding to 
the active site. In contrast, if ion binding at a site or sites that 
stabilizes folding occurs at a lower affinity, then the observed 
dependence of the reaction on metal ion concentration will 
represent necessarily the binding to the structural, but 
nonetheless activating, site.

Additionally, RNA–metal ion interactions take two thermo-
dynamically important forms: 1) diffuse binding in which the 
ions interact electrostatically with the negatively charged phos-
phodiesters backbone and 2) site-specific binding in which metal 
hydrate makes inner sphere or direct H-bond contact with an 
RNA functional group (39). Although diffuse ion interactions 
may weaken contact, they are numerous and overall have a high 
thermodynamic contribution to folding and to apparent catalytic 
activity. Also, site-bound interactions are fundamentally impor-
tant, but may only occur at certain steps in the ribozyme mech-
anism, or be linked indirectly to catalysis by stabilization of 
structure. The difficulty in characterizing these interactions bio-
chemically is described picturesquely as equivalent to searching 
for them amid the “sea” of other RNA–metal ion interactions.

Accordingly, titration experiments are difficult to interpret in 
terms of specific metal ion binding sites, and development of 
structural and biochemical probes for functional metal ion in-
teractions has been a considerable focus of ribozyme research.

**Evidence for Acid/Base Catalysis in the Hepatitis Delta Virus 
Ribozyme Active Site**

The hepatitis delta virus (HDV) ribozyme is a member of the 
class of small ribozymes and functions as a self-cleaving RNA 
sequence critical to the replication of the virus’ RNA genome 
(1, 8, 40). HDV ribozymes are proposed to employ several 
catalytic strategies that include an important example of general 
acid/base catalysis that involves a specific cytosine residue in 
the active site. Indeed, a milestone in our understanding of 
RNA catalysis was the observation that HDV and other small 
ribozymes could function in the absence of divalent metal 
ion cofactors, provided that high (molar) concentrations of 
monovalent ions are present (41, 42). These high monovalent 
ion concentrations are believed to stabilize the active RNA 
conformation, which implies that the primary role of divalent 
metal ions is in structural stabilization (42).

An additional insight into HDV ribozyme active site inter-
actions came from the determination of the structure of the 
ribozyme including the 3′ cleavage product. The 5′-OH termi-
nus of this product is located in a cleft formed by the intersection 
of conserved helical elements. Within this cleft, the 5′-OH is 
sufficient proximity to form a hydrogen bond with the N3 
of C75 (C76 in the antigenomic version of the ribozyme). If 
C75 becomes protonated, it could act as a general acid to affect 
rate of cleavage (43) (Fig. 5). Participation of titrat-
able groups in the HDV reaction is indicated by a bell-shaped 
pH-rate profile that reveals two groups with apparent pKa val-
ues. ~6.5 and 9 (44). However, in the absence of divalent metal 
ions, the pH dependence reversed, with the rate of cleavage in-
creasing with lowering of pH (42, 45, 46), which indicates that 
a metal ion hydrate could perform as a general base. Although 
deletion of C75/C76, or mutation to G or U, reduces cleavage 
adynamic activity drastically (45, 47) a C76A mutation retains signif-
icant activity but displays an altered apparent pKa for the pH 
dependence of the self-cleavage reaction (44, 45). Functional 
group modification experiments with C analogs with altered 
N3 pK as demonstrated the importance of C75 ionization in the 
HDV cis-cleavage reaction. Inhibition of ribozyme function also 
occur at C75 with a C analog that lacks the N4 amino group, 
which implicates the exocyclic amine in critical interactions.

**Catalytic Modes in Natural Ribozymes**

**Evidence for involvement of a cytosine 
nucleobase in proton transfer**

The hepatitis delta virus (HDV) ribozyme is a member of the 
class of small ribozymes and functions as a self-cleaving RNA 
sequence critical to the replication of the virus’ RNA genome 
(1, 8, 40). HDV ribozymes are proposed to employ several 
catalytic strategies that include an important example of general 
adcid/base catalysis that involves a specific cytosine residue in 
the active site. Indeed, a milestone in our understanding of 
RNA catalysis was the observation that HDV and other small 
ribozymes could function in the absence of divalent metal 
ion cofactors, provided that high (molar) concentrations of 
monovalent ions are present (41, 42). These high monovalent 
ion concentrations are believed to stabilize the active RNA 
conformation, which implies that the primary role of divalent 
metal ions is in structural stabilization (42).

An additional insight into HDV ribozyme active site inter-
actions came from the determination of the structure of the 
ribozyme including the 3′ cleavage product. The 5′-OH termi-
nus of this product is located in a cleft formed by the intersection 
of conserved helical elements. Within this cleft, the 5′-OH is 
sufficient proximity to form a hydrogen bond with the N3 
of C75 (C76 in the antigenomic version of the ribozyme). If 
C75 becomes protonated, it could act as a general acid to affect 
rate of cleavage (43) (Fig. 5). Participation of titrat-
able groups in the HDV reaction is indicated by a bell-shaped 
pH-rate profile that reveals two groups with apparent pKa val-
ues. ~6.5 and 9 (44). However, in the absence of divalent metal 
ions, the pH dependence reversed, with the rate of cleavage in-
creasing with lowering of pH (42, 45, 46), which indicates that 
a metal ion hydrate could perform as a general base. Although 
deletion of C75/C76, or mutation to G or U, reduces cleavage 
adynamic activity drastically (45, 47) a C76A mutation retains signif-
icant activity but displays an altered apparent pKa for the pH 
dependence of the self-cleavage reaction (44, 45). Functional 
group modification experiments with C analogs with altered 
N3 pK as demonstrated the importance of C75 ionization in the 
HDV cis-cleavage reaction. Inhibition of ribozyme function also 
occur at C75 with a C analog that lacks the N4 amino group, 
which implicates the exocyclic amine in critical interactions.
Figure 5  Proposed acid/base catalytic interactions in the HDV ribozyme active site. The cleavage site for the HDV ribozyme is shown with the nucleotide 5′ to the site of bond cleavage shown in red and the nucleotide 3′ to that site shown in blue. Two proposed mechanisms for the function of the C75 cytosine nucleobase and a hydrated active site metal ion are shown in which C75 acts as either an acid (left) or base (right) as described in the text.

in the active site (48). These data develop a model of general acid/base catalysis, with C75 potentially acting as the acid as indicated by the ribozyme–product structure and a hydrated divalent metal ion as the general base.

Remarkably, addition of exogenous cytosine or certain nucleobase and imidazole analogs can rescue partially the activity of the nearly inactive C75U mutant ribozyme (44, 49). Structure-function studies of the rescuing analogs provided additional insight into mechanism. The pH-rate profiles for the rescued reactions were bell shaped with one pKa attributed to ionization of the exogenous base. When a second potential ionizable nucleobase (C41) was removed, one leg of the bell-shaped curve was eliminated (50). The purported ionizable base, C41, is distant from the active site but may enhance cleavage rates through structural stabilization. The rescue of cleavage activity in the C75 deletion mutant, and by analogy the native ribozyme, can be explained by a model in which three ionizable groups exist (8, 34). These include: 1) the 2′OH group nucleophile (or a general-base catalyst with a relatively high pKa that acts at the 2′OH group, 2) a C75 or an analogous rescuing buffer that acts as a proton donor, and 3) a nucleobase (C41) associated with structure formation.

Thus, the role of C75 seems to act as a general acid in its protonated form. The most likely candidate based on the product complex and on the chemistry of phosphodiesters would be protonation of the 5′OH leaving group. Recently, Das and Piccirilli (51) provided strong evidence for this model using a chemically modified "activated" RNA substrate in which the 5′OH leaving group is replaced by sulfur, which has a lower pKa and thus is a better leaving group. The activated 5′S substrate specifically suppressed the effects of C75 mutation and modifications that alter its pKa, consistent with it providing general acid catalysis, which mediates proton transfer to the leaving group through a protonated N3-imino nitrogen. However, this model for C75 function is distinct from that derived from crystallographic studies of precursor forms of the genomic ribozyme in which catalysis is blocked by the C75U mutation, chelation of divalent metal ions, or 2′-deoxy substitution of the 2′-OH nucleophile (52). Importantly, these structures reveal electron density for a divalent cation coordinated to the 5′-O, whereas the position of the N3 of U75 is equidistant from the 2′O and the 5′O. Comparison of the precursor and product structures reveals differences in the active site structure after cleavage. Interpretation of these differences in terms of the catalytic mechanism suggests that a conformational change occurs and results in dissociation of the catalytic metal ion. Thus, although strong evidence exists that the C75 nucleobase functions in acid/base catalysis, likely in its protonated form, the specific transition state interaction is still of considerable interest as is the interplay of catalysis and metal ion binding.

Other small ribozymes

The HDV ribozyme, however, is only one example of the class of small ribozymes that catalyze self-cleavage, which includes the hammerhead, hairpin, V5, and glmS riboswitch ribozymes (1, 2). Like the mechanistic detail described above, crystal structures have been solved for several modified versions of these ribozymes, and the structures thus derived provide candidate
nucleobases that could act as acid/base catalysis. Biochemical tests of these interactions in the transition state have been made in several cases to support the involvement of nucleobases in proton transfer. Nonetheless, the same limitations with respect to kinetic ambiguity limit the detail with which specific active site mechanisms can be articulated. Riboswitches are a newly described class of self-cleaving RNA in which the catalytic activity is linked directly to regulation of gene expression (53, 54). The glmS ribozyme is a self-cleaving RNA that occurs in the 5' untranslated region of the mRNAs of the gene that encodes glucosamine-6-phosphate (GlcN6P) synthase in some Gram-positive bacteria and also acts to regulate gene expression in an analogous fashion. Self-cleavage represents gene expression when this ribozyme is activated by binding of GlcN6P (55). Recent structural and biochemical studies result in a model for this RNA in which it binds its activator such that the amine group of GlcN6P is in position to interact with the reactive phosphate potentially to provide nucleophilic activation. Such coenzyme usage raises the potential for an expanded catalytic repertoire for ribozymes. Yet, obtaining a complete description of the transition state interactions for the glmS and other small ribozymes continues to present structural biologists, theoreticians, and experimentalists with significant challenges.

Analysis of Divalent Metal Ion Interactions in the Group I Self-Splicing Intron Active Site
Structural and biochemical evidence for catalytic metal ion interactions

Group I (GI) introns are large ribozymes that function as self-splicing introns that catalyze two successive phosphorothioate reactions using a single conserved active site (4, 12). In the first reaction, the 3'-O of a G cofactor attacks the phosphodiester bond at the 5' splicing site; the 3'-OH of the last exon nucleotide is the leaving group. In the second reaction, the 3'-O product from the first reaction attacks the 3' splice site that generates the spliced exons; the 3'-O of the last intron nucleotide (termed ωG) is the leaving group. Note that the 3'-O leaving group in the first step is the nucleophile for the second step, and the interactions with the leaving group in the second step will replicate interactions involved in nucleophilic activation in the first step (Fig. 6) (56). Recent crystal structures of a ribozyme intermediate formed after the first splicing step show that the substrate is in a highly constrained conformation with complete reversal of strand direction at the 3' splice site (57, 58). A network of hydrogen bonding interactions positions the 3'-OH of the 5' exon proximal to the scissile phosphate, and the geometry is that expected for intrinsic nucleophilic attack. With respect to active site interactions with the nucleophile/leaving group oxygens, most attention has focused on divalent metal ion interactions, and now strong evidence exists that nucleophilic activation at both 3'-O nucleophiles involves direct coordination to active site Mg(2+).

The GI ribozyme reaction depends on the presence of divalent metal ions, but as indicated above, the binding of these ions plays multiple roles that include folding and enhancing substrate binding affinities (59). The rate of the chemical step is Mg(2+) dependent, but these data do not distinguish between direct or indirect roles, or a combination of both. As indicated above, distinguishing active site metal ions from what has been referred to as the "sea" of other functionally important metal ion interactions presents a considerable challenge. For the GI ribozyme and other catalytic RNAs, site-specific evidence for active site metal interactions comes primarily from analyses of thiophilic metal ion rescue of phosphorothioate and other substrate modifications (e.g., References 60 and 61). These analyses rely on the fact that substitution of a substrate phosphate for the phosphorothioate weakens the affinity of coordinated Mg(2+) ions.

Figure 6 Proposed divalent metal ion coordination interactions involved in metal ion catalysis by the GI intron ribozyme active site. The coordination interactions determined by substrate Ph and 3'-amino modification and quantitative metal rescue are depicted on the left. The individual interacting functional groups are shown in red. The observed metal ions (green spheres) in the crystal structure of the GI ribozyme active site is shown on the right and the metal ligands identified by functional studies are shown as red spheres. Adapted from Reference 56.

Adapted from Reference 56.
substantially such that catalysis is inhibited by several orders of magnitude. Because softer metal ions such as Mn(2+) and Cd(2+) coordinate more readily to sulfur, if inclusion of these ions in the reaction rescues activity, then a direct coordination interaction at the substituted position is likely. By comparing quantitative analyses of the concentration dependence of the metal ion rescue for different substrate modifications, a network of different metal ion coordination reactions has been revealed. The model developed from a series of elegant biochemical studies of this kind combined with a wealth of additional structure-function and kinetic data can be summarized as follows: Metal ions coordinate to both the 3'-oxygen leaving group and to the 3'-oxygen on the G nucleophile in the first step (MA and MB, respectively), and a third metal ion interacts with the 2'-hydroxyl of the G nucleophile (MC). Two metal ions (MA and MC) also contact the pro-SP oxygen of the scissile phosphate. Extension of this analysis to encompass phosphate oxygens within the intron itself provide evidence that these metal ions are positioned by coordination interactions with nonbridging oxygens that constitute, in part, the active site of the ribozyme (12, 61, 62).

Recently, high-resolution structures of the ribozyme, which includes one that contains the 2'-OH of the substrate G, a key metal ion ligand, have been reported that provide the most consistent model to date of active site interactions that include metal ion coordination to the functional groups identified biocchemically. In the most complete structure, two metals coordinate to nonbridging oxygens in the intron and substrate that include all biochemically identified interactions (56, 63). One Mg(2+) ion makes inner sphere contacts with both the 2'-O and the 3'-O nucleophile leaving group of the G cofactor. One obvious difference in this model is that one metal ion makes the contacts attributed to metal ions MB and MC in the model derived biocchemically. A through differences in the electrostatic environment of the active site because of different functional group modifications may complicate interpretation of the biochemical analyses, the remarkably high degree of correspondence between the structural and functional analyses provide the most detailed account to date for ribozyme active site metal ion interactions.

Other large ribozymes

The presence of active site metal ion interactions seems to be a general feature of the class of large ribozymes that catalyze phosphoryl transfer. Both Group I intron and RNase P RNA ribozymes that catalyze transesterification and phosphodiester bond hydrolysis, respectively, are thought to use metal ion catalysis as well. Although ample evidence exists for higher order structure that provides specificity via positioning the reactive phosphate, high-resolution structure of other examples of this class are limiting currently. Information that concerns active site structures comes largely from kinetic analysis of substrate modifications including the kinds of PS rescue experiments described above. Both GI1 and P RNA are modeled to have two metal ions coordinated to the pro-RP oxygen of the reactive phosphate and to the nucleophile and leaving group oxygens. However, the number and the position of coordinated metals and their ribozyme contacts are yet to be described fully. Under reaction conditions that favor the chemical step, both ribozymes react with rates consistent with general base catalysis but whether these ribozymes, like their smaller counterparts, employ functional groups in proton transfer remains unknown.

Substrate Positioning and Substrate Assisted Catalysis in the Peptidyl Transferase Center of the Ribosome

A landmark achievement in our understanding of biology has been the elucidation of the structure of ribosomes and their complexes with substrates, cofactors, and inhibitors (15). In addition to providing a wealth of functional insights, the structures showed that the peptidyl transferase active site involves RNA functional groups exclusively (64). Comparisons with model reactions indicate that the ribosome active site provides ~100-fold catalysis that is promoted significantly by positioning the reactive groups for catalysis. However, mechanistic studies have been providing evidence both for and against the involvement of active site functional groups in additional catalytic interactions (65, 66). Indeed, the extent to which the ribosome employs catalytic interactions with the amine nucleophile and the oxygenation of the tetrahedral intermediate has framed several important studies that concern ribozyme mechanism (67).

Structures of the large subunit with low-molecular weight substrates and products bound are highly suggestive regarding peptidyl transferase active site interactions (64, 67-69). Three groups exist in the neighborhood of the reactive amino group that could conceivably interact to assist in positioning and deprotonation of the nucleophile. These include the 2'-OH of A76 at the 3' end of the P-site bound peptidyl tRNA itself, the N3, and the 2'-OH of A2451 (Escherichia coli) of the large subunit tRNA. Replacement of the 2'-OH group of A76 in the P-site bound tRNA with 2'-H or 2'-F resulted in a dramatic decrease in the rate of peptide bond formation but did not affect ground state substrate binding (70, 71). Combined with the structural perspective, a role in nucleophilic activation is proposed; however, the precise role of this substrate residue in assisting catalysis is not resolved entirely. Site-specific incorporation of nucleoside analogs into 23S RNA of thermus aquaticus 50 S subunits (72) provided evidence that the ribose 2'-OH of A2451 is also important for catalytic function. Also, mutation of A2451 reduces the rate of the chemical step of peptide bond formation by ~100-fold (73, 74).

Using puromycin as an A-site substrate, the rate of the peptidyl transferase reaction is strongly pH-dependent, and pH-rate profiles suggest two ionizing groups with apparent pKa values of 7.5 and 6.9 (75). The second pKa of 6.9 has been attributed to the amine nucleophile of the puromycin substrate. In the single-protonated state in which the ribosomal group with pKa 7.5 is protonated whereas that with pKa 6.9 is deprotonated, the reaction is 100-fold faster compared with the uncatalyzed reaction. Protonation of a ribosomal group with pKa 7.5 thus seems to contribute a factor of ~100 to the observed rate that is...
Catalytic Modes in Natural Ribozymes

Figure 7. General chemical scheme for peptidyl transfer by the ribosome. The scheme shows nucleophilic attack by the amine group of the amino acid (with side chain R2) esterified to the tRNA in the ribosomal A site (right) on the ester linkage of the aminoacyl tRNA (with amino acid chain R1). The 2′OH of the peptidyl tRNA participates in the reaction and seems to transfer a proton to the 3′O leaving group, either directly or potentially via a solvent bridge. Adapted from Reference 76.

likely to be caused by additional catalytic interactions. Kinetic analysis of a A2451U mutant showed that the pH dependence caused by the ribosomal ionizing group was eliminated, which is consistent with a direct role of A2451 in catalysis (75). Thus, the peptidyl transferase active site clearly makes contacts that position the aminoacyl and peptidyl tRNA termini in the appropriate geometry for nucleophilic attack. In addition to substrate-assisted catalysis by the adjacent tRNA 2′OH, the ribosome active site also provides a network of noncovalent interactions that assists in providing additional rate enhancement (Fig. 7) (76).

Conclusions

All enzymes are constrained necessarily by the intrinsic chemistry of the reactions that they catalyze and the availability of functional groups for interfacing with this chemistry. As transesterification of phosphodiesters and carbon esters involves proton transfer, acid/base catalysis is a fundamental strategy for which strong support now exists for ribozymes. Yet, a continuing challenge is “following the protons” to and from active site functional groups and the substrate. Similarly, it is accepted widely that metal ions are involved directly in phosphotransfer reactions catalyzed by large ribozymes, but the site and number of active site metal ion interactions and their precise contribution to rate enhancement is still the subject of intense interest. Confronting these challenges and developing new technical and intellectual tools for understanding RNA catalysis represents a significant focus of the current investigations of ribozyme chemistry.

Acknowledgments

We thank Drs. Eric Christian, Michael Been, and Scott Silverman for helpful discussions regarding to the concepts discussed in this article. We are grateful particularly to Drs. Scott Strobel, Donald Burke, and Phil Bevilacqua for review and insightful comments on the manuscript. We are indebted to the community of enzymologists who investigate RNA catalysis and whose foundational work is alluded to, but not referenced explicitly, because of limitations of space and scope of this article.

References

Catalytic Modes In Natural Ribozymes


Further Reading


Chemical Damage to Nucleic Acids

Errol C. Friedberg, Laboratory of Molecular Pathology, Department of Pathology and University of Texas Southwestern Medical Center, Dallas, Texas

Thomas J. Kodadek, Division of Translational Research, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, Texas

doi: 10.1002/9780470048672.wecb412

Nucleic acids are highly reactive macromolecules that can undergo a variety of chemical changes that may be deleterious to their function. In particular, deoxyribonucleic acid (DNA) undergoes multiple spontaneous chemical changes and is reactive with reactive oxygen species generated during oxidative metabolism and with naturally occurring chemicals, ultraviolet radiation, and a myriad of environmental chemicals. This review highlights many of these chemical changes.

Biologic Background

The fact that nucleic acids are highly reactive and are thus subject to frequent chemical change, i.e., DNA damage, necessitated the evolutionary selection of multiple cellular protective mechanisms in order to sustain life. These mechanisms include multiple ones designed to (1) dispose of free radicals, (2) restore the normal chemistry and nucleotide sequence of damaged DNA (DNA repair), and (3) tolerate persistent DNA damage that threatens the viability of cells as a result of arrested DNA replication or transcription, or because of high levels of mutation. This review is largely restricted to a consideration of chemical alterations of the nitrogenous bases, the informational elements of DNA. However, consideration is also given to the formation of strand breaks in DNA. The article does not consider interactions of nucleic acids with synthetic compounds, of which a very large number is known to react with both DNA and RNA. Nor do we consider the mispairings of bases that arise from the occasional errors associated with normal DNA replication. Finally, we do not address chemical alterations in RNA, since most of the RNA in cells turns over rapidly and the biological significance (if any) of RNA damage is apparently limited. The nature and mechanism of the numerous biologic responses that cells can mount to cope with the consequences of DNA damage is also beyond the scope of this limited review. We refer interested readers to a recent comprehensive textbook on these topics and to numerous references quoted therein (1).

Spontaneous Alterations in DNA Chemistry

Tautomeric shifts

All of the bases commonly found in DNA (adenine (A), cytosine (C), guanine (G), and thymine (T)) can undergo spontaneous, pH-dependent chemical changes called tautomeric shifts (2). The base T normally exists in the keto form (C=O) in the C-4 position, but it can occasionally exist in the rare enol form (C–OH) (Fig. 1). Under such circumstances it can anomalously pair with G in the usual keto state. Conversely, the C-6 position of G in the rare enol form can pair with T in the usual keto form (Fig. 1). The same is true of the N-6 position of A, usually in the amino form (NH2), switching to the imino form (NH) tautomer; in which case, it can mispair with cytosine in the amino form (Fig. 1). Reciprocally C can mispair with A when the C-4 position switches to the rare imino form (Fig. 1).
Chemical Damage to Nucleic Acids

Figure 1  Anomalous base pairing involving rare tautomeric forms of the bases in DNA. When either T or G are in the rare enol form, they can pair. Similarly, when A or C are in the rare imino form, they can pair. (Reproduced from Reference 1 with permission.)

Tautomerism in the bases commands interesting prominence in the history of the elaboration of the structure of DNA by James Watson and Francis Crick. When in the early 1950s Watson was attempting to identify structurally compatible base pairing schemes using simple two-dimensional cardboard models of the bases, he inadvertently deployed the rare tautomeric states until the chemist Jerry Donahue pointed out this error to him.

The tautomeric forms I had copied out of Davidson's book were, in Jerry's opinion, incorrectly assigned. My immediate retort that several other texts also pictured guanine and thymine in the enol form cut no ice with Jerry. Happily he let out that for years organic chemists had been arbitrarily favoring particular tautomeric forms over their alternatives on only the flimsiest of grounds (3).

Spontaneous Deamination of Nitrogenous Bases in DNA

The common DNA bases A, C, and G, as well as the less common base 5-methylC, are endowed with exocyclic amino groups. These bases can be spontaneously lost as a function of the ambient pH and/or temperature (deamination), thus altering their ability to correctly pair with other bases (4).

Spontaneous Deamination of Nitrogenous Bases in DNA

The common DNA bases A, C, and G, as well as the less common base 5-methylC, are endowed with exocyclic amino groups. These bases can be spontaneously lost as a function of the ambient pH and/or temperature (deamination), thus altering their ability to correctly pair with other bases (4).

Deamination of cytosine (C) to generate uracil (U) in DNA

Perhaps the most extensively studied example of spontaneous deamination, certainly in terms of its biologic outcome, is that of C to yield uracil (U) in DNA (Table 1). U, which is chemically identical to T except for the methyl group in the C-5 position of the latter, most often pairs with A, resulting in permanent G:C → A:T transition mutations during subsequent rounds of DNA
Table 1  Endogenous DNA lesions arising and repaired in a mammalian cell* in 24 h

<table>
<thead>
<tr>
<th>Endogenous source</th>
<th>100% double-stranded DNA</th>
<th>98% double-stranded DNA</th>
<th>2% single-stranded DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrolysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>depurination</td>
<td>9000</td>
<td>9000</td>
<td></td>
</tr>
<tr>
<td>depyrimidation</td>
<td>300</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>cytosine deamination</td>
<td>50</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>5-methylcytosine deamination</td>
<td>5</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Oxidation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-hydroxyguanine (8-oxoG)</td>
<td>500</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Ring-saturated pyrimidines (thymine glycol, cytosine hydrates)</td>
<td>1000</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>Lipid peroxidation products (M1G, etheno-A, etheno-C)</td>
<td>500</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>Nonenzymatic methylation by 5-adenosylmethionine</td>
<td>7-methylguanine</td>
<td>3000</td>
<td>3000</td>
</tr>
<tr>
<td>3-methyladenine</td>
<td>600</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>1-methyladenine &amp; 3-methylcytosine</td>
<td>n.d.</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Nonenzymatic methylation by nitrosated polyamines &amp; peptides</td>
<td>O6-methylguanine</td>
<td>10 - 50</td>
<td>10 - 50</td>
</tr>
</tbody>
</table>

*Estimates are for a 3 × 109 bp genome. (Reproduced from Reference 1 with permission.)

replication (5). Additionally, the presence of U instead of T in the promoter regions of coding genes may affect the binding of regulatory proteins, thereby altering gene expression (1).

The chemistry of C deamination has been extensively studied (6, 7). It has been suggested that at neutral pH hydroxyl ions can directly attack the C-4 position of C (Fig. 2), especially in single-stranded DNA. An alternative explanation offered is an addition-elimination reaction during which dihydrocytosine is generated as an intermediate. This reaction may involve the formation of dihydrocytosine and dihydrouracil as intermediates (Fig. 2).

Cytosine in DNA is stable, with an estimated half-life of about 200 years (4). This value derives for a calculated rate constant for cytosine deamination in single-stranded DNA of ~2 × 10^-10/s based on the direct measurement of deamination in nucleotides and polynucleotides in vitro at various temperatures and pH (8). Evaluating the biologic significance of these values is problematic because they derive from in vitro measurements using single-stranded polynucleotides, whereas in living cells, most DNA is, of course, in the duplex configuration and the rate of spontaneous deamination of cytosine in duplex DNA in vitro is believed to be ~1% of that in single-stranded DNA (8). However, limited regions of single-stranded DNA are transiently generated in the course of various DNA metabolic transactions, in particular replication, transcription, and recombination. Additonally, considerable evidence exists that duplex DNA undergoes spontaneous localized denaturation, or "breathing," that could further promote this process (9). (On the other hand, single-stranded DNA can interact with specific DNA-binding proteins that may protect these regions of the genome from chemical change.)

The literature documents various situations in which C deamination is enhanced. These situations include the presence of C in cyclobutane pyrimidine dimers (a form of DNA damage caused by exposure to UV radiation; see later discussion) or in mispairings with other bases or with alkylated bases (1). Cytosine deamination is also promoted in the presence of nitrous acid, a reaction that although not considered in this review, has lent much to our understanding of possible chemical mechanisms of spontaneous deamination (1). While considering the presence of U in DNA, it is relevant to point out that this base can also arise in DNA by its misincorporation (instead of T) during normal DNA replication as a consequence of the existence of a small pool of dUTP (1). Not unexpectedly, perturbations that increase the pool size of dUTP relative to that of TTP promote this misincorporation (1). However, the biologic consequences of A:U "mispairs" in DNA are limited because such base pairs have the same coding potential as A:T pairs (1).

Essentially all life forms studied are endowed with enzymes that recognize and remove U from DNA. These enzymes, called uracil-DNA glycosylases (UDGAs), are historically noteworthy because their discovery identified a new class of DNA repair enzymes that collectively embrace many more substrates than uracil in DNA. The several DNA glycosylases that selectively remove uracil from DNA comprise a superfamil of proteins that are represented in all life forms (1).
Chemical Damage to Nucleic Acids

Figure 2  Plausible chemical mechanisms for the base-mediated deamination of cytosine. Top: Deamination resulting from 1,4-addition. Bottom: Deamination resulting from 1,2-addition. In both cases, only the cytosine ring is shown subsequent to hydroxide addition.

Deamination of 5-methylcytosine

Cytosine that is methylated at the 5 position (5-methylC) is naturally encountered in the genomes of some organisms, both prokaryotic and eukaryotic (10) (Table 1). It has been estimated that the rate of deamination of 5-methylC is significantly faster than that of C in single-stranded polynucleotides, and that it could account for as much as 10% of the spontaneous deamination events in the genome of mammalian cells under physiologic conditions (4). Since deamination of 5-methylC generates T in DNA, this spontaneous reaction is considered to be an important source of T:G mispairs that can lead to G:C → A:T transition mutations, especially in CpG-rich promoter regions of many genes, so-called CpG islands (1).

For many years it was considered that T:G mispairs generated by the deamination of 5-methylC would not be repaired by mismatch repair mechanisms because of the intuitive notion that conventional mismatch repair would not distinguish T in T:G base pairs from that in normal T:A base pairs, potentially leading to wholesale loss of TMP from DNA. However, the discovery of DNA glycosylases as a general class of base excision repair enzymes led to the detection of a DNA glycosylase that specifically recognizes T in T:G mispairs (1). Additionally, some bacteria are endowed with a specialized form of mismatch repair that distinguishes between T:G and T:A base pairs (1).

Deamination of adenine and guanine

The deamination of A and G in DNA to hypoxanthine and xanthine, respectively, transpires at rates about 50-fold less than that of C (9). The biologic significance of such deamination reactions is also overshadowed by data indicating that they transpire about 10,000 times less frequently than their complete loss from the genome during spontaneous depurination (see later discussion).

Spontaneous Loss of Bases from DNA (Depurination and Depyrimidination)

The spontaneous loss of bases from DNA after hydrolysis of the N-glycosyl bond linking them to the sugar-phosphate backbone is a frequent source of spontaneous DNA damage, with potentially important biologic consequences (1) (Table 1). It is estimated that the genome of the average mammalian cell sustains the loss of about 10,000 purines a day, generating many apurinic/apyrimidinic (AP) sites (9). AP sites are very unstable, often leading to hydrolysis of the sugar-phosphate backbone and the formation of strand breaks (1, 4, 9).

The kinetics of base loss has been extensively studied in vitro using polynucleotide substrates incubated at varying pH and temperatures. Such studies indicate that G is lost from DNA about 1.5 times more rapidly than A (11). Extrapolation of Arrhenius plots generated by direct measurements at varying temperatures yields a rate constant for the depurination of duplex DNA of $k = 3 \times 10^{-11}$/s at physiologic ionic strength and temperature (11). This result translates to the loss of $\sim 1$ purine/E. coli genome equivalent. However, thermophilic bacteria that thrive at high temperatures and often at very acidic pH may sustain the loss of several hundreds of purines each generation absent specific biologic mechanisms to offset this loss. As just mentioned, mammalian cells may sustain the loss of $\sim 10,000$ purines/day. The chemistry of depurination at acidic pH is believed to involve protonation of bases followed by elimination of the base in an S$_{N}$2-like process that probably involves a ribose-derived oxonium ion (Fig. 3). Pyrimidines are spontaneously lost from polynucleotides at about 1/20th the rate of the loss of purines, but this process may translate into the loss of hundreds of pyrimidines during each mammalian cell cycle (12).
Chemical Damage to Nucleic Acids

Figure 3  Plausible chemical mechanism for acid-catalyzed depurination and subsequent base-mediated strand cleavage, shown in the context of an A–T base pair. Only the adenine-containing segment of the molecule is shown subsequent to the initial reaction. The depurination reaction proceeds through an oxonium ion intermediate (top right) that is quenched by water. The resultant hemiacetal exists in equilibrium with the open-chain aldehyde, thus making the 2′-hydrogen relatively acidic and facilitating base-catalyzed elimination that leads to strand cleavage.

The deoxyribose residues that are generated at sites of base loss in DNA exist in equilibrium between the closed furanose form and the open aldehyde form (13) (Fig. 3). The 3′ phosphodiester bonds associated with the latter are labile and are readily hydrolyzed by β-elimination during which the pentose carbon to the aldehyde is activated at alkaline pH and at increased temperatures (13). The same reaction proceeds at a reduced rate at neutral pH, at physiologic pH and temperature, a site of base loss in DNA has an estimated average lifetime of ∼400 h (14).

Spontaneous Oxidative Damage to DNA

Products of oxygen breakdown, such as superoxide radicals, hydroxyl radicals (the most reactive oxygen radical known to chemistry) and hydrogen peroxide, collectively referred to as reactive oxygen species (ROS) (15, 16), are continuously generated as byproducts of mitochondrial respiration in all aerobic organisms, as well as from lipid peroxidation and from the metabolic activity of phagocytic cells. ROS can also be generated when cells are exposed to various exogenous agents, notably ionizing radiation and numerous chemicals (especially redox-cycling drugs) (1). ROS are highly reactive with DNA, as well as with proteins and lipids. Indeed, the so-called oxygen paradox points out that although oxygen is fundamental to aerobic life forms, ROS pose what is perhaps the most pervasive natural threat to the integrity of both the nuclear and the mitochondrial genomes of cells (15). Numerous disease states (in particular cancer) as well as the process of aging have been implicated as consequences of oxidative damage to DNA (1).

To cope with this fundamental problem, cells have evolved a myriad of defense mechanisms designed to both reduce the burden of free ROS and to repair damage generated in DNA by ROS (1). The former defense mechanisms primarily comprise multiple cellular antioxidants; antioxidant enzymes such as superoxide dismutase, glutathione peroxidase, and catalase (strategically located in or near the mitochondrial membrane), as well as naturally occurring antioxidants such as vitamins A, E, and C; glutathione and thioredoxin. Cellular oxidative stress, a state characterized by (but not limited to) extensive oxidative damage to DNA, results when levels of ROS exceed these various antioxidant defense mechanisms. The following discussion focuses on the chemistry and to a lesser extent the functional consequences of oxidative base damage to DNA. Details of the repair of such damage and of the regulation of various antioxidant enzymes are reviewed elsewhere (1).

Reactive oxygen species (ROS)

Molecular oxygen (dioxygen \(O_2\)) exists in a triplet ground state and hence is relatively unreactive with the singlet state of biologic molecules. But the unpaired orbitals of dioxygen can sequentially accommodate single electrons, yielding superoxide radicals \(\cdot O_2^-\), hydrogen peroxide \(H_2O_2\), hydroxyl radicals \(\cdot OH\), and singlet oxygen \(\cdot O_2\).
Chemical Damage to Nucleic Acids

I. OH, and water as shown in the following equation (15, 17):

\[ (e^-\cdot e^-) + (2H^+) \rightarrow H_2O_2 \rightarrow OH + H_2O \]

The parameters that most significantly influence the potential of ROS to interact with and damage DNA are their reactivity, their half-life, and their diffusibility in cells. Being highly reactive, OH can readily oxidize organic molecules. However, their extreme reactivity dictates a high probability that they will rapidly encounter a reactive cellular component other than DNA. Thus, they have limited diffusibility, and OH generated directly from leaky mitochondrial respiration is probably not a major source of DNA damage.

The generation of OH in the very close proximity required for its interaction with DNA is believed to transpire by the Fenton reaction (1), first described around the turn of the nineteenth century by Henry John Horstman Fenton, who discovered that hydrogen peroxide is a much more potent oxidant in the presence of various metals, notably the ferrous ion (FeI) than in their absence. Subsequently, Fritz Haber and coworkers proposed that the potent oxidant generated during the Fenton reaction is in fact OH, produced according to the following reaction (18):

\[ Fe^{2+} + H_2O_2 \rightarrow Fe^{3+} + OH + H_2O \]

Even trace concentrations of iron support a robust Fenton reaction if reducing agents such as NADH (17), ascorbate (19), or superoxide are available to recycle FeII to the active FeIII form. Although the Fenton reaction typically involves FeII, it is also supported in the presence of other transitional metals such as copper.

Like hydrogen peroxide, the free superoxide radical (O2-) is not highly reactive. However, this chemical moiety can readily generate H2O2, especially in the presence of the enzyme superoxide dismutase, by the reaction:

\[ 2O_2^- + 2H^+ \rightarrow O_2 + H_2O_2 \]

Superoxides can also reduce and liberate FeII from fenton, or liberate FeIII from iron-sulfur clusters (17, 20), thereby generating free iron that facilitates the formation of highly reactive oxygen species from H2O2 and O2- by way of the Fenton reaction.

Hydroxyl radicals can abstract electrons from residues of organic macromolecules (RH) by the following general reaction:

\[ RH + OH \rightarrow RH + H_2O \]

This reaction can initiate chain reactions that may generate DNA damage at considerable distances from the initial chemical event (21). Peroxidation of unsaturated lipids initiated by reactive free radicals such as OH represents a prominent example of such a chain reaction. The phospholipids of all membranes contain high concentrations of polyunsaturated fatty acids, and chain reactions involving hundreds of phospholipids can react in each oxidation event (22). The initial products of such fatty acid oxidation are lipid hydroperoxides, but these are relatively short-lived; they are either reduced by glutathione peroxidases to unreactive fatty acid alcohols or they react with metals to produce products such as epoxides and aldehydes, which are themselves reactive.

The major aldehyde products of lipid peroxidation are malondialdehyde and 4-hydroxyxynonenal (Table 1, Fig. 4). Malondialdehyde can react with DNA to generate adducts at the bases A, C, and G. The mutagenic adduct M1G (pyrimido[1,2-a]purine-10(9)H)one has been detected at levels as high as 1 adduct per 109 nucleotides in human tissues. M1G is a reactive electrophile that can undergo further modification, leading to cross-linking of an adducted DNA strand to the opposite strand, or to some protein (22). Exocyclic etheno adducts can also arise from lipid peroxidation, possibly by reaction of an epoxide of 4-hydroxyxynonanal with A, C, or G in DNA.

In addition to oxidants that are generated by the Fenton reaction, superoxide radicals (O2-) readily react with nitric oxide (NO), generating peroxynitrite anion (ONOO-) in the following reaction:

\[ NO + O_2^- \rightarrow ONOO^- \]

The protonated form of peroxynitrite anion, peroxynitrous acid, is highly reactive with biologic molecules. Hence, the production of nitric oxide from nitric oxide synthase (a complex enzyme containing several cofactors, and a heme group that is part of the catalytic site), which catalyzes the formation of NO from oxygen and arginine, can render cellular components such as DNA susceptible to superoxide-mediated damage (1).

DNA damage by ROS

Hydroxyl radicals can add to the DNA bases or abstract hydroxyl atoms, generating multiple types of DNA damage. The C4-C5 double bond of the pyrimidines C and T is particularly sensitive to attack by hydroxyl radicals, generating a large spectrum of oxidized pyrimidines, notably thymine glycol, urea residues, 5-OHdU, 5-OHdC, and hydantoin (Fig. 5).

Table 1. Products of DNA Oxidative Damage

<table>
<thead>
<tr>
<th>Product</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1G</td>
<td>Mutagenic adduct of lipid peroxidation</td>
</tr>
<tr>
<td>4-HNE</td>
<td>4-Hydroxyxynonenal</td>
</tr>
<tr>
<td>NO</td>
<td>Nitric oxide</td>
</tr>
<tr>
<td>ONOO^-</td>
<td>Peroxynitrite anion</td>
</tr>
</tbody>
</table>

Tables 1 and 2. Similarly, interaction of hydroxyl radicals with the purines A and G generates multiple purine oxidative products. The altered base 7,8-dihydro-8-oxo-guanine (23, 24), sometimes referred to as 8-oxoguanine (8-oxOG), is a biologically important form of base damage caused by oxygen free radicals (1) (Table 1). When present in DNA, this residue readily assumes a conformation that can base pair with A, resulting in transcription mutations after DNA replication (25).

Oxidative damage to the nitrogenous bases in DNA is readily generated during the handling of this biologic macromolecule under standard laboratory conditions. Despite the extensive use of antioxidants during the isolation and handling of DNA, this problem has seriously complicated accurate measurements of basal steady-state levels of oxidative damage in cells, especially with the advent of sensitive techniques for its measurement, and early estimates reported in the literature should not be relied on (1).

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Figure 4. (a) Malonaldehyde is a major product of lipid peroxidation that is reactive with A, G, and C in DNA. (b) 4-Hydroxy-2-nonenal, another major product of lipid peroxidation, can generate exocyclic etheno adducts of A, C, and G in DNA. (Reproduced from Reference 1 with permission.)

Treatment of DNA with hydrogen peroxide or other free radical-generating systems (such as xanthine/xanthine oxidase) can also result in the formation of an imidazole ring-opened derivative of guanine, designated 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FaPy) (26). This lesion is prominent among the forms of base damage induced by H$_2$O$_2$. Both this reaction and that for the analogous lesion in A proceed through a carbinolamine-type intermediate.

Close to 100 products of DNA base damage caused by ROS have been identified (27), usually by treating bases, nucleosides,
Chemical Damage to Nucleic Acids

Figure 5 Some of the many products of oxidation of the bases in DNA.

-or oligonucleotides in vitro with strong oxidants or ionizing radiation (27). It has been estimated that endogenous ROS may result in as many as 200,000 base lesions/day in the genome of the average mammalian cell (1).

Hydrogen abstraction typically affects the deoxyribose sugar units, leading to fragmentation of the sugars, base loss, and strand breaks with a terminal sugar residue fragment (28). Such breaks most often affect single polynucleotide chains of the DNA duplex. In contrast, ionizing radiation, which also generates ROS (1) typically produces clusters of hydroxyl radicals that can cause double-strand breaks. There is an extensive literature on ionizing radiation damage to DNA. This topic is not included in this review because absent unshielded travel in space, organisms are not typically exposed to ionizing radiation. Furthermore, as just noted, many types of DNA damage caused by ionizing radiation are identical to those generated during oxidative stress.

Anaerobic, or hypoxic, cells are resistant to the deleterious effects of ionizing radiation. Ionizing radiation, such as that used in radiotherapy, kills cells by producing DNA damage, particularly DNA double-strand breaks. This damage results from ionizations in the DNA or in water molecules very close to the DNA that produce a radical on the DNA (DNA•).

This radical then enters into a competition for oxidization primarily by oxygen (which fixes, or makes permanent, the damage), or reduction primarily by –SH-containing compounds that can restore the DNA to its original form. Thus, DNA damage, including single- and double-strand breaks, is less in anaerobic cells. This effect of oxygen in sensitizing cells to radiation is quantitated as the ratio of dose in the absence of oxygen to dose in the presence of oxygen needed to obtain the same surviving fraction of cells. For mammalian cells, this ratio is usually 2.5–3.0. The oxygen partial pressure that produces sensitivity midway between the oxic and the anaerobic responses is approximately 3 mm Hg. Clinical trials, particularly with head and neck cancers, have demonstrated that the more hypoxic tumors are more radioresistant than the less hypoxic tumors.

How cells cope with ROS

As mentioned, cells can survive the large amounts of ROS generated during normal cellular metabolism both by directly inactivating ROS by a variety of reactions, and by repairing oxidative damage to DNA. With regard to the former category
Chemical Damage to Nucleic Acids

Table 2  Radiation yields (nmol J⁻¹) of DNA base products under oxygen-free (N₂O) and oxygenated conditions (N₂O + O₂).

<table>
<thead>
<tr>
<th>Parent base</th>
<th>Product formed</th>
<th>Chromatin</th>
<th>N₂O</th>
<th>N₂O + O₂</th>
<th>Naked DNA</th>
<th>N₂O</th>
<th>N₂O + O₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5,6-dihydrothymine</td>
<td></td>
<td>0.26</td>
<td>n.i.</td>
<td>5.6</td>
<td>n.i.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-hydroxy-5,6-dihydrothymine</td>
<td></td>
<td>0.38</td>
<td>n.i.</td>
<td>7.7</td>
<td>&lt; 0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thymine glycol</td>
<td></td>
<td>0.094</td>
<td>0.40</td>
<td>10.2</td>
<td>43.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-hydroxymethyluracil</td>
<td></td>
<td>0.064</td>
<td>0.052</td>
<td>n.d.</td>
<td>n.d.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-hydroxy-5-methylhydantoin</td>
<td></td>
<td>0.042</td>
<td>n.d.</td>
<td>n.d.</td>
<td>n.d.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-hydroxy-5,6-dihydroxyasinosine</td>
<td></td>
<td>0.17</td>
<td>n.i.</td>
<td>2.6</td>
<td>&lt; 0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cytosine glycol</td>
<td></td>
<td>0.95</td>
<td>2.30</td>
<td>10.7</td>
<td>25.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5,6-dihydroxycytosine</td>
<td></td>
<td>b</td>
<td>0.33</td>
<td>n.d.</td>
<td>n.d.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-hydroxyhydantoin</td>
<td></td>
<td>b</td>
<td>0.22</td>
<td>n.d.</td>
<td>n.d.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-oxoA (7,8-dihydro-8-oxoadenine)</td>
<td></td>
<td>0.96</td>
<td>1.02</td>
<td>8.7</td>
<td>5.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-oxoG (7,8-dihydro-8-oxoguanine)</td>
<td></td>
<td>0.80</td>
<td>3.50</td>
<td>5.3</td>
<td>15.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-oxoG (7,8-dihydro-8-oxoguanine)</td>
<td></td>
<td>1.81</td>
<td>1.81</td>
<td>12.4</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Yields from the literature are for determinations by GC-MS for DNA isolated from chromatin and naked double-stranded DNA. Nil, not detected; b, not detected above levels present in control sample; n.d., lesion measurement not determined. (Reproduced from Reference 1 with permission.)

of cellular defenses, the following have been most extensively studied.

1. Superoxide dismutase (SOD) enzymes that eliminate superoxide by the reaction:

   \[ 2O_2^- + 2H^+ \rightarrow H_2O_2 + O_2 \]

   The several SODs in mammalian cells constitute a major form of protection against ROS. Since superoxide dismutase generates H₂O₂, its detoxifying effect primarily results from preventing the accumulation of free Fe²⁺ and the production of peroxynitrite.

2. Removal of hydrogen peroxide by catalase in the reaction:

   \[ 2H_2O_2 \rightarrow 2H_2O + O_2 \]

   In mammalian cells, catalase is largely contained in peroxisomes and does not seem to be nearly as biologically important as SOD (137).

3. The reduction of hydrogen peroxide by organic reductants (RH) such as glutathione, ascorbate, and cytochrome c:

   \[ H_2O_2 + 2RH \rightarrow 2H_2O + 2ROH \]

4. Yet another mode of defense against oxidant stress in mammalian cells is by transcriptional induction of heme oxygenase, which acts to generate heme metabolites with anti-oxidant properties (29).

Other Types of Spontaneous Damage to DNA

Spontaneous alkylation of DNA

Alkylating agents are electrophilic compounds that react efficiently with nucleophilic centers in organic macromolecules, including DNA (1). Alkylating agents can be highly toxic to cells and are widely used in the treatment of cancer primarily for this reason. Alkylating agents can be either monofunctional or bifunctional. The former have a single reactive group that reacts covalently with one of the many nucleophilic centers in DNA. Bifunctional agents have two reactive groups, so each molecule can potentially react with two sites in DNA, generating crosslinks. Numerous potential reaction sites for alkylation have been identified in all four nitrogenous bases, but not all of them have equal reactivity. The sites of reaction in DNA for many monofunctional alkylating agents include the following: in A, N1, N3, N6, and N7; in G, N1, N2, N3, N7, and N9; in C, N4, N6, and N7; and in T, N1, N3, N6, and N7.
0\textsuperscript{2}; in C, N3, N\textsuperscript{4}, and O\textsuperscript{2}; and in T, N3, O\textsuperscript{2}, and O\textsuperscript{4}. The ring nitrogens of the bases are in general more nucleophilic than the oxygens, with the N7 position of guanine and the N3 position of adenine being the most reactive (1). Alkylation of oxygen in phosphodiester linkages results in the formation of phosphotriesters (1).

Cells are endowed with several distinct enzyme-catalyzed mechanisms that specifically effect the repair of various alkylated bases (1). The alkylating agents used experimentally to generate the multiple substrates attacked by these enzymes are, in the main, not naturally occurring. However, it is intuitively compelling that these enzymes must have evolved to cope with spontaneous forms of alkylation damage to DNA (Table 1). Support for the notion of spontaneous alkylation damage to DNA comes from several studies.

1. Some microorganisms are endowed with a particular methyltransferase activity that catalyzes the methylation of halide ions (31), thereby naturally generating methyl chloride (MeCl\textsubscript{1}), a mutagenic alkylating agent that can react with various atoms in the nitrogenous bases. Large amounts of methyl chloride are also naturally generated during the burning of biomass (32), and reactive methyl radicals can arise by the irradiation or oxidation of methyl compounds such as methylhydrizine (33).

2. Some antibiotics are alkylating agents. A prominent example is streptozotocin, an antibiotic produced by the soil bacterium Streptomyces achronogenum. This compound is a 2-deoxy-1-glucose derivative of the potent alkylating agent methylnitrosourea (MNU), among naturally occurring chemical in this category, steroids are becoming increasingly recognized as a source of spontaneous DNA damage. Steroids are derived from cholesterol,

Spontaneous DNA–DNA and DNA–Protein Crosslinks

The covalent interaction of a chemical with bases on both strands of the DNA duplex can result in the formation of interstrand DNA crosslinks. Such lesions can interfere with the separation of the two polynucleotide chains that is required during various DNA metabolic transactions, especially replication and transcription. The biologic consequences of DNA crosslinking have been mainly investigated using synthetic bifunctional chemicals such as bifunctional alkylating agents (nitrogen and sulfur mustards are notable examples). However, several possible endogenous sources of interstrand crosslinks have been identified. For example, dietary nitrites can be converted to nitrous acid under the acidic conditions in the stomach. Nitrous acid generates interstrand crosslinks preferentially between the exocyclic N\textsuperscript{2} amino groups of guanine at CG sequences (36), and it has been estimated that a DNA interstrand crosslink may result for every four deaminations generated by nitrous acid (1). Aldehydes constitute another potential source of interstrand crosslinks in cells. Acetaldehyde, a product of normal cellular glycolysis, can cause DNA interstrand crosslinks (1). Malon-dialdehyde (another aldehyde) can be generated as products of lipid peroxidation.

DNA–protein crosslinks can constitute another source of spontaneous DNA damage, although one that has not been as extensively studied as the types of DNA damage already discussed. Nonetheless, aldehydes have been shown to cross-link histones and other DNA-bound proteins to polynucleotides. Cerulein and related aldehydes can also be endogenously generated by oxidative degradation of unsaturated lipids. Guanine adducts of acrolein, crotonaldehyde, and trans-4-hydroxynonenal can form crosslinks with peptides (37). This reaction is mediated by a Schiff base linkage between a ring-opened aldehyde moiety and the N-terminal \(\alpha\)-amine of the peptide (37).

Relatively unreactive chemicals can be metabolically activated to forms that are highly reactive with DNA.

A variety of relatively nonpolar compounds can undergo metabolic activation that is effected by specific metabolizing enzymes (38). The biologic function of these enzymes is to convert potentially toxic, lipid-soluble nonpolar chemicals to more reactive water-soluble and thus excretable forms, frequently esters (38). However, many of these more reactive water-soluble compounds can interact with nucleophilic centers in DNA (38), generating various types of base adducts. A series of membrane-bound proteins endowed with mono-oxygenase activities have been extensively characterized as components of these detoxification systems. These heme-containing proteins strongly absorb light at \(\sim 450\) nm and in combination with one or more membrane-bound flavoprotein reductases are often referred to as the cytochrome P-450 system (1).

A mong naturally occurring chemical in this category, steroids are becoming increasingly recognized as a source of spontaneous DNA damage. Steroids are derived from cholesterol,
which is converted to 17-hydroxyprogesterone by a series of specific oxidation-reduction reactions. Additional such reactions generate the adrenal cortical steroids, as well as the sex hormones estrogen, progesterone, and testosterone. Estrogens, including the natural hormones estradiol and estrone, can be hydroxylated at several positions by cytochrome P-450 enzymes. Hydroxylation at the 4 position to generate catechol estrogens is particularly important with regard to DNA damage. In cells, 4-hydroxylated catechol oxides are oxidized to semiquinones and then to DNA-reactive quinones. These quinone intermediates react at the 3- and 7-positions of the purine bases in DNA, creating unstable bulky adducts that are readily depurinated. The catechol estrogen metabolites have the potential to produce the full spectrum of oxidative DNA damage discussed above.

Although not widespread, aflatoxins, among the most potent liver carcinogens known, represent another example of DNA-damaging agents that have their origin as products of natural metabolism. Aflatoxins are toxins produced by the fungi Aspergillus flavus and Aspergillus parasiticus. Contamination of cereal grains such as rice, peanuts, and corn by these fungi can result in human and animal exposure (39). Chemically, the aflatoxins consist of a difurofuran ring system fused to a substituted coumarin moiety, with a methoxy group on the benzene ring (39). Among the aflatoxins of fungal origin, aflatoxin B1 is the most potent. Aflatoxin B1 can enter cells by passive diffusion and is metabolically converted by mixed function oxygenases of the cytochrome P-450 system, generating aflatoxin B1-8,9-epoxide as a major product (39). G residues in double-stranded DNA are preferentially attacked by this reactive electrophilic epoxide (1). The major product is an N7 adduct on G (8,9-dihydro-8-(N7-guanyl)-9-hydroxyaflatoxin B1) (40). The positively charged imidazole ring weakens the glycosyl bond, promoting depurination and subsequent sites of base loss. Mildly alkaline conditions can result in the formation of formamidopyrimidine derivatives containing an opened imidazole ring, which is a reaction that has also been identified in vivo (39).

**DNA Damage Caused by Natural Environmental Agents**

**Ultraviolet radiation from the Sun**

UV radiation from the Sun is a ubiquitous and important source of DNA damage that contributes to the high incidence of skin cancer in exposed individuals. The UV radiation spectrum has been arbitrarily divided into the UV-A component (wavelengths 320–400 nm), the UV-B component (wavelengths 295–320 nm), and the UV-C component (wavelengths 100–295 nm). Penetration of the atmospheric ozone layer is weak for wavelengths below 300 nm. Thus, solar UV radiation primarily comprises the UV-A and UV-B components. However, experimental studies typically involve the use of UV-C radiation from germicidal lamps with a sharp emission maximum at 254 nm. Since this wavelength is very close to the absorption maximum of DNA (260 nm) and is inefficiently absorbed by proteins, UV-C radiation-induced damage is relatively specific for DNA (1).

**Major photoproducts in DNA**

When cells are exposed to UV-C radiation, the quantitatively most abundant form of DNA base damage involves saturation of the pyrimidine 5,6-double bonds, resulting in covalent linkages between adjacent pyrimidines through a four-membered ring structure (1) (Fig. 6). This lesion is called the cyclobutane pyrimidine dimer (CPD). CPD can exist in cis-syn, cis-anti, trans-syn, or trans-anti configurations, but in double-stranded B-form DNA, they exist predominantly in the cis-syn form (1). However, trans-syn dimers are generated to some extent, mainly in denatured DNA (1), and can be detected in single-stranded regions and in duplex DNA with special conformations, such as the junctions between B-DNA and Z-DNA (1).

The thermodynamic and spectroscopic properties of short segments of duplex DNA containing a single CPD indicate that cis-syn CPD can be accommodated in the double helical structure of B-DNA such that hydrogen bonding with the opposite A residues is possible (41). The crystal structure of...
UV light is one in which the C6 position of a 5′-oligonucleotides (1). Although Watson-Crick base pairing can occur at the 3′ T in a CPD, base pairing of the 5′ T with A is severely weakened because one hydrogen bond cannot be made. CPD are extraordinarily stable to extremes of pH and temperature and to total acid hydrolysis of DNA. They can thus be readily resolved from thymine in such hydrolysates by a variety of chromatographic techniques, and quantitated (1).

The formation of CPD during irradiation of DNA is a reversible process that can be represented by the following reaction:

\[ \text{Py} + \text{Py (DNA)} \rightleftharpoons \text{Py}_{\text{add}} + \text{Py (DNA)} \]

This reaction reaches photochromic equilibrium when the thymine-containing CPD content (T \( \rightarrow \) T, T \( \rightarrow \) C, and C \( \rightarrow \) T dimers) reaches ~7% of the total thymine content of DNA (1). This steady state reflects a dynamic equilibrium in which the rate of CPD formation (which is pseudo-zero order, to good approximation) and that of CPD reversal (which is first order in dimer content) are equal (1). The yield of T \( \rightarrow \) T CPD is highest, whereas that of C \( \rightarrow \) C CPD is lowest (1). CPD are not randomly distributed in DNA. Numerous studies at the DNA sequence level have shown that their yields depend on DNA sequence context. In general, the equilibrium level of CPD is higher for TT sites flanked on both sides by A compared with such sites flanked on the 5′ side by A and on the 3′ side by G or C (1).

A another major form of damage generated in DNA by UV light is one in which the C6 position of a 5′ pyrimidine is covalently linked to the C4 position of the 3′ adjacent pyrimidine (Fig. 6). These lesions (which are readily detected by their lability in alkali conditions at 80-100°C (43)) are called pyrimidine-pyrimidine (6-4) adducts, or simply (6-4) photoproducts (6-4PP). The pyrimidine planes in (6-4PP) are almost perpendicular. Hence, they result in prominent distortions of the double-helical structure of DNA. (6-4PP) can involve adjacent TC, CC, or (less often) TT sequences. Their formation at CT sequences is infrequent. Cytosine methylation at the C5 position inhibits the formation of (6-4PP) (44). The yield of (6-4PP) is proportional to the incident UV fluence in the range 100-500 J/m² and continues to increase after exposure to several thousand J/m² (1). In UV-C-irradiated DNA, the ratio of CPD (6-4PP) is ~3:1 (1).

**Minor photoproducts in DNA**

Exposure of DNA to UV radiation also generates oxidative base damage. In particular, thymine glycol (5,6-dihydroxy-5,6-dihydrothymine) can be generated by saturation of the 5,6-double bond of some pyrimidines. Under anhydrous conditions, the so-called "spore photoproduct" or 5,6-dihydro-5-(a-thyminyl)-thymine, another type of dimeric pyrimidine photoproduct, can be generated by addition of a methyl group of a thymine to the C5 position of a neighboring thymine. The formation of this major photoproduct in UV-irradiated bacterial spores of B. subtilis (1) is dependent on the configuration of the DNA, being generated more readily in DNA in the compacted A-form (1). Consequently, the spore photoproduct is readily formed in dry DNA films and in DNA solutions containing high concentrations of ethanol (1).

Another type of chemical reaction involving a pyrimidine base is the addition of a molecule of water across the 5,6 double bond of C to yield a 5,6-dihydro-6-hydroxy derivative and yield the cytosine hydrate. The quantum yield for the formation of cytosine hydrates in UV-irradiated DNA is greater in single-stranded than in duplex-DNA (45). Hydrates of cytosine, deoxycytidine, CMP, or dCMP are unstable, readily reverting to the parent form by dehydration (45). However, their half-life is dramatically increased in DNA, and cytosine hydrate may be the major nondimer C photoproduct. Cytosine hydrate can undergo deamination and dehydration to yield uracil (1). The hydrate of 5-methylcytosine may undergo deamination to yield 5-thymine hydrate, which can convert to thymine upon dehydrogenation (1).

A photoproduct called 8,8-adeno dehydrodimer is observed after irradiation of poly(dA). This lesion involves a single 8,8 bond linking the imidazole rings (46). Another uncharacterized lesion involving adenosine has been reported after UV irradiation of simian virus 40 DNA (1). The lesion is alkali-labile and occurs at ACA sequences. Other photoninduced lesions involving purines have been identified (47). The fact that they are recognized by DNA repair enzymes suggests that they are physiologically significant. In general, the generation of purine photoproducts lesions is enhanced if they are flanked by two or more contiguous pyrimidines on their 5′ side (1).

**DNA crosslinks and strand breaks**

UV radiation can result in the cross-linking of DNA to proteins (DNA–DNA crosslinks) have also been observed, mainly when DNA is irradiated in the dry state, in an extremely densely packed condition (such as in the heads of salmon sperm), or in other special conformations (1). The spore photoproduct, formed in UV-irradiated dry DNA as described above, can arise not only between adjacent pyrimidines in the same chain but as a reaction between pyrimidines in different DNA chains (1).

Irradiation of DNA with very high doses of UV-C light can also result in breakage of the polynucleotide chain (1). However, the amount of UV radiation required to reduce the molecular weight of Streptococcus pneumoniae DNA by 50% is about 100 times that required to reduce the transforming activity of the streptococmycin resistance marker of that organism to the same extent (1). No chain breaks are detected in phage T7 DNA exposed to doses of UV radiation that inactivate almost 100% of the phage population (1). Thus, no conclusive evidence shows that the direct formation of DNA strand breaks by UV-C radiation is of biologic consequence. The frequency of strand breaks and DNA-protein crosslinks is dramatically increased by irradiation at longer wavelengths (1). In cells, most strand breaks observed after UV-irradiation are caused by biologic processing, either from interruption or breakdown of stalled
DNA replication forks, or as intermediates in the repair of photoproducts.

Conclusions

As discussed, much is known regarding the chemical reactivity of DNA and how these transformations lead to mutation. As discussed, much is known regarding the chemical reactivity of DNA and how these transformations lead to mutation. However, important challenges remain. For example, most of the estimates regarding the rates of DNA damaging processes are derived from simple model studies using oligonucleotides.

Do these rates reflect the reality of what occurs in the context of chromatin? These questions are challenging and new technologies will be required to address them. Although not covered in detail in this review, there is much to be done to increase our understanding of the enzymatic mechanisms employed by repair proteins to undo the damage brought about by everyday life. Thus, we anticipate that the field of DNA damage and repair will continue to fascinate both chemists and biologists for some time to come.

References


Further Reading


See Also

Chemical Views of Biology
Biomolecules Within the Cell
Chemistry of Biological Processes and Systems
Applications of Chemical Biology
DNA Transposable elements are segments of DNA capable of moving from one locus in a genome to another, or from one genome to another, using mechanisms that do not depend on large regions of sequence homology between the TE and the target DNA site. Their effects were observed first in plants in the middle of the twentieth century (1), and they were characterized at the molecular and mechanistic levels, largely in bacteria, in the last three decades (2). With the birth of whole genome sequencing, a growing recognition has occurred of the ubiquity of TEs, of their diversity, and of importance in shaping both eukaryote and prokaryote genomes and in influencing genome function. In some cases, their numbers are so high that it is tempting to consider that, together with bacteriophages, plasmids, and mobile introns, they form part of a genomic “ecosystem.” In the prokaryotes alone, nearly 2000 different insertion sequences (ISs; the simplest form of autonomous TEs) have been identified (IS-finder: www-is.biotoul.fr). TEs have been classified in various ways (see http://bioinformatics.org/wikiposon/doku.php). For example, in eukaryotes, a major division can be drawn between those elements that transpose via an RNA intermediate (Class I: retroviruses, retrotransposons) and those that transpose via a DNA intermediate (Class II: DNA transposons). Nevertheless, perhaps the most pertinent classification, for the purposes of this article, that deals exclusively with DNA transposons is based on the reaction mechanisms that they have adopted for their movement (3). These reactions involve cleavage of the DNA at the ends of the TE in the donor DNA molecule and transfer of these ends into a target DNA. They are catalyzed by a TE-encoded enzyme, the transposase. In the special case of retroviruses, the (DNA) provirus is liberated from its donor site not by DNA cleavages but by transcription to generate an RNA copy. During the viral lifecycle, this provirus is subsequently reverse-transcribed into a double-strand DNA genomic copy. This copy is then processed by an enzyme that resembles a transposase, the retroviral integrase or IN, which also assures subsequent integration into the host genome. Currently, the following five types of DNA transposase have been recognized: the DDE enzymes (most are identified so far) (3, 4); the tyrosine transposases (related to tyrosine site-specific recombinases of the phage λ Int family) (2); the S-transposases (related to serine site-specific recombinases of the γδ resolvase family) (2); the Y2-transposases (related to Rep proteins involved in rolling circle replication and to Relaxases, which are involved in conjugative gene transfer in bacteria) (5); and, finally, the Y1-transposases (also related to Rep and Relaxase proteins). The Y1-transposases were discovered only a few years ago (6–8), and it seems likely that other types of enzyme will be identified in the future. Each class of enzymes catalyzes a distinct chemistry that we describe below.

DDE Transposases and Integrases

The DDE enzymes are named for their characteristic triad of acidic amino acids (aspartate and glutamate). These enzymes bind in a sequence-specific way to the ends of the TE that generally carry terminal inverted repeat sequences and include the transposase recognition sequence. The simplest TEs, such as bacterial ISs, tend to carry only single transposase-binding sites at each end, whereas other more complex transposable elements...
may carry arrays of such binding sites. These arrays may be arranged in different patterns at each end and provide a means for distinguishing one end from the other. This arrangement could be involved, for example, in regulating transposon gene expression or in forming an asymmetric synaptic complex. It could lead to end-specific ordered cleavages or strand transfers that would determine the order of events that lead to integration.

Many DDE transposases carry a DNA sequence-specific binding domain in their N-terminal regions and at least one domain involved in multimerization. Generally, the catalytic domain is located in the C-terminal part of the protein. The DDE domain is by far the most studied and best understood binding domain in their N-terminal regions and at least one that would determine the order of events that lead to integration. The acidic pocket coordinates two Mg²⁺ ions jointly between DNA and RNA. Phosphodiester backbone by water. With the appropriate DNA substrate, the target phosphate can be shown to undergo stereochemical inversion in the course of the reaction, which implies a direct single-step, in-line nucleophilic attack (17). It is important to note that no covalent enzyme-DNA intermediate is formed during catalysis by DDE enzymes.

The acidic pocket coordinates two Mg²⁺ ions jointly between the nucleic acid substrate and the catalytic acidic residues of the enzyme (Fig. 1). This idea was based on the proposition of Staats (the two-ion model), which concerns the reaction chemistry of exonuclease activity of DNA polymerase (16). This coordination enhances nucleophilic attack of the nucleic acid phospho-sugar backbone and guarantees substrate recognition and catalytic specificity (16).
Despite their diversity, DDE transposases catalyze only two chemical reactions: cleavage and strand transfer (Fig. 1a and 1b). These reactions involve single-strand DNA cleavages. DDE enzymes do not catalyze double-strand cleavage. The cleavage reaction occurs at both ends of the TE, generally using H2O as the nucleophile, to generate 3′-OH ends (19, 20) (Fig. 1a). In the retroviral integrase (IN) proteins, this reaction is known as processing. For several transposons, this reaction is a major regulatory checkpoint because, in these cases, cleavage does not occur on isolated TE ends but requires prior formation of a transposase complex that involves both ends. In several systems, it has been shown that complex assembly evolves through several steps and becomes increasingly stable during this process. This evolution reflects the highly organized architecture of the complex, known as the transpososome (21).

A dynamic model of the catalytic reactions that lead to strand cleavage has been proposed based on structural considerations (22). A summary is thought to involve coordinate binding of transposase, its DNA substrate(s), and two divalent metal ions. On transposase DNA binding, the two metal ions (H and T for hydrolysis and transfer, respectively; Fig. 1a) find their appropriate positions in the active site and are poised for catalysis. Metal ion H orients and activates the water molecule (depicted as −OH) for nucleophilic attack. Metal ions T (which
is coordinated with an irregular geometry) and H then move closer to each other. The target phosphorus atom adopts a pentavalent transition state (Fig. 1a2), which is then converted into product as the metal ions move back and away from each other (presumably by charge repulsion between T and H). The resulting 5′-phosphate and 3′-OH then dissociate (Fig. 1a3).

In the case of nucleases, the reaction terminates at this step. For transposase, however, the liberated 3′-OH of the transposon end is then used as a nucleophile in a second reaction: trans-esterification. In this reaction, the 3′-OH attacks the target DNA, which results in strand transfer or joining of the transposon strand to its target strand. The detailed picture of the DDE cleavage reaction (Fig. 1a) might also be extended to the strand transfer reaction (Fig. 1b). The nucleophile of this reaction would now be the 3′-OH group from the preceding cleavage reaction, which would be coordinated by the T cation (Fig. 1b1). Therefore, the trans-estearification reaction is comparable to that of cleavage but, here, the in-line nucleophilic attack transfers the free 3′-OH end of the transposon in the target DNA (red in Fig. 1b), which creates a transposon-target joint.

Although all transposons with a DDE transposase (“DDE transposons”) use this type of chemistry, a large diversity exists in the overall transposition mechanism. As explained above, DDE transposases catalyze only single-strand cleavage and transfer of the 3′-OH transposon ends (the transferred strand). However, to liberate the transposable element from donor DNA, the transposase must deal with the second DNA strand (also called the nontransferred strand (3, 24). A subclassification of DDE transposases is based on the mechanisms used to manage this.

In the well-characterized bacterial IS4 family (which includes IS50 and IS10), it is the liberated 3′-OH of the transferred strand that is used to attack and to cleave the second strand. This generates a hairpin intermediate (Fig. 1c), in which transferred and nontransferred strands are joined, and both DNA strands that flank the ends are removed. Hairpin formation involves a considerable torsion of the DNA, which is aided by extrusion of a subterminal thymine residue from the DNA helix—a flipped-out T (25). This formation has been studied in detail for both IS10 and IS50 and requires a specific group of aromatic residues within the DDE domain. These residues are conserved within the family (26), and they serve to initiate and to stabilize this sequence-specific distortion (27). A second round of cleavage then removes the interstrand hairpin that regenerates the initial 3′-OH, which is used subsequently in the final strand transfer reaction.

This type of second strand management is found in eukaryotic transposable elements, such as the HAT group (28), and in V(D)J immunoglobulin-gene rearrangements (2) (Fig. 1d). In these cases, however, hairpin formation occurs on the equivalent of the transposon flanks rather than on the transposon itself. In the case of V(D)J, no specific subterminal T exists, and in the case of the HAT transposons, the hairpin is formed on the transposon flank and can vary between different HAT copies. Thus, hairpin formation in these cases occurs on sequences that are not necessarily conserved, and it seems unlikely that it involves a specific flipped-out T residue.

For the widely dispersed Tc-Mariner transposon group, the transposase first cleaves within the 5′ end of the transposon—the nontransferred strand. This activity resembles the nuclelease activity (see above) that simply terminates at the cleavage step and does not take in charge the strand transfer step. Moreover, unlike transposition reactions, this cleavage does not require formation of a synaptic complex (29). The transferred strand is cleaved at the very tip of the TE. The fact that the nontransferred strand is cleaved within the transposon results in retention of the few TE-specific bases in the donor molecule after TE excision.

After rescaling and repair, the donor backbone retains several additional base pairs derived from the TE (called a scar) that marks the passage of the transposon (2).

Interestingly, transposon Tn7 behaves in a similar way but, in this case, the 5′ endonuclease activity is supplied by a separate enzyme whose structure resembles that of a type II restriction enzyme (30), and cleavage occurs clearly at the transposon tip rather than within the TE. Transposition of Tn7, like most bacterial elements, does not leave a scar. In both the Tc-Mariner transposon group and the Tn7 family of transposons, the transposase then cleaves and transfers the 3′ end in a true DDE transposition reaction (Fig. 1e). Finally, cases exist in which the second strand is not processed at all. These include bacteriophage Mu and the Tn3, IS1, IS30, and IS256 families. Here, replication is involved intimately in the transposition process itself. In a first step, the transposase cleaves and transfers a transposon end to a target DNA. For phage Mu and Tn3, both transposon ends are transferred directly into target DNA that links both donor and target molecules at each of the transposon ends (2) whereas for IS2 (3), IS30 and IS256 family members, the 3′OH of one transposon end attacks the opposite end. In both cases, a branched structure is generated around which a replication fork is then assembled to resolve this structure (Fig. 1f). In the case of Mu and Tn3, strand transfer is followed by replication and leads to duplication of the transposon. In the case of IS1, IS30, and IS256 family members, replication leads directly to formation of a transposon circle intermediate and regenerates the original donor locus. The transposon circles, which carry abutted transposon ends, are highly recombinogenic transposition intermediates. They then undergo insertion readily into a suitable target DNA in a reaction that involves concomitant transposon-catalyzed cleavage and concerted transfer of both ends (31).

The Tyrosine and Serine Transposases

Although the reactions catalyzed by the 5- and Y-site-specific recombinases are well characterized, little is known about the biochemistry of their cousins, the Y- and S-transposases. These transposases are thought to catalyze strand breakage and transfer as do their site-specific recombinase relatives. Neither of these requires divalent metal ions for catalysis. Although these transposases show site-specificity for the ends of the donor transposon, they seem more flexible than classic site-specific
recombinases in the DNA target sequences they recognize and use as the partner during integration. Strand cleavage catalyzed by the Y- and S-recombinase enzymes occurs using a tyrosine or serine hydroxyl group, respectively. This generates a phospho-
roline or phosphoserine intermediate. The Y-recombinases form a 3'-phosphoryl residue bond, whereas the S-recombinases create a 5'-phosphosphate bond. Various host accessory proteins are required in these reactions to generate protein-DNA complexes with the correct architecture for catalysis.

The Y2-Rep and Relaxase proteins are less well understood, and the Y2 IS151 transposase is proving to be difficult to work with. Again, it is assumed largely that the Y2 enzyme behaves the same as the related plasmid and phage Rep and plasmid Relaxase proteins. The mechanism of the final group, the Y1 transposases, is now relatively well understood.

S-recombinases

Cleavage and strand transfer are mediated by successive tran-
esterification reactions within a synaptic complex that includes the two partner DNA sequences (called core sequences) bound by a recombinase tetramer. In the transposases, the DNA se-
quences would be the transposon ends. The nucleophile hydro-

droxyl of the catalytic serine residues initiate recombina-

tion by attacking specific phosphodiester bonds (the scissile phosphates) of the DNA backbone (Fig. 2a). Cleavage re-

sults in covalent attachment of the proteins to the 5' ends of the cleaved DNA strands and production of 3'-OH free ends. Cleavage (Fig. 2a) and strand transfer (Fig. 2b) is mediated by successive trans-
esterification reactions within a synaptic complex that includes the two partner DNA sequences (called core sequences) bound by a recombinase tetramer. In the transposases, the DNA sequences would be the transposon ends. The nucleophile hydroxyl of the catalytic serine residues initiate recombination by attacking specific phosphodiester bonds (the scissile phosphates) of the DNA backbone (Fig. 2a). Cleavage results in covalent attachment of the proteins to the 5' ends of the cleaved DNA strands and production of 3'-OH free ends. The first cleavage and/or strand exchange allows the reaction to proceed to the second pair of strand exchanges that occurs at the other end of the central region of the core site (Fig. 2b). Exchange of the second pair of strands is separated temporally and spatially from the first. This exchange, together with the inverted polarity of cleavage, represents the major differences to S-recombinase driven catalysis.

Y transposases are heterogeneous. They share the capacity to be transferred by a conjugative mechanism from cell to cell and are referred to as conjugative transposons (CTns) (2). Oftentimes, they carry antibiotic resistance determinants (usually the tetM gene) and are important vectors for disseminating this antibiotic resistance. Some are self transferable (e.g., Tn916 from Streptococci and Tn525 from Enterococci), whereas others are mobilized by functions provided in trans by other elements such as conjugative plasmids or other CTns (e.g., the NUS elements from Bacteroides spp.). They can have a very broad host range and seem to contribute significantly to gene transfer in complex bacterial populations.

Y2-transposases

These enzymes are members of a family of nucleases (34) known as the HUH superfamily because they include a con-
served histidine-hydrophobic-histidine motif that provides two of three ligands to an essential divalent metal ion cofactor. Gen-
erally, these enzymes are monomeric. Members include proteins that initiate conjugative plasmid transfer from cell to cell or catalyze rolling circle replication (RCR) in certain bacteriophages and plasmids. Prokaryotic members include the IS11 family and the newly identified ISCR group (35), whereas eukaryotic members include the helitrons (22).

The similarity between RCR and rolling circle transposition (RCT) is underlined by the fact that the “left end” of IS1 resembles a rolling circle replication origin—the structured region that is recognized and undergoes single-strand-specific cleavage to initiate replication. The enzymes carry five conserved blocks of amino acids, one of which includes a pair of tyrosine residues involved in catalysis. The best-characterized reactions that use Y2 enzymes are those involved in plasmid replication (e.g., pK174). They use two active site tyrosines and cleave the DNA by releasing the 3'-OH and 5'-phosphotyrosyl bond during cleavage.
Figure 2 Proposed mechanism of Y- and S-transposases. (a) The S-transposases. Proteins are shown in yellow and transposon ends in green. This figure is based on the known mechanism of serine site-specific recombinases. The serine nucleophiles are indicated (S) as are the 5′-OH liberated on formation of the phosphotyrosine bonds. (b) Synapsis and attack by the tyrosine. The top right and bottom left subunits are active for cleavage. (b2) Cleavage liberates the 3′-OH, which then attacks the phosphotyrosine bond on the opposite (horizontal as drawn) subunit to reveal the DNA strand in a first strand transfer step. (b3) The first strand transfer product. (b4) Second strand cleavage. The top left and bottom right transposase subunits are activated and cleave the opposite strands. (b5) Second strand transfer results in formation of a transposon joint and a donor backbone from which the transposon has been deleted. (b6) The Y-transposases. Proteins are shown in yellow and transposon ends in green. This figure is based on the known mechanism of tyrosine site-specific recombinases. The tyrosine nucleophiles are indicated (Y) as are the 5′-OH. (c) Synapsis and attack by the serine nucleophile. (c2) Rotation of the substrates (large arrow). (c3) Realignment and 3′-OH attack of the phosphoserine bonds. (a4) Resealing the broken strands generates a transposon circle and a reclosed donor backbone from which the transposon has been deleted. (c) The Y1 transposases. The left and right transposase subunits are activated and cleave the opposite strands. (c4) Excision of IS generates a transposon circle and a reclosed donor backbone from which the transposon has been deleted. (c5) Single stranded cleavage and transfer results in formation of a transposon joint and a donor backbone from which the transposon has been deleted as a transposon circle. (c) The Y1 transposases. The left and right transposase ends are shown in red and blue, respectively. (c1) Diagram that shows the potential secondary structures at the left and right ends. (2) TnpA binding and cleavage. The top strand is recognized by TnpA (yellow ellipses) that catalyzes cleavage via a 5′-phosphotyrosine linkage (black arrow). This cartoon shows that the covalent bond is formed at the 5′ end of the left end and at the 3′ end of donor flank at the right end. (c3) Cleavage is then followed by exchange between the left and right ends to generate a Right End–Left End transposon joint and to liberate a single strand transposon circle. (c4) Excision of IS and closure of the flanking backbone would leave a deletion bubble in the bottom strand. (c5, c6) Replication of the donor plasmid (arrow) would generate a copy of the original donor plasmid with an intact, double strand IS copy (c6), and a copy of the plasmid devoid of the IS (c5).
The RCT mechanism confers several interesting features on the IS91 family members. These elements do not carry terminal inverted repeats nor do they generate target site duplications as do the DDE transposons. IS91 itself inserts with a specific orientation at the 3' end of a conserved tetrancleotide sequence (5'CTTG3' or 5'GTTC3'), which is probably involved in transposition initiation and termination and also required for additional efficient transposition. Deletion of the downstream ("right") end results in "one-ended" transposition in which different lengths of vector DNA neighboring the deletion accompany the element to its new target site. These terminata with a 5'CTTG3' or 5'GTTC3' tetrancleotide located in the vector (2).

Y1-transposases

These enzymes were identified and characterized more recently (3). They are also part of the HUH superfamily of nucleases but currently characterized members have only one catalytic tyrosine and form obligatory dimers in both the bacteria and the archaebacteria (37). IS608 does not possess terminal IRs. The left (LE) and right ends (RE) include palindromic repeats, which form DNA hairpins located at some distance from the cleavage sites. The cleavage site at the left end is 19b upstream from the 5' end of the flanking donor DNA at RE (Fig. 2c1). IS608 insertion occurs 3' to a tetrancleotide TTAC contained in the flanking DNA, which abuts LE (38) directly (Fig. 2c2). Moreover, it is also required for subsequent transposition (7). IS608 transposition is strand specific. It occurs by precise excision of the "top" transposon strand to generate a circular intermediate with abutted copies of RE and LE (the transposon joint or RE–LE junction; Fig. 2c3) and resealing of the flanking DNA, which preserves the target TTAC (7). TnpA behaves as a dimer in solution and a crystal structure indicated that the molecule forms an elongated and flat dimer (8). The crystal structure of the complex formed by TnpA and a 22 nt long, single-stranded oligonucleotide that represents the RE palindrome showed a DNA hairpin (an imperfect palindromic dimer) bound to each of the two recognition sites in the TnpA dimer (8).

TnpA binds single strand (ss) DNA that carries either "top strand" LE or RE (see Fig. 2c2) much more strongly than double strand (ds) DNA ends. It did not bind the ss "bottom strand." TnpA catalyzes efficient strand-specific cleavage of a single "top" strand oligonucleotides, which includes LE or RE. IS608 transposition necessitates formation of an enzyme-substrate intermediate in which TnpA is attached covariantly at a DNA end via a 5' phosphotyrosine bond. Because of the conserved polarity of cleavages required at the ends, TnpA is joined to the 5' end of the transposon at LE but to the 5' end of the flanking donor DNA at RE (Fig. 2c2) (8). Nucleophilic attack of the LE phosphotyrosine bond by the free RE 3'OH results in resealing of the ss transposon to form a circle (Fig. 2c4), whereas attack of the RE 3' OH releases the donor backbone (Fig. 2c6). The transposon circle that carries an RE–LE junction can undergo insertion into a suitable TTAC target. The resulting donor molecule intermediate (Fig. 2c4) could be resolved by replication to generate a copy of the original donor plasmid molecule (Fig. 2c5) and a copy from the deleted transposon (Fig. 2c6).

Conclusions

Transposition is a general term that covers a variety of different TEs and includes several diverse chemistries. It should be underlined that although several partial structures of DDE enzymes have been determined, the structure of only a single complete DDE transposase has been obtained complexed with appropriate ends. For other types of transposase, two structures of Y1 transposases are available (8, 36), one of which includes its DNA substrate (25). A detailed structural information will be critical to understand how these fascinating enzymes function. The information presented above is derived largely from in vitro experiments, although a body of in vivo information also exists. One major challenge in the field of transposition at present is to determine how these elements interact with and are regulated by their respective hosts.

References


Further Reading
Group II Intron: Structure, Function, and Catalysis

Michael Roitzsch, Howard Hughes Medical Institute, Yale University, New Haven, Connecticut
doi: 10.1002/9780470048672.wecb679

The self-splicing group II introns are among the largest naturally occurring ribozymes and have been found primarily in organellar genomes of plants, fungi, bacteria, proteobacteria, and blue algae. More recently, a group II intron was found in a primitive metazoan. More than 200 different group II introns have been identified to date. All group II introns have a conserved secondary structure composed of six stem-loop structures termed domains D1–D6, which are arranged around a central wheel and provide distinct contributions to intron function. Group II introns catalyze their own excision from pre-RNA by two consecutive transesterifications and concomitant splicing of the flanking exons. The reaction requires Mg2+ ions, which are directly involved in catalysis and are also needed for proper folding of the intron. Intriguingly, group II introns share several mechanistic and structural features with the eukaryotic spliceosome, which suggests an evolutionary relationship. In vivo, group II intron splicing is generally assisted by proteins. With the help of proteins, some group II introns can also reverse the splicing reaction and integrate themselves into target genomes; this process resembles transposition of non-LTR retrotransposons and can be exploited for biotechnological applications.

It was long believed that protein enzymes were the only cellular machines that could conduct biological catalysis. In 1982, however, Thomas R. Cech described the self splicing group I intron from the ciliated protozoan *Tetrahymena thermophila*, which can excise itself from pre-rRNA and join the flanking exons in vitro (1). Cech coined the term "ribozyme" for this new class of biocatalysts. At the same time, Sidney Altman reported that the RNA components of RNase P from *Escherichia coli* and *Bacillus subtilis* could cleave tRNA precursor molecules in the absence of protein cofactors (2). Over time, it has become obvious that ribozymes are widespread in nature. Ribozymes that catalyze a broad range of reactions have been detected in numerous different organisms, which include plants, lower eukaryotes, bacteria, and viruses (3). More recently, a ribozyme was even found in the human genome (4).

Background

Group II introns are a distinct subgroup within the naturally occurring ribozymes

With exception of the ribosome that catalyzes peptidyl transfer reactions, all naturally occurring ribozymes catalyze phosphoryl transesterifications on their parent RNA (in cis) or other RNAs (in trans) and can be roughly divided into three categories based on their catalytic properties.

1. Nucleolytic ribozymes. Many smaller (40–150 nucleotides) ribozymes catalyze site-specific cleavage in cis. Mechanistically, they activate a 2'-OH group to allow nucleophilic attack of the adjacent 3'-phosphate, which results in formation of a cyclic 2',3'-phosphodiester and release of the downstream RNA strand (Fig. 1a). The reaction resembles non-specific base-catalyzed RNA cleavage. Members of this category are the hammerhead ribozyme, the hepatitis delta virus ribozyme, the structurally related mammalian CPEB3 ribozyme, the hairpin ribozyme, the varkud satellite ribozyme, and the glmS riboswitch. Their function is to process replication intermediates or, in the case of the glmS riboswitch, the metabolite-dependent control of gene expression.

2. RNase P. RNase P is a trans-acting ribozyme (150–500 nucleotides) that processes pre-RNAs by endonucleolytically cleaving off leader sequences from the 5'-end. RNase P positions and activates a water molecule to attack the scissile phosphate,
Figure 1 Schematic of phosphodiester bond-cleavage strategies used by different natural ribozymes. (a) In self-cleaving ribozymes, the 3′-OH group of the 3′-terminal nucleotide attacks the scissile phosphate, which results in a cyclic 2′,3′-phosphodiester. (b) RNase P activates a water molecule to allow a nucleophilic attack on the scissile phosphate, which forms a free 3′-OH group and a 5′-monophosphate at the resulting fragments. (c) In the first step of splicing of group I introns, the attacking nucleophile is provided by the 3′-OH group of a guanosine moiety. As a result, excised group I introns have an additional G at the 5′-terminus. (d) The attacking nucleophile in the first step of splicing of group II introns (branching pathway) is the 2′-OH group of a dedicated adenosine moiety (Abp) located close to the intron 3′-end in domain D6. This reaction results in a lariat form of the excised intron.

which results in a free terminal 3′-OH group

Figure 1b Although they are usually associated with proteins, RNase P ribozymes have been shown to catalyze strand cleavage in the absence of their protein cofactors.

3. Self-splicing ribozymes. These ribozymes are non-coding intervening sequences (introns) that can catalyze their own excision from the parent pre-RNA. Two groups of self-splicing ribozymes exist: the group I and group II introns. They can be very large with sizes of 200–1500 nucleotides (group I introns) and 400–2500 nucleotides (group II introns). Ribozyme-catalyzed self-splicing is a two-step reaction.

In group I introns, the attacking nucleophile in the first step is the 3′-OH group of an external guanosine (Fig. 1c). In the second step of splicing, the 5′-exon terminus attacks the 3′-scissile phosphate, which results in spliced exons and a linear intron with a 5′-terminal guanosine.

In group II introns, the attacking nucleophile in the first step is either the 2′-OH group of an adenosine located close to the intron terminus (Fig. 1d) or a water molecule (analogous to Fig. 1b). The second step is similar to that of group I introns and involves attack of the 5′-exon terminus on the 3′-splice site and ligation of the two exons.

Similarities to the eukaryotic spliceosome

In higher organisms, splicing is catalyzed by a ribonucleoprotein complex termed spliceosome. Several mechanistic and structural similarities suggest that the spliceosome may have derived from ancestral group II introns. Both group II introns and the spliceosome catalyze splicing in a two-step reaction that results in formation of excised lariat introns, and they share the same stereospecific preferences at the splice sites (vide infra). In both systems, the two splicing steps proceed through inversion at the scissile phosphates, which is consistent with an SN2 reaction mechanism (5). More strikingly, the highly conserved group II intron domain D5 has metal ion-binding properties and structural features similar to the stem loop of the spliceosomal small nuclear RNA U6. These similarities are so substantial that an isolated domain (D5) from a group II intron can functionally replace a subsection (U6) of the spliceosomal RNA in an in vivo assay (6). Group II introns are therefore considered a valuable model for studying functions related to the more complex spliceosome.

Protein assisted splicing

Group II introns are essentially unreactive under physiological salt conditions (vide infra). Therefore, splicing in vivo requires protein cofactors (7). Various proteins have been associated with splicing. These proteins can be roughly divided into three groups.
1. Intron encoded maturases. These proteins, which are normally encoded by an open reading frame (ORF) within the intron, bind specifically their own RNA and promote formation of critical tertiary contacts. In addition, most of these proteins have endonuclease and reverse-transcriptase domains, which allow intron mobility (i.e., retrotransposition and retrohomolog) by a mechanism termed targeted primed reverse transcription (TPRT) (vide infra).

2. Proteins recruited from the host. Many proteins have to the intron terminus. This reaction leads to a covalently linked branch point adenosine and formation of a new covalent link-age. To join the 5′-end of the 5′-oned 5′-exon (attacks in the first splicing step. In the branching pathway are two different pathways, which differ in the nucleophile that catalyzes self-splicing can proceed through

Reactions catalyzed by group II introns

Group II intron-catalyzed self-splicing can proceed through two different pathways, which differ in the nucleophile that attacks in the first splicing step. In the branching pathway (Fig. 2a), this nucleophile is the 2′-OH group of the branch point adenine, which is a bulged nucleotide located close to the intron terminus. This reaction leads to a covalently linked 2′,5′-branched intron structure with a lariat form. The second reaction step involves attack of the 3′-terminal OH-group of the cleaved 5′-exon on the 3′-splice site. The reaction products are the excised lariat intron and the ligated exons.

The alternative splicing mode is the hydrolytic pathway (Fig. 2b). As the name indicates, a water molecule is the nucleophile that attacks the 5′-splice site in this reaction mode, which generates a linear intron. The second splicing step is identical to that of the branching pathway.

It is remarkable that group II introns can also bind the spliced exons and use them as substrate for different reactions. The simplest of these reactions is the spliced exon reopening, which is an irreversible hydrolytic cleavage of the exons at the original splice site (Fig. 2c). A more interesting reaction is the complete reversal of both steps of splicing, which allow group II introns to reinvert themselves into the spliced exons (8). Reversal of the second step of splicing is an energetically neutral transesterification, which can be performed by linear and lariat introns. The situation is different for reversal of the first step of splicing, which has only been observed for the lariat intron. The lariat intron can reverse the first step of splicing by a second transesterification, which includes breaking of the 2′,5′-phosphodiester at the branch point adenosine and formation of a new covalent linkage with the 5′-exon. The linear intron cannot perform a second transesterification because it lacks the 2′,5′-phosphodiester linker.

Structure, Function, and Catalysis

1. Domain 1 (D1)

D1 is absolutely essential for catalytic activity. It provides a scaffold that helps the intron fold into the native tertiary structure. Several long-range interactions within D1, like the 3′-ε–1′-κ and the 3′-ω–ε–κ interactions, are part of a larger, functional substructure that also involves the II1 loop in the stem of D1. This substructure was named "3′-anchor" and is important for intron function because it positions the 5′-splice site near the bulge of D5. In all group II introns, D1 contains the exon-binding site EBS1; this site is a short sequence that base pairs with a complementary sequence termed intron binding site (IBS) located at the 3′-end of the 5′-exon. Group IIA and IIB introns have an additional second stretch of nucleotides (EB52) that interact with the 5′-exon. In addition, D1 contains a site termed EBS3 in group IIB and IIC and h in group IIA introns, which binds to the first nucleotide of the 3′-exon. The EBS1-IBS and, in group IIA introns, the h-ε–κ interactions are of fundamental importance for the introns as they allow recognition and binding of the exons and determination of the splice site.
Group II Introns: Structure, Function, and Catalysis

Figure 2  (a) In the branching pathway the phosphate at the 5′-splice is attacked by the 2′-OH group of the branch point adenosine. In the second step of splicing, the 3′-terminal OH-group of the 5′-exon attacks the phosphate at the intron 3′-exon junction, which results in spliced exons and excised lariat intron. (b) In the hydrolytic pathway, the first step nucleophile is a water molecule that cleaves the 5′-intron, which forms a linear intron. The second step of splicing is essentially the same as in the branching pathway. (c) Spliced exon reopening requires binding of the substrate. The substrate is then irreversibly cleaved by activation of a water molecule.

Domain 2 (D2)
In contrast to D1, D2 is not necessarily required for splicing, and most of this domain can be removed without severe effects on the catalytic activity. However, it stabilizes the tertiary structure by forming long-range interactions to D1 (θ–θ′) and D6 (η–η′).

Domain 3 (D3)
D3 functions as a catalytic effector that strongly enhances the chemical rate of catalysis. To date, the μ–μ′ interaction is the only identified tertiary interaction between D3 and D5.

Figure 3  All group II introns share a common secondary structure with six domains that radiate from a central wheel. Based on differences in some features, group II introns can be divided into subgroups IA, IB, and IC. Long-range tertiary interactions are indicated by greek letters. The folding control element is indicated. The catalytic triad in domain D5 is marked by a rectangle. The branch point adenosine in domain D6 is shown as a bold “A.” Exons are depicted as gray boxes and intron binding sites are indicated (adapted from Reference 20).
Domain 4 (D4)

D4 is not needed for splicing. In some introns, it contains an ORF encoding a maturase protein.

Domain 5 (D5)

D5 is a short stem-loop structure typically composed of 34 (group IIA and IIB) or 32 (group IIC) nucleotides with a conserved two-nucleotide bulge. It is the most conserved intronic domain and absolutely necessary for catalysis. It forms several contacts (i-ι', κ-κ', γ-γ') to D1 as well as a contact to D3 (ι-ι'). A very important feature of D5 is a highly conserved stretch of three consecutive nucleotides named the catalytic triad, which is located at the 5'-end of D5 and is absolutely required for catalysis. The sequence of the catalytic triad is AGC in group IIA and group IIB introns and CGC in group IIC introns. The catalytic triad and the nucleotides at the bulge form several direct contacts to divalent metal ions, which help to position them into the catalytic core.

Domain 6 (D6)

D6 is a hairpin structure that harbors the branch point adenosine, whose 2'-OH group acts as the nucleophile for the first step of splicing in the branching pathway. In addition, D6 contains an contact that forms the η-η' contact with D2. This element is a GWRA tetraloop-receptor interaction that is believed to mediate a conformational switch between the two steps of splicing.

J2/3 linker

Besides the six domains, the region joining domains D2 and D3 (J2/3) is particularly conserved and catalytically important. It forms the γ-γ' interaction, which joins J2/3 with the intron terminus and contains two consecutive nucleotides, GA, in group IIA and IIB introns and GC in group IIC introns, which are critical for intron function. Photocrosslinking studies have shown that J2/3 is located in the proximity of the catalytic triad in D5. A recent crystal structure confirms these findings and shows that J2/3 forms base triplets with the catalytic triad and the bulge of D5, which helps to bring together catalytically important residues of the intron (12).

Trans splicing constructs and kinetic investigations

In vitro kinetic assays are an important tool to study the catalytic mechanisms of ribozymes. Intron mutant and deletion constructs have been used in kinetic assays to study the role of particular nucleotides and to distinguish whether a mutation causes a defect in chemistry or binding of exonic substrate or intronic components.

A very interesting feature of group II introns is their modularity; the intronic domains can be provided separately to form a functional ribozyme (13). For example, the exD123 construct consists of a 5'-exon and D1, D2, and D3. This construct alone is inactive, but addition of the catalytically essential domain D5 in trans generates an active two-piece ribozyme that cleaves off the 5'-exon; it is therefore a mimic for the hydrolytic pathway of the first step of splicing. Similarly, an intron construct that consists of D5 and D6 (D56) can be combined with the exD123 construct to obtain a branching pathway mimic for the first step of splicing. To date, numerous other multipiece constructs have been successfully studied and some examples of naturally split group II introns have even been found plants (9). An important feature of trans splicing constructs in kinetic assays is that the individual constructs can be folded separately under appropriate conditions without reacting, and hence they allow the separation of folding from catalysis.

Group II introns are metal-dependent enzymes

Group II introns are metal-dependent enzymes that require Mg\(^{2+}\) for both folding and catalysis (14). Most group II introns are essentially unreactive at physiological salt concentrations. Hence, for in vitro experiments, nonphysiologically high salt concentrations, typically 500 mM of a monovalent and 100 mM Mg\(^{2+}\) are required to obtain optimal splicing conditions. In vivo, these high metal requirements are believed to be relieved by intron-encoded maturases and other cofactor proteins recruited from the host organism.

Metal ions have been localized by independent methods, which include phosphorothioate substitution (13), Tb\(^{3+}\) cleavage (14), X-ray crystallography (12), and NMR (15). Metal ion-induced metal contacts are provided by the phosphate backbone. Although the bridging phosphates are not chiral by chemical definition, the two nonbridging oxygen atoms are distinct in the tertiary structure. The nomenclature used to define these atoms is derived from phosphorothioate substitution experiments. Substitution of a single oxygen atom with sulfur at a bridging phosphate generates a chiral phosphorothioate diastereomer named as the Rp- and Sp-isomers according to the Cahn-Ingold-Prelog system. The nonbridging oxygens of a bridging phosphate can be identified accordingly as pro-S or pro-R oxygens.

Metal ions are required for folding

Studies with the group IIB intron ai5 from the mitochondrial genome of S. cerevisiae have suggested that group II introns fold slowly into the native state via an on-pathway intermediate (16). In contrast, most other ribozymes proceed along a "rugged" folding pathway with stable misfolded intermediates ("kinetic traps"). From the unfolded state, group II introns first fold slowly into a compact intermediate state and then, in a fast step, into the native state. At low Mg\(^{2+}\) concentrations, the introns can only reach the compact state, which was exploited to study this state in more detail. Using nucleotide analog interference mapping with compaction as the selection criterion, a small substructure in the heart of D1 was identified as a folding control element (Fig 2). This element must adopt the correct conformation before other tertiary contacts of D1 can be established, leading to a highly compact tertiary structure of D1. Once the D1 scaffold is formed, high Mg\(^{2+}\) concentrations or a cofactor can facilitate rapid docking of the other domains into their designated binding sites thus forming the native (i.e., catalytically active) ribozyme.
Figure 4 Chemical mechanism of group II intron splicing (branching pathway). The sequences and numbering are according to the ady intron from S. cerevisiae. The solid arrows between the panels mark the forward splicing direction, the dotted arrows the reverse splicing direction. (a) The first step is initiated by attack of the 2′-OH group of the branch point adenosine (A880) at the scissile phosphate as indicated by the arrow. The dotted lines indicate coordinative bonds to metal ions. Metal ion coordinations from the catalytic triad (A816-C818) and the bulge in DS (C837-C839) are indicated in this panel but omitted in the following panels for clarity. (b) In the transition state of the first step of splicing, the scissile phosphate is expected to have a trigonal bipyramidal arrangement. (c) In the splicing intermediate, the metal coordinated oxygen of the 5′-scissile phosphate has formally changed from pro-RP to pro-SP. The second step of splicing is initiated by attack of the 3′-terminal oxygen of the 5′-exon at the 3′-scissile phosphate as indicated by the solid arrow. The dotted arrow indicates the reverse splicing reaction. (d) The transition state of the second step of splicing is expected to have a trigonal bipyramidal arrangement at the scissile phosphate similar to the transition state of the first step of splicing. (e) Schematic of the products immediately after the second chemical step. This arrangement is also the initial configuration expected for the first step of reverse splicing as indicated by the dotted arrow.
Metal ions in the catalytic core

Many ribozymes and proteins that catalyze phosphoryl-transfer reactions use a mechanism employing two metal ions, and early on group II introns were hypothesized to use a similar mechanism (17). However, evidence for the existence of two metal ions in the catalytic core was only found very recently (12, 13, 18). A direct Mg$^{2+}$ coordination of the pro-S$_5$ oxygen of the first nucleotide of the catalytic triad is evident based on phosphorothioate substitution experiments (18). Recently, an intact group II was successfully crystallized for the first time (12). This structure confirms the metal contact from the first nucleotide of the catalytic triad and shows additional contacts to this metal ion from the second nucleotide of the catalytic triad and from the first nucleotide upstream of the bulge (Fig. 4a). This nucleotide is additionally coordinated to the second Mg$^{2+}$ ion in the core. The distance between the two Mg$^{2+}$ ions in the crystal structure is 3.9 Å, which is in agreement with the proposed two-metal-ion mechanism (17).

Because the data from phosphorothioate substitution experiments, which provide information about the metal binding situation before splicing, match the crystal structure that captures the situation after splicing, it may be concluded that the two metal ions are tightly bound in the catalytic center in an arrangement that remains essentially unchanged through both steps of splicing.

During splicing, the two metal ions form additional contacts to the substrate. Evidence from phosphorothioate substitution experiments are coordinative bonds to the pro-R$_5$ oxygens of the scissile phosphates at both splice sites and to the 3′-oxygen of the leaving groups in both steps of splicing (Fig. 4).

The chemical mechanism of group II intron splicing

The following paragraph will discuss the detailed chemical mechanisms of both steps of splicing by the branching pathway. Before the first chemical step of splicing takes place, the substrate is properly arranged in the catalytic core, which is mainly facilitated by the EBS-DSS and S-S interactions, and the metal ions are positioned as described above (Fig. 4a).

The first step of splicing

In the first step of splicing, the 2′-OH group of the branch point adenosine attacks the scissile phosphate at the designated 5′-splice site. It is not established how the nucleophile is activated; however, it is likely that the 2′-OH is coordinated to a metal ion in the core that could facilitate deprotonation (17).

Both steps of splicing are known to proceed with inversion at the scissile phosphates, which is consistent with an Sn2 substitution (5). In the transition state, the scissile phosphate is therefore expected to adopt a trigonal bipyramidal arrangement (Fig. 4b). Metal ion coordination to a nonbridging oxygen of this phosphate might help to compensate the emerging negative charge in the transition state. The leaving group (i.e., the 3′-oxygen of the 5′-exon) is likewise stabilized by metal ion coordination. In the resulting intermediate, the metal-ion coordinated oxygen atom of the 5′-splice site phosphate has formally changed from pro-R$_5$ to pro-S$_5$ because of the inversion (Fig. 4c).

The second step of splicing

The leaving group of the first step of splicing, which is the 3′-terminal OH group of the 5′-exon, becomes the attacking nucleophile in the second step of splicing. Metal coordination properties of the participating reaction centers are comparable to the first step. The pro-R$_5$ oxygen of the scissile phosphate, the 3′-oxygen of the 3′-terminal intron nucleotide (the leaving group) and the 3′-terminal nucleotide of the 5′-exon (the attacking nucleophile) are coordinated to divalent metal ions (Fig. 4c).

The transition state of the second step of splicing resembles that of the first step. The scissile phosphate adopts a trigonal bipyramidal arrangement and is stabilized by metal coordination to a non-bridging oxygen atom (Fig. 4d). The leaving group, i.e., the 3′-oxygen of the terminal intron nucleotide, is likewise stabilized by metal coordination. The resulting products are the lariat intron and the spliced exons (Fig. 4e). This arrangement is also the starting conformation for the reverse splicing reaction, which is indicated by dotted arrows in Fig. 4.

Future Research Topics

The story of group II introns is not at the end. The recent crystal structure of an intact group II intron will presumably lead to several follow-up studies. For example, it would be interesting to crystallize and compare group II introns from different subgroups or to capture different reaction states of the intron.

Group II introns naturally catalyze a broad range of reactions. The catalytic repertoire of other ribozymes was previously extended by in vitro evolution. It could be very exciting to apply this method to the versatile group II introns.

The ability to invade duplex DNA has led to useful technological applications. By manipulating the exon-binding sites in domain D1 of the intron, which recognize potential target sites, group II introns can be reprogrammed to attack a given sequence selectively. This method was used to develop group II introns into gene-targeting vectors ("targetrons") (19). Targetrons may have a great potential for medical applications. They could be used to knock out or to repair a malfunctioning gene of choice. A nother idea is to engineer the intron ORF in D4 to contain additional coding sequences; group II introns could then be used to introduce new genes into an organism.

Acknowledgments

The author thanks Anna M. Pyle, Amanda Solem, Nora Zingler, and Olga Fedorova for critical reading of the manuscript and useful discussions.

References

Group II Introns: Structure, Function, and Catalysis


Further Reading


http://www.fp.ucalgary.ca/group2introns/

See Also

Bacterial DNA polymerases, chemistry of Origins of life: emergence of the RNA world RNA, noncoding

Further Reading


http://www.fp.ucalgary.ca/group2introns/
Noncoding RNAs
Allison A. Henry and Jörg Vogel, Max Planck Institute for Infection Biology, Berlin, Germany
doi: 10.1002/9780470048672.wecb522

After decades of being either underestimated as a simple messenger in the expression of proteins from deoxyribonucleic acid (DNA), or revered as the molecule from which all life potentially originated, ribonucleic acid (RNA) is at last enjoying a well-deserved golden age as modern techniques allow researchers to uncover the myriad of subtle roles that this molecule plays in the cells of today. Noncoding RNAs, by definition, do not become translated into protein. This broad category includes cellular RNAs that range in size from 20 to 20,000 nucleotides, can contain chemical modifications, adopt a wide variety of secondary and tertiary structures, associate with proteins and other RNA molecules, and have physiological roles, in all organisms, that run the gamut from catalysis to regulating gene expression. As this article is intended to complement the other RNA-related sections of this volume, a brief introduction to all types of noncoding RNA will be followed by a comprehensive discussion of three important categories: riboswitches, small RNAs in bacteria, and large RNAs in mammals. Riboswitches are cis-acting RNA elements that modulate gene expression in a highly specialized manner by means of an aptamer domain, an RNA sequence that tightly and selectively binds a given metabolite. Small RNAs are 70–200 nucleotide molecules in bacteria that regulate gene expression, primarily at a posttranscriptional level. In mammalian cells, large RNAs of up to 19,000 nucleotides are essential to X-chromosome inactivation and genomic imprinting.

To a chemist, who is accustomed to thinking about the enormous diversity of structures and functions to be found in small molecules, ribonucleic acid (RNA) may seem humdrum. It is, after all, a polymer composed of only four subunits, which comprise a ribose ring and a purine or pyrimidine, that are linked together by phosphodiester bonds. From a simple-minded perspective, RNA has fivefold fewer types of functionality than proteins, which are made up of 20 different residues; in fact, RNA is similar in composition to deoxyribonucleic acid (DNA), the preferred molecule of genetic information storage that does not have a reputation for carrying out interesting chemistry inside the cell. Such an observer would likely be content with the notion that the roles of RNA are limited to coding for proteins and serving as viral genomes, in accordance with the biological dogma that existed until about three decades ago. Since then, new technologies have enabled significant progress in our understanding of the diversity of structures and functions of RNA molecules that exist in the cells of all organisms. Concomitant with our discovery and characterization of new RNA molecules, the nomenclature in the field has expanded, and for the sake of clarity, we risk stating the obvious: RNA that codes for protein is referred to as messenger RNA (mRNA) and all others are noncoding RNAs, by definition. Three categories of noncoding RNA, riboswitches, regulatory small RNAs in bacteria (sRNAs), and large noncoding RNAs in mammals, will be discussed in some detail in this article, but first a summary of the features and functional repertoire of noncoding RNAs in general is presented.

Sizing Up Cellular Noncoding RNAs
Noncoding RNAs populate virtually all polymer lengths between 20 and 20,000 nucleotides and are found in viruses, prokaryotic cells, and eukaryotic cells (Table 1). After they become transcribed, these molecules can be cleaved and chemically modified. As many as 100 potentially functionally meaningful nucleotide modifications have been reported, including nucleobase and sugar methylations, the substitution of sulfur for oxygen, C-glycosidic linkages like the one in pseudouridine, and even condensation products of amino acids and nucleobase exocyclic amino groups (1–3). However, most types of RNA
## Table 1. Some characteristics of noncoding RNAs within the cell

<table>
<thead>
<tr>
<th>RNA</th>
<th>Size (nt)</th>
<th>Organism</th>
<th>Physiological role</th>
</tr>
</thead>
<tbody>
<tr>
<td>hTR</td>
<td>150–1300</td>
<td>eucarya</td>
<td>telomerase-mediated telomere synthesis</td>
</tr>
<tr>
<td>MLH1</td>
<td>350–450</td>
<td>universal</td>
<td>ribonuclease P-mediated pre-rRNA cleavage</td>
</tr>
<tr>
<td>group I introns</td>
<td>200–1000</td>
<td>bacteria, eucarya</td>
<td>catalyze self-excision from pre-mRNAs, rRNAs, and tRNAs</td>
</tr>
<tr>
<td>group II introns</td>
<td>600–1000</td>
<td>bacteria, eucarya</td>
<td>catalyze self-excision from pre-mRNAs, rRNAs, and tRNAs, and reverse splicing into DNA</td>
</tr>
<tr>
<td>hepatitis δ ribozyme</td>
<td>85</td>
<td>hepatitis δ virus</td>
<td>catalyzes self-cleavage within viral genomic RNA during rolling-circle replication</td>
</tr>
<tr>
<td>hairpin ribozyme</td>
<td>50</td>
<td>tobacco ringspot virus</td>
<td>catalyzes self-cleavage and ligation within viral genomic RNA during rolling-circle replication</td>
</tr>
<tr>
<td>hammerhead ribozyme</td>
<td>78</td>
<td><em>S. mansoni</em></td>
<td>catalyzes self-cleavage and ligation within viral genomic RNA during rolling-circle replication</td>
</tr>
<tr>
<td>srRNA</td>
<td>100–200</td>
<td>eucarya</td>
<td>spliceosome-mediated pre-mRNA splicing</td>
</tr>
<tr>
<td>snRNA</td>
<td>60–300</td>
<td>archaea, eucarya</td>
<td>pre-rRNA processing and modification; srRNA chemical modification</td>
</tr>
<tr>
<td>gRNA</td>
<td>70</td>
<td>trypanosome protists</td>
<td>uridine deletion and insertion within mRNA</td>
</tr>
<tr>
<td>rRNA0</td>
<td>120–4.7 kb</td>
<td>universal</td>
<td>catalyzes peptide bond formation in protein biosynthesis</td>
</tr>
<tr>
<td>peRNA</td>
<td>70–90</td>
<td>universal</td>
<td>protein biosynthesis; transcriptional regulation of aminoaacyl-tRNA synthetases</td>
</tr>
<tr>
<td>trnRNA</td>
<td>363</td>
<td>bacteria</td>
<td>rescues stalled ribosomes</td>
</tr>
<tr>
<td>4.5 S/5S</td>
<td>79–522</td>
<td>universal</td>
<td>signal recognition particle-mediated protein secretion</td>
</tr>
<tr>
<td>siRNA</td>
<td>20–23</td>
<td>eucarya</td>
<td>miRNA cleavage</td>
</tr>
<tr>
<td>miRNA</td>
<td>20–23</td>
<td>eucarya</td>
<td>translational repression; miRNA cleavage</td>
</tr>
<tr>
<td>rasiRNA</td>
<td>23–28</td>
<td>eucarya</td>
<td>transcriptional silencing; regulation of chromatin structure</td>
</tr>
<tr>
<td>tncRNA</td>
<td>19–23</td>
<td><em>C. elegans</em></td>
<td>unknown</td>
</tr>
<tr>
<td>piRNA</td>
<td>26–31</td>
<td>mammals</td>
<td>suggested antagonists of Piwi proteins</td>
</tr>
<tr>
<td>cis-acting RNA (riboswitches)</td>
<td>100–400</td>
<td>bacterial</td>
<td>transcriptional and translational regulation of gene expression; respond to temperature, metabolite, protein, or RNA</td>
</tr>
<tr>
<td>sRNA</td>
<td>70–250</td>
<td>bacteria</td>
<td>miRNA accumulation; translational activation; mRNA degradation; translational repression</td>
</tr>
<tr>
<td>Xist</td>
<td>17–19 kb</td>
<td>mammals</td>
<td>initiates silencing on appropriate X-chromosome</td>
</tr>
<tr>
<td>Tsix</td>
<td>30 kb</td>
<td>mammals</td>
<td>blocks initiation of silencing by Tsix on appropriate X-chromosome</td>
</tr>
<tr>
<td>Xite</td>
<td>12 kb</td>
<td>mammals</td>
<td>mediates persistence of Tsix throughout X-chromosome inactivation</td>
</tr>
</tbody>
</table>

1See text for references.
2RNA component of ribonucleoprotein complex; in order of appearance in the text.
3Approximate.
4Specifically, in the organelles of eukaryotic cells.
5Although biologists are used to thinking about tRNA, rRNA, and noncoding RNA as altogether different classes of RNA, for historical reasons, tRNA and rRNA are named here as types of noncoding RNA because they do not contain open reading frames that encode proteins.
6Isolated example of cis-acting elements in eukaryotic mRNAs have been reported, but this type of regulatory mechanism appears to be broadly used only by bacteria.
7Excluding internal open reading frames.
Noncoding RNAs

with the notable exception of transfer RNA (tRNA), are not heavily modified. In keeping with such complexity in primary structure, RNA is known to adopt an impressive array of secondary and tertiary structures, and many of these have been elucidated by molecular modeling, biochemical structure determination, and spectroscopic and crystallographic techniques. Two common RNA secondary structure motifs are the helix, defined as two or more consecutive GC, AU, or GU (canonical) pairs, and the loop, defined as a region in which the nucleotides do not form canonical pairs. Loop structures are further categorized as hairpin, bulge,内部branch, or pseudoknot (4). Beyond Watson–Crick base pairing, ribonucleotides also interact with one another by means of their Hoogsteen and sugar-edge surfaces, and combinations of such pairs as these define specific RNA motifs (5). RNA motifs mediate not only tertiary RNA–RNA contacts, but also quaternary RNA–protein interactions. It should be noted here that, without exception, RNAs within the cell are associated with proteins; however, in the interest of brevity, the discussion below will be limited primarily to the RNA component of each ribonucleoprotein.

Maintenance of Telomeric DNA

Telomerase is a ribonucleoprotein complex that exists in eukaryotic cells for the apparently sole purpose of synthesizing telomeric DNA, which consists of tandemly repeated sequences that contain clusters of G-residues and forms the ends of chromosomes. Telomerase comprises two essential core components, a protein subunit that has reverse transcriptase (RT) activity and an RNA sequence (hTR) that contains clusters of G-residues and serves as the template substrate for the RT (6). The template RNA and its G-rich RNA annex to form a partial duplex with DNA as the primer. RT-mediated polymerization of dGTP and other complementary triphosphate substrates produces a DNA terminus that has been extended by around six nucleotides. The new end can become a substrate for either another round of telomerase-mediated elongation or primase/polymerase-mediated ligation-synthetic synthesis.

Processing of RNA

RNA molecules that catalyze RNA strand scission and RNA ligation form one class of ribozymes, or RNA-based enzymes. Well-characterized examples of naturally occurring ribozymes include the RNA moiety of ribonuclease P (M1 in bacteria, H1 in mammals) (7), group I introns (8), group II introns (9, 10), and the hepatitis delta virus (11), hairpin (12), and hammerhead ribozymes (13). Other classes of noncoding RNA that either guide or participate directly in the cleavage, ligation, or chemical modification of various RNA molecules have emerged more recently. Small nuclear RNA (snRNA) is known for its nucleus of eukaryotic cells, and as a component of the spliceosome, which, in conjunction with 5m proteins, processes pre-mRNA and prepares it for export to the cytoplasm (14). Small nuclear RNA (snRNA), comprising many distinct individuals, and small Cajal body-specific RNA (scRNA) are also responsible for processing RNA in eukaryotic cells. snRNAs reside in the nucleolus and guide the cleavage and chemical modification, including sugar methylation and pseudouridylation, reactions that are necessary for large and small ribosomal RNA subunit assembly and export to the cytoplasm (15). Certain snRNAs are similarly acted upon by snoRNAs (16). snoRNAs are localized to foci in the nucleoplasm, known as Cajal bodies, and are involved in the chemical modification of most snRNAs that are transcribed by RNA polymerase II (17). Finally, the guide RNAs (gRNAs) that are found in bysynosome protists direct the site-specific insertion and deletion of uridines in mRNA sequences, a phenomenon referred to as editing (18).

Synthesis and Trafficking of Protein

Ribosomal RNA (rRNA) was long thought to play a passive, structural role in protein synthesis, but it is now known to also be a ribozyme that catalyzes peptide bond formation (19). The newest form of noncoding RNA that participates directly in protein synthesis is rRNA, the so-called adapter molecule found in all organisms that contains both the anticodon RNA sequence and its cognate amino acid residue as a chemically activated aminoacyl–methyl (20). Two housekeeping RNAs, tmRNA (21) and 4.5 S RNA (22), a component of the signal recognition particle, function to maintain the competence of the ribosome and to ensure the secretion of certain proteins, respectively.

RNA Interference

RNA interference, a mechanism of gene silencing in eukaryotic cells that relies on very short sequences of single-stranded RNA, is arguably the one area of recent noncoding RNA research that has come most to fruition. Small interfering RNAs (siRNAs), microRNAs (miRNAs), repeat-associated small interfering RNAs (rasiRNAs), and tiny-noncoding RNAs (tnRNAs) are short RNA molecules, approximately 22 nucleotides, that have been best characterized in the nematode, C. elegans (23, 24), but also studied in mammals, fish, insects, and plants. The genes that encode siRNAs and miRNAs are transcribed into double-stranded (dsRNA) or short hairpin primary (pri-mRNA) RNAs, respectively. Whereas dsRNA is exported directly to the cytoplasm, pri-mRNA is initially processed to pre-miRNA at rRNAase III-like endonuclease, Drosha, in the nucleus. In the cytoplasm, dsRNA and pre-miRNA are acted upon by the RNase III-like endonuclease, Dicer, to yield short dsRNA molecules with 2-nucleotide 3′ overhangs and 5′ phosphates that are known as siRNAs. In both cases, one strand of siRNA goes on to form a complex with a member of the Argonaute family of proteins, where it acts as a guide in binding to the miRNA target. This ribonucleoprotein complex is referred to as RISC in the siRNA pathway and miRNP in the miRNA pathway. The single-stranded siRNA then binds to its miRNA target by means of perfect (siRNA) or partial (miRNA) complementarity, and it brings about translational repression that may or may not be accompanied by protein-mediated endonucleolytic cleavage of the miRNA target. The newest class of these small mammalian RNAs, Piwi-interacting RNAs (piRNAs), was named for the interaction of its members with Piwi proteins, a subgroup of the Argonaute family (25, 26). Piwi proteins play a role in cell differentiation, and preliminary evidence suggests that piRNAs act as antagonists of their functions.
Riboswitches

RNA sequences that tightly and selectively bind to a molecule of choice, small or large, can be engineered with relative ease; they are referred to as aptamers. For example, in vitro selections have enabled the isolation of RNA sequences that bind to biologically interesting molecules like nucleotides (27), coenzymes (28), amino acids (29), sugars (30), and many others (31). In light of this observation, it is perhaps not surprising that this intrinsic property of RNA is also exploited by nature for the purpose of regulating gene expression with the utmost economy. Riboswitches are cis-acting regulatory RNAs that are found in the 5' untranslated region (5' UTR) of some mRNA molecules and are able to modulate gene expression in response to small molecule metabolites by means of their aptamer and expression platform domains (32). The structures of the domains are coupled in such a way that binding of the ligand directly changes the efficiency of transcription or translation of the mRNA, allowing the cell to circumvent the need for a protein-based metabolite sensor (33). Metabolite-responsive riboswitches represent one type of cis-acting regulatory RNA. Other types respond to temperature, protein ligands, and RNA ligands. Although these types will not be further discussed, recent reports regarding their structure and function are recommended to the reader (34–38).

The nature of the expression platform domain of the riboswitch, and the way in which its structure is coupled to the aptamer domain, in the presence and absence of ligand, dictates whether the riboswitch will affect the extent of transcription or the extent of translation of the mRNA of which it is part (39). Illustrative examples are given below for both transcription-based and translation-based mechanisms. Other less general riboswitch strategies have been reported, including RNA splicing (40, 41), self-cleavage (30), antisense (42), a potential role for RNase P (43), and ribosome-mediated attenuation (44), but these strategies will not be discussed further.

Transcription Effects

Genes are encoded by DNA and become transcribed into RNA through the work of RNA polymerases. This process has three distinct phases: initiation, elongation, and termination (45). On average, elongation in bacteria is highly processive, with about 50 nucleotides per second being incorporated over the length of the transcription unit; however, there do exist natural pauses 50 nucleotides per second being incorporated over the length average, elongation in bacteria is highly processive, with about distinct phases: initiation, elongation, and termination (45). On through the work of RNA polymerases. This process has three genes are encoded by DNA and become transcribed into RNA secondary structure formation occurs when the RNA segment favors a conformation in which an antiterminator (AT), rather than a terminator (T), is present and polymerase-mediated transcription proceeds. In the case of negative regulation, which is not depicted, the presence of ligand (L) causes the RNA segment to favor a conformation in which a terminator (T), rather than an antiterminator (AT), is present, halting polymerase-mediated transcription. The filled bars serve only as orientation guides and do not imply a certain extent of complementarity or stability within the hairpins.

Figure 1 Representation of transcriptional regulation by the binding of a ligand to a cis-regulatory element of nascent mRNA. Depicted is positive regulation: In the presence of ligand (L), the RNA segment favors a conformation in which an antiterminator (AT), rather than a terminator (T), is present and polymerase-mediated transcription proceeds. In the case of negative regulation, which is not depicted, the presence of ligand (L) causes the RNA segment to favor a conformation in which a terminator (T), rather than an antiterminator (AT), is present, halting polymerase-mediated transcription. The filled bars serve only as orientation guides and do not imply a certain extent of complementarity or stability within the hairpins.
in a fully bound state and, importantly, glycine binding is cooperative: the binding affinity of the second glycine is 100- to 1000-fold greater than that of the first glycine. This cooperativity of ligand binding, common in protein molecules, has the effect of increasing the sensitivity of the guanine riboswitch, allowing dramatic changes in transcription to take place in response to small changes in the concentration of guanine. It was suggested that this mechanism, found in the same study to also exist in the putative sodium and alanine sminporter encoded by the V. cholerae, evolved to ensure that excess glycine is efficiently used as an energy source.

Riboswitches negatively regulate gene expression if binding of the metabolite to the aptamer domain results in the formation of a terminator in the expression platform. Examples of riboswitches that conform to this mechanism are more numerous, as are associated structural studies. High resolution crystal structures exist for the liganded forms of both the 68 nucleotide aptamer domain of the guanine-sensing pur riboswitch of B. subtilis and the 5′-adenosylmethionine (SAM)-sensing riboswitch from T. tengcongensis (48, 49). In the G-riboswitch structure, the ligand is stacked within a five-tiered arrangement of base triples, and it participates in eight direct H-bonding interactions with four aptamer nucleotides; three H-bonds represent a canonical Watson-Crick base pair between guanine and a C nucleotide, determining specificity of the riboswitch; three H-bonds are formed between the minor groove heterocatoms of guanine and the Watson-Crick base edge of a U nucleotide; the two remaining H-bonds involve the minor groove oxygen and 2′ OH of two independent U nucleotides. In the SAM-riboswitch structure, the ligand adopts a compact conformation that enables both the formation of a cation-pi interaction between its main chain amino group and adenine moiety, and the possibility of participating in two distinct sets of aptamer interactions: the side of the ligand that presents the Watson-Crick face of adenine and the main chain atoms of methionine is engaged in H-bonding and stacking interactions with one helix while the ribose ring and side chain of SAM have favorable van der Waals interactions with a different helix. Despite their differences in ligand binding, the two riboswitches share a remarkably similar overall structure in which a key helix, the P1 helix, is formed as part of either a three-helix-turning-fork-type structure (G-riboswitch) or a four-way junction structure (SAM-riboswitch). In the presence of the cognate metabolite, each structure stabilizes an adjacent terminator helix and prevents transcription of the downstream genes. In the absence of the cognate metabolite, the P1 helix and the terminator dissociate, allowing the nucleotides that otherwise form the 3′ side of the P1 helix to form an anti- terminator helix with the nucleotides that otherwise form the 5′ side of the terminator. This mechanism is thought to be shared by other riboswitches that operate at the level of transcription, such as the FMN-responsive riboswitch that negatively regulates expression of the ribDEAHt operon in B. subtilis (50) and the divergent magnesium-sensing riboswitch that negatively regulates mpt expression in S. typhimurium (51).

**Translation Effects**

Protein biosynthesis in prokaryotes also proceeds through distinct phases known as initiation, elongation, and termination, which are not to be confused with the phases of transcription (see above). Initiation is characterized by the assembly of the small subunit (30S, comprising 21 proteins and the 1500 nucleotide 16 S rRNA) and the large subunit (50S, composed of 34 proteins, the 120 nucleotide 5S rRNA, and the 2900 nucleotide 23S rRNA) of the ribosome, together with the initiator tRNA that is charged with N-formylmethionine, on the 30 nucleotide ribosome-binding site (RBS) of a given mRNA (52). The RBS, located within the 5′ UTR of the transcript, has several important features that are common among most mRNAs. From 5′ to 3′, there is a pyrimidine-rich region that interacts with ribosomal protein S1; the Shine-Dalgarno (SD) sequence, typically GGAGG, forms base pairs with the 16 S rRNA; the initiation codon is AUG in the majority of cases, and UUG or UUA in a few cases; finally, there is a downstream box (DB) that also forms base pairs with the 16 S rRNA. Therefore, the precise sequence and structure of the mRNA are of paramount importance to the efficiency with which it becomes translated into protein. Many riboswitches regulate gene expression at the stage of translation initiation. Their mechanisms are analogous to riboswitch regulation at the transcriptional level, described above, except that ligand binding results in the formation of a structure that either sequesters or liberates the RBS.

A tautomeric structure of the liganded forms of both the 71 nucleotide adenine-sensing aptamer domain of the add riboswitch in V. vulnificus, an example of positive translational regulation, were recently solved. The structure of the A-riboswitch is very similar to that of the G-riboswitch, described above, except that the specificity-determining pyrimidine is a U rather than a C, and that the tuning fork structure containing the P1 helix coexists with a single-stranded downstream region in which the RBS is available to bind the 30S subunit of the ribosome (48). In this case, the expression of adenine deaminase is increased in the presence of adenine. In contrast, translation of this mRNA is reduced in the presence of TPP, and this negative regulatory mechanism is also made clear by inspection of the X-ray crystal structure (53). In the complex, TPP adopts an extended conformation in which it acts as a bridge between two parallel helical domains that are connected to the P1 helix by means of a three-way junction. The 4-amino-5-hydroxymethyl-2-methylpyrimidine moiety of the ligand engages in interaction with the P1 helix and base stacking, and base triple formation with one helical domain while the pyrophosphate group is bound to the other helical domain by H-bonding and electrostatic interactions that are mediated by two hexacoordinated divalent magnesium ions. The P1 helix stabilizes a downstream hairpin in which the SD sequence and start codon of thiM are sequestered, precluding translation initiation. It is likely that similar structures play a role in the negative translational regulation of yjaa in B. subtilis by a flavin mononucleotide (FMN)-responsive riboswitch (50), and that of bsbA in E. coli by a coenzyme B12-responsive riboswitch (54).
Chromosomally Encoded Regulatory sRNAs in Bacteria

sRNAs that regulate gene expression in bacteria are a newly identified class of noncoding RNAs. These molecules are commonly 70-200 nucleotides in length and are found encoded within the chromosome in either a cis (antisense) or trans location relative to their target protein-coding genes. Although sRNAs can influence gene expression by interacting with protein targets, most of the individuals that have been characterized thus far act by means of either extensive or limited base pairing with mRNA targets. This class of regulators was discovered long ago in the plasmids harbored by enterobacteria, in bacteriophage, and in transposons; all of these are cis-encoded antisense RNAs that exploit their perfect complementarity to bind to the cognate transcript, which encodes a replication factor, toxin, or transposase, to name a few, and bring about transcription attenuation, mRNA degradation, or inhibition of translation (55). Within the past 5 years, much more extensive work has been done to identify and characterize sRNAs that are encoded within the genomes of bacteria. The history of our knowledge of such sRNAs has been described as a chronology with two distinct periods: the classic and the modern (56). In the classic age, about 12 novel RNA molecules were discovered, either intentionally by means of metabolic labeling or serendipitously. In the modern age, which began in 2001, the existence of new information, generated by new technology, prompted eight research groups to carry out comprehensive, systematic searches for more sRNAs (57-64). Both computational methods and direct experimental detection were used, and these contributions increased the number of confirmed chromosomally encoded sRNAs in E. coli to about 70 (65). Although characterization of the physiological roles and mechanisms of action of these sRNAs in E. coli has not been fully reported (65-64), the following summary of the work done so far illustrates that their expression patterns and functions are too diverse to be generalized.

Regulation of sRNAs

It is known that the steady-state level of many sRNAs depends on the phase of growth or environmental conditions of the cell, and obviously the concentration of the sRNA at any given time represents a sum of its expression, processing, and degradation. With respect to expression, some regulatory mechanisms have been solved. RpoS sRNA is positively regulated by the CpxR-CpxB phosphorelay system (66). OxyS sRNA is induced in the presence of hydrogen peroxide by the transcriptional regulator, OxyR (67). isiR-2 sRNA is encoded adjacent to a LexA site, and its expression is induced by treating cells with the DNA-damaging agent, mitomycin C (68). RhlB sRNA is encoded adjacent to a Fur protein-binding site, and its expression is induced in response to iron depletion (69). OmrA (RygA), OmrB (RygB), and MicF sRNAs are regulated in part by the EnvZ-OmpR two-component system, which responds to changes in osmolarity (70). Apparently, many sRNAs are present in the cell in more than one length; for example, GadY was shown to exist simultaneously as 105, 90, and 59 nucleotide-RNA molecules (71). However, the details of sRNA biogenesis regarding active and inactive forms and the existence or extent of chemical modification are not well understood. Finally, the half-lives of sRNAs differ significantly from one to the next (61). As the protein factors that are known to play a role in the degradation of some sRNAs tend to be the same as those that are involved in the sRNA-mediated regulation of gene expression, they are described in the following section.

Regulation by sRNAs: RNA–RNA Interaction

Posttranscriptional regulation of gene expression can be achieved by base pairing interactions between sRNAs and their target mRNAs. It is important to note that the two sequences need not be perfectly complementary over a long patch; in fact, relatively short regions of complementarity interrupted by mismatches, small loops, or longer intervening sequences are the norm. The sRNA–mRNA binding event has been shown to result in a variety of outcomes, including mRNA accumulation, activation of translation, repression of translation, and mRNA degradation. Thus far, the involvement in these processes of three main proteins has been demonstrated, although the extent and nature of their roles cannot be generalized. First, the Sm-like host factor for Qp (Hfq) is an RNA-binding protein. It has been shown to have a high binding affinity for more than one third of the 70 known sRNAs, and to increase the half-lives and steady-state levels of certain sRNAs (72). Moreover, multiple studies have shown that it enhances interaction between sRNAs and their target mRNAs (73-75). Structural studies of Hfq have contributed to the development of a model in which the protein binds to relatively structureless A/U-rich segments of an RNA molecule that are in the vicinity of highly structured regions, and that the binding event has long-range effects on these secondary structures that allow intermolecular RNA–RNA interactions to take place (73, 76). Second, Rnase E is a hydrolytic endonuclease that cleaves single-stranded RNA (77). It has been shown to catalyze the degradation of certain sRNAs and their target mRNAs, some examples of which are given below. Whereas its N-terminal catalytic domain acts to cleave RNA, the well-conserved across species, the C-terminal domain of Rnase E is neither generally structured nor conserved, consisting rather of 15-40 amino acid stretches, referred to as microdomains, that bind to known protein partners like RhlB, enolase, and PNPase, components of the degradosome, but may also interact with other protein and RNA factors or enable self-assembly (78). For example, Hfq itself has been shown to interact with the C-terminal exonuclease (79). Third, Rnase III is a hydrolytic endonuclease that cleaves double-stranded RNA (80). It has also been shown to catalyze the degradation of certain sRNAs and their target mRNAs, some examples of which are given below. An sRNA can increase expression of its target gene, an outcome that appears to be relatively rare, by increasing the amount of target mRNA, increasing the efficiency with which it becomes translated into protein, or doing both. The cis-encoded antisense sRNA, GadY, was recently shown to increase the level of gadX mRNA in E. coli by means of perfectly complementary base pairing interactions involving the 3′ UTRs of both RNAs (71). The increase in gadX mRNA, which encodes a transcriptional
regulator of acid resistance genes, eventually results in an increase in the levels of two glutamate dehydrogenases, GadA and GadB. Examples of sRNAs that make use of a translational activation mechanism are also few; however, DsrA (85 nucleotides) and RpaA (105 nucleotides) are two sRNAs that act to positively and nonredundantly regulate the translation of rpsO mRNA into its product, the stationary phase sigma transcription factor, in E. coli (66, 83). In the absence of DsrA and RpaA, the intramolecular hairpin structure adopted by the unusually long (567 nucleotides) 5’ UTR of rpsO mRNA sequences its RBS, precluding translation initiation. When present, DsrA and RpaA form base pairs with the upstream region of the 5’ UTR of rpsO mRNA, competing with interactions it otherwise engages in with the downstream region of the 5’ UTR and liberating the RBS. Both DsrA-mediated and RpaA-mediated translational regulation of rpsO mRNA are Hfq-dependent (82).

In most cases described to date, base pairing of an sRNA to its target mRNA results in a negative effect on gene expression. By analogy to positive sRNA-mediated gene regulation, the RNA regulator can achieve its negative effect by decreasing the amount of target mRNA or, decreasing the efficiency with which it becomes translated into protein, or doing both. The strategy in which an sRNA represses translation by directly binding to the RBS of its target mRNA, in an Hfq-dependent manner, is highly represented in the regulation of expression of outer membrane proteins (Omp) in E. coli. ompC, ompF, andompA are selectively regulated in this way by MicC, MicF, and MicA sRNAs, respectively (83-86). Other sRNAs have been shown to bring about target mRNA cleavage by means of Hfq and RNase E. For example, RyhB sRNA leads to the degradation of sodB (encodes iron-containing superoxide dismutase) and sdiA (encodes iron-containing succinate dehydrogenase) mRNAs under iron-depleted conditions, and Sgrs sRNA leads to the degradation of ptsG mRNA, which encodes a major glucose transporter, under sugar phosphate stress (69, 87, 88). Two important aspects of this mechanism have recently come to light. First, the C-terminal domain of Rnase E interacts with Hfq, and a ribonucleoprotein complex composed of these two proteins and an sRNA is sufficient to bring about specific mRNA degradation (79). Second, although target cleavage often accompanies sRNA-mediated translational repression, making the effect irreversible, it may not be necessary to achieve the desired regulatory effect (89). In some cases, degradation of target mRNA in the presence of an sRNA has been shown to depend on Rnase III rather than Rnase E. For example, IsrR-1 sRNA and tisAB mRNA, which encodes a toxic peptide that is in-encoded with respect to its target, isrA, in the cyanobacterium, Synechocystis sp. PCC 6803 (90). The common feature shared by the IsrR-1/tisAB and IsrR/risA regulatory systems is a high degree of complementarity between the sRNA and its target mRNA, which likely enables the formation of a double-stranded RNA structure that is recognized as a substrate by RNase III.

### Regulation by sRNAs: RNA-Protein Interaction

Although the majority of chromosomally encoded regulatory sRNAs in bacteria, whose functions have been described, work by base pairing directly to an RNA target, there are also cases in which sRNAs have been shown to bind to a protein target to change the expression of genes at either the transcriptional or the posttranscriptional level. For example, 6S sRNA is known to impact global partitioning of transcription by directly interacting with a transcription factor. After becoming transcribed, together with an open reading frame as a dicistronic unit, and processed, to a 184 nucleotide sRNA, 6S goes on to tightly and specifically bind RpsD protein (91). RpsD, also known as σ70, is the vegetative transcription factor that directs RNA polymerase to transcribe the subset of genes that ought to be expressed under normal conditions. By folding into a long hairpin with a central internal loop that very much resembles an open promoter of DNA, 6S RNA is thought to compete with actual promoter DNA for the nucleic acid-binding site of RpsD, which in turn allows other regulators that use alternative sigma factors to become more expressed. A further testament to the versatility of sRNAs in bacteria is the paradigmatic CsrA protein/CsrB RNA system of posttranscriptional regulation. CsrA is a posttranscriptional regulator that binds to mRNAs, which code for a diverse set of proteins but share a conserved core sequence that the protein recognizes. By changing the stability or efficiency of translation of its targets, CsrA can have either positive or negative effects on their expression (92). CsrB and CsrC are sRNAs, 360 and 245 nucleotides in length, respectively, that bind CsrA by means of repeated hairpins that terminate in loops composed of the same conserved core sequence found in the mRNA targets of CsrA (93, 94). Whereas CsrB and CsrC comprise 18 and 9 such loops, respectively, the mRNA targets of CsrA have only one to three per molecule. In binding CsrA, CsrB and CsrC antagonize the numerous and diverse effects that this protein otherwise has as a posttranscriptional regulator.

### Large Noncoding RNAs and Mammalian X-Chromosome Inactivation

X-chromosome inactivation (XCI) is the process by which female mammalian cells, early in development, silence nearly all of the genes on all but one X-chromosome to achieve dosage parity between the sexes (95). Four distinct processes are associated with XCI: X-chromosome counting, chromosome choice, initiation and propagation of silencing, and maintenance of heterochromatin. All of these events are regulated by an 80-450-kb X-linked locus called the X-inactivation center (Xic). Interestingly, this locus can be described as having a death
The first large noncoding RNA gene of the Xic to be identified was Xist (98). In mice, it encodes a 17.4-kb untranslated RNA that is essential for XCI. Xist is expressed from the inactive X-chromosome and interacts physically with it, cis, via chromatin and the nuclear matrix, "coating" the chromosome. According to the current model, two conserved sequences in Xist, Repeat A and Repeat C, affect transcription from both the Xist promoter and promoters subject to XCI and recruit specific silencing proteins to the Xic, respectively. Likely protein partners in this process are enzymes that methylate or ubiquitylate chromatin. As Xist is transcribed, a process that could last up to 30 minutes, it spreads along the silenced chromosome and is thought to propagate these transcriptional and protein-recruiting effects as it goes. Xist is repressed by the antisense large noncoding RNA, Tsix (98). Before silencing, Tsix and Xist are both expressed, with Tsix being present in excess, from both the X chromosome to remain active and the X chromosome to be silenced. Once chromosomal choice has occurred, Xist is only expressed on the active X-chromosome until the process of XCI has ended. Tsix is encoded antisense to Xist, its transcription start site is downstream of Xist, and its function requires that it be transcribed to at least within the Xist gene. The persistence of Tsix is aided by another large noncoding RNA, Xite (X-inactivation intergenic transcription elements), which has multiple start sites clustered in two regions upstream of Tsix (98).

Although it is not yet known exactly how these and other large noncoding RNAs help to regulate XCI, and the related process of genomic imprinting, their mechanisms are likely to be complex. It is interesting to speculate that this particular physiological role is ideally suited to RNAs rather than proteins. That is, because protein-coding RNAs are transported to the cytoplasm for translation, their protein products are not usually traceable to the chromosome of origin of their corresponding transcript. Differently, noncoding RNAs remain in the nucleus to achieve the ends for which they evolved.

References


Further Reading

Noncoding RNAs


http://www.prl.msu.edu/PLANTncRNAs/

http://www.sanger.ac.uk/Software/Rfam/

http://research.imb.uq.edu.au/rnadb/

See Also

Micro RNA (miRNA), Chemistry of
Ribozymes Selected by Nature, Chemistry of
RNA, small nuclear (snRNA) and small nucleolar (snoRNA), Chemistry of
small interfering RNA (siRNA), chemistry of
mRNA Untranslated Regions (UTRs)
Molecules composed of RNA perform many roles that are vital for the life and reproduction of all cellular organisms and viruses, one of which is the role of messengers that carry from genes the information required for the production of protein. A typical feature of messenger RNA (mRNA) is its relative instability, which ensures that the production of any particular protein soon comes to an end when it is no longer required and synthesis of the message is downregulated. Moreover, while a transcript is being produced, its rate of decay is as important as its rate of synthesis in determining its cellular concentration and thus, to a large extent, its level of function. The turnover of mRNA, and indeed all other types of RNA, is mediated by multiple ribonucleases. The activity of these enzymes in cells can be inhibited rapidly by chemical treatments; thus, for many experimental systems their presence does not pose a significant barrier to the study of gene regulation at the RNA level. However, the development of RNA-based therapeutics such as small interfering RNAs and aptamers has required the introduction of modifications to prevent their degradation by ribonucleases within the cells and body fluids of humans. This article describes the properties of RNA that make it susceptible to ribonucleases and the techniques used to determine the stability of RNA molecules. It also provides examples of how RNA stability in living cells is being modulated for biotechnology and clinical applications, and it reviews the strategies that are being adopted to increase the in vivo stability of RNAs with therapeutic potential.

The information encoded in the genomes of all forms of life is transmitted via mRNAs (L. 2) to the ribosomes, which are the macromolecular machines that produce proteins through a process called translation (see the WECB review on translation: topics in chemical biology). The mRNAs in effect are copies of the genes, the units of heredity within the genetic blueprint, hence the initial description of these intermediates as DNA-like RNA (3). The process of copying the information from a gene into an RNA is called transcription; (4–6) (see also the WECB reviews on the RNA polymerase, chemistry of and the Transcription, Initiation of.). In general, genes are transcribed only when their products are required (see also the WECB reviews on transcription factors and transcription, activators and repressors of). Moreover, most mRNAs are turned over at a high rate relative to the time required to generate new cells. This instability ensures that protein production comes to an end soon after the transcription of the corresponding gene ceases. Moreover, while a transcript is being produced, its cellular concentration, which in large measure can determine its level of function (e.g., the amount of protein produced by the translation of mRNA), is critically dependent on its rate of decay as well as its rate of synthesis. Thus, the instability of RNA is central to the control of gene expression (Fig. 1). The turnover of mRNA, and indeed all other types of RNA, is mediated by exoribonucleases (7–9), which seem to function in close concert with endoribonucleases in all eubacteria studied to date (10–12) as well as chloroplasts (13).
RNA Stability, Chemistry, Measurement and Modulation of

The details of the nucleases and mechanisms that facilitate the decay and processing of intracellular RNA in bacteria (10-12, 14), organelles (13, 15), plants (16), as well as yeast and animals (17-20) are not covered here because excellent, up-to-date reviews are available (see also the WECB review on the mRNA localization and mRNA levels, control of).

Molecules composed of RNA are integral components also of the ribosome, whereas others are involved in the transfer of amino acids to growing polypeptide chains through a process that is mediated by the base pairing of the transfer RNAs (tRNAs) with codons in the mRNA. The ribosomal RNAs are bound by numerous proteins, whereas the transfer RNAs fold into compact tertiary structures (21, 22). Consequently, these RNAs in their mature form are less accessible than most mRNAs to nucleases and have longer half-lives (23, 24). The latter makes biological sense because it would be wasteful to turnover rapidly RNA components of an apparatus that is essential for growth no matter the physiology of the organism.

Determining the Rates of RNA Turnover

The value that is used universally to record and compare the rate of decay of cellular RNA species is the “half-life,” the time required for the initial recorded level of an RNA to be reduced by half. The typical half-lives of mRNA in the bacterium Escherichia coli tend to be on the order of a few minutes (25, 26), whereas in the yeast Saccharomyces cerevisiae it is not uncommon for half-lives to be measured in hours (27). In plants and mammalian cells, the values of half-life range from less than 1 hour to days, with the average being several hours (28-30).

Inhibiting transcription

In the study of eubacterial RNAs, half-life values are obtained most commonly by measuring RNA levels after blocking transcription via the addition of rifampicin, a semisynthetic antibiotic that binds and inhibits the endogenous RNA polymerase 

After blocking transcription, RNA samples are collected at different time points and immediately frozen in liquid nitrogen (39) or mixed with an agent that rapidly permeates cells and inactivates enzymes, thereby stopping RNA turnover. Traditionally, phenol has been used as a chemical-halting agent, but now available is a proprietary agent called RNA later (Ambion, Inc., Austin, TX), which is aqueous and nontoxic and stabilizes RNA in a range of bacterial species as well as eukaryotic cells and tissues (40). Cells then are disrupted by mechanical, chemical, or enzymatic means, and the RNA is isolated; the precise procedures depend on the source of the biological material (41, 42).

Determining the relative levels of transcripts at different time points

The abundance of particular RNAs in samples taken at different time points can be determined by using relatively straightforward techniques, such as Northern blotting, nuclease protection, and primer extension assays (Fig. 2). A laboratory manual that describes these techniques in detail is available (43); thus, only a brief overview is provided here. Northern blotting (Fig. 2A) involves separating RNA molecules on the basis of their size by using denaturing gel electrophoresis, blotting the RNA to a membrane support, and probing for a specific transcript using a complementary nucleic acid that is labeled. In nuclease protection assays (Fig. 2B), a labeled, single-stranded nucleic acid that is complementary to a specific transcript is mixed with a sample of RNA and the two are allowed time to hybridize before the excess (unprotected) probe is removed by incubating it with a nuclease such as S1 that can cut single-stranded, but not double-stranded, nucleic acid. The amount of RNA in the sample then can be inferred from the amount of labeled probe that remains. Primer extension assays (Fig. 2C) involve incubating an RNA sample with a labeled oligonucleotide that is complementary to a particular transcript. When annealed to its target, the oligonucleotide serves as a primer that can be extended by an RNA-dependent DNA polymerase called reverse transcriptase to produce complementary DNA (cDNA). The amount of extended cDNA product that is synthesized is proportional to the amount of transcript in the original sample. In addition, with the recent development of polymerase chain reaction (PCR) machines that incorporate fluorescence detection, it is possible to determine quantitatively the levels of transcripts in different samples by using reverse-transcription PCR (44). As described for primer extension analysis, the first step in this approach is the synthesis of cDNA. This product is amplified then by PCR in the presence of a gene-specific fluorescent probe or a dye, such as SYBR Green I (an asymmetrical cyanine derivative), that fluoresces strongly when it binds to double-stranded DNA (45). By measuring the level of fluorescence at the end of each
RNA Stability, Chemistry, Measurement and Modulation of

Figure 1 The cycle of gene expression. Nucleotides are joined to form mRNA by a polymerase that uses a protein-coding gene as the template. This process is called transcription. The RNA polymerases initiate transcription at sites in genomes called promoters. In eubacteria, the primary product of transcription of most genes can be used as a message by ribosomes to synthesize polypeptides from amino acids (in the form of charged transfer RNAs). This process is called translation. However, in eukaryotes, modification of the primary transcript is required. Modifications include capping the 5′ end with 7-methyl guanosine via a 5′-to-5′ link, polyadenylating the 3′ end, and splicing out segments (introns) that separated protein-coding regions (exons). This step is called RNA processing. The mRNA has a relatively short lifetime and is converted to nucleotides by the combined actions of multiple ribonucleases. This process is called RNA degradation or decay.

cycle (i.e., in “real-time”), the kinetics of amplification can be determined, from which the amount of transcript in a sample can be determined (see also the WECB review on XXXX) (44).

Transcriptional pulse–chase experiments

The individual steps in the decay of a particular transcript can be followed by engineering the corresponding gene to place it under the control of an externally regulatable promoter that can be used to switch on and off transcription within minutes (46). Some of the most commonly used regulatable systems are based on elements derived from the tetracycline-resistance (tet) operon of Escherichia coli. Systems have been engineered so that transcription can be switched on or off rapidly in a reversible manner by adding tetacycline or a derivative, such as doxycycline. These systems are being used widely to study both eubacteria and eukaryotes (47). The creation of a pulse of transcription means that the bulk of the population of a particular transcript enters the decay pathway within a narrow time window. This approach can synchronize the process sufficiently for the order of steps in the pathway to be determined (46). To date, transcriptional pulse–chase experiments have been used mainly to study mRNA decay in S. cerevisiae and mammalian cells in culture. However, in principle they could be used to study the decay of RNA in any system, provided a transcriptional pulse can be created that is short relative to the lifetime of the RNA being studied. Externally regulatable promoters have been used also to switch off transcription of a specific gene so that the half-life of its mRNA can be determined (34). A n advantage of this technique is that should the decay require a factor(s) (RNA or protein) that is relatively short-lived, the continuing transcription of all other genes in the background will ensure that the abundance of any such factor does not diminish and become rate limiting. It is also possible to estimate the rate of decay of a particular transcript in a background of continuous transcription within organisms for which an easily regulatable promoter system is not yet available. This estimation can be achieved by measuring the “approach to steady-state labeling,” which involves monitoring the incorporation of 32P-labeled phosphate into specific transcripts (34). Although this method is rather cumbersome, the values obtained are comparable with those derived by other approaches (33).

The stability of transcripts on a genome-wide scale

More recently, with the advent of gene arrays, it has been possible to measure the rates of RNA decay on a genome-wide scale. This technology involves labeling an RNA population and then hybridizing it against a library of thousands of immobilized spots of DNA, each of which serves as a specific probe for a particular transcript (48–50). The amount of label that associates
with each spot is proportional to the amount of the corresponding transcript in the original RNA sample (see also the WECB reviews on Transcript Profiling, Tools for and Array-Based Tools for Nucleic Acids). Interestingly, this approach is starting to reveal new relationships between the functions encoded by mRNAs and their stability. For example, it seems that the shortest-lived mRNAs in E. coli tend to correspond to genes that as of yet have not been characterized (25, 26). Interestingly, genes with some of the shortest-lived mRNAs in eukaryotes have known functions. These genes include the c-myc and c-fos transcription factor genes in mammalian cells (51), several that are inducible by auxin in plants (52, 53), and many that are associated with the mating switch in yeast (54). Therefore, by analogy with eukaryotic systems, many uncharacterized E. coli genes with unstable mRNAs may be regulators. As E. coli has been characterized extensively in the laboratory, it is likely that the processes controlled by genes with unstable mRNAs are required only under some particular environmental condition (26). The decay rates of eukaryotic mRNAs also are being analyzed on a genome-wide scale. For example, this approach has revealed that similar to auxin-inducible genes (52, 53), genes responsive to mechanical stimulation and the circadian clock are rapidly degraded (37). The study of S. cerevisiae has revealed that, similar to what has been found for E. coli, genes involved in central metabolism tend to have longer half-lives (27).

Modulating mRNA Turnover

The measurement of the half-lives of individual transcripts has revealed that for all organisms studied, the values for different transcripts can vary over an order of magnitude. This finding has led to many investigations of the cis-acting elements within RNA and trans-acting factors that determine the rate of RNA decay (10–13, 15–20) (see also the WECB review on the mRNA localization and mRNA levels, control of). Moreover, the knowledge gained is being used to modulate gene expression. Extensive coverage of this area is beyond the scope of this article; however, the potential to manipulate gene expression at the level of RNA stability is illustrated below by using two different examples.

Enhanced protein production in bacteria

E. coli has long served as a host for the production of proteins (see the WECB review on protein expression, systems for ). Recently, it has been manipulated genetically to stabilized transcripts that are synthesized as part of the commonly used pET expression system. In this system, cloned genes are transcribed from a promoter that is recognized by the RNA polymerase of bacteriophage T7 (55). However, although the cloned gene is transcribed efficiently, the transcripts that are produced by T7 RNA polymerase can be highly unstable. This instability occurs because the bacteriophage enzyme transcripts genes more rapidly than the endogenous RNA polymerase of E. coli. As a
consequence, ribosomes cannot keep pace (in eubacteria, transcription and translation are coupled) and endoribonucleases gain access to "naked" stretches of RNA (56, 57). To counter this effect, strains of E. coli that host the pET system now are offered with a mutation that reduces the rate at which transcripts are cleaved by RNase E (58), an endoribonuclease that initiates much of the mRNA decay in E. coli (59). It also has been shown that mRNA produced using a T7 promoter can be stabilized by using a mutant polymerase that transcribes DNA at a slower rate (60).

The modulation of mRNA decay by a small molecule

A substantial proportion of the genetic and acquired diseases of humans stems from nonsense mutations, which create codons that signal the premature termination of translation. Normally the mRNA of genes that contain such codons is degraded rapidly by a "quality control" mechanism, the details of which are covered in several excellent reviews (17, 61-63). Recently, a small, orally available molecule (called PTC124) has been identified that reduces the fidelity of the ribosome such that it inserts an amino acid when it encounters a premature termination codon (PTC), but not a bona fide translational stop codon (64). This action suppresses rapid mRNA decay, which thereby permits the production of full-length polypeptides. Moreover, at least some molecules so produced have activity as PTC124 has been shown through two Phase 2 clinical trials to have efficacy in the treatment of patients with forms of Cystic Fibrosis (CF) or Duchenne Muscular Dystrophy (DMD) caused by nonsense mutations (65). However, further research is required to study the long-term effects of suppressing nonsense-mediated mRNA decay (NMD). In this context, it is perhaps worth noting that very recent studies indicate that NMD has a role in controlling the expression of a subset of normal (i.e., unmutated) genes (66, 67). Thus, although suppressors of NMD will likely have side effects, it is hoped that these side effects will be outweighed by alleviating the effects of severe disease in those individuals where the root cause is a nonsense mutation. For details of other strategies for the treatment of PTC diseases, readers are referred to a comprehensive review of this topic (68).

Small, Untranslated RNAs and RNA-Based Therapeutics

Nonsense-mediated decay is just one of many examples of gene regulation that occurs at a step after the initiation of transcription. In bacteria and eukaryotes, classes of small, untranslated RNAs (sRNAs) have been identified that regulate the translation of their targets through a physical interaction that involves intermolecular base pairing (see also the WECB review on RNA, noncoding). In such cases, the sRNA is described as being antisense (i.e., its sequence is the reverse complement of its target).

Antisense RNA and interference

In eubacteria, antisense RNAs originate from genetic loci that either overlap or are distant from the genes of their targets (14, 69, 70). The former, cis-encoded antisense RNAs are found associated with ancillary mobile genetic elements, such as viruses, and are synthesized always from a segment of the DNA strand opposite that used for synthesis of the mRNA target. In contrast, the trans-encoded antisense RNAs display only partial complementarity to their targets, can affect multiple targets, and seem to regulate core cellular functions in response to suboptimal or stressful growth conditions (14, 69, 70). Antisense RNAs that target mRNA targets have been shown to affect translation, which in turn can influence the rate of mRNA turnover (71). It is known that reducing ribosome coverage can provide ribonucleases with increased access to susceptible sites (56, 72). Plants, fungi, and animals have a class of sRNAs called micro RNAs (miRNAs) that are analogous to bacterial antisense RNAs insofar as intermolecular base pairing alters the translation, decay, or both of mRNA targets (73-75) [see also the WECB review on the MicroRNA (miRNA)]. However, unlike bacterial antisense RNAs, the actual RNA that targets the mRNA, the "guide" RNA, is generated via a double-stranded (ds) intermediate of 20-25 bp. Importantly, it has been found that short, chemically synthesized ds RNA, when introduced to plants and animals, can mimic effectively the short double-stranded intermediates of miRNAs and consequently interfere dramatically and selectively with the expression of genes that have a complementary sequence (76, 77). These short, interfering RNAs (siRNA), which can be synthesized readily by using phosphoramidite chemistry (78, 79) (see the WEBC review on the solid-phase synthesis of biomolecules), have revolutionized the study of eukaryotic gene function both in cell culture and in living organisms (80-82) and, as described below, have potential as therapeutic agents for the treatment of human disease (83-87) [see also the three WECB reviews on RNA Interference to Treat Human Diseases, Applications of; gene silencing techniques and the small interfering RNA (siRNA); chemistry of]. The machinery that mediates RNA interference in eukaryotes does not exist in bacteria, and the introduction of RNAi to prokaryotes exogenously does not provide a routine means of altering the expression of a targeted gene; however, it should be noted that the production of antisense RNA intracellularly has been shown to induce gene silencing in many bacteria and that libraries of constructs that produce such RNAi have been used (although not extensively) to study gene function on a genome-wide scale (88, 89).

Additional classes of sRNA

Antisense RNAs are not the only class of sRNAs that have key cellular and regulatory functions. In eubacteria, for example, an RNA called 6S RNA binds directly to the form of RNA polymerase that transcribes the general housekeeping genes; this interaction causes a reduction in the expression of these genes as part of a wider scheme that redirects the activity of the transcriptional machinery during periods of nutrient starvation (90, 91). Non-antisense sRNAs also target steps in gene expression other than transcription (92). When the progress of
a ribosome is arrested on a bacterial mRNA because of truncation of the coding region, an sRNA can enter the ribosome in a manner analogous to a normal tRNA, stimulate the release of the ribosome, and complete the cycle of translation by serving as a template that encodes a short peptide. The peptide added by the “trans-translation” of the aptly named transfer-messenger RNA (tmRNA) marks the improperly terminated protein as substrates for rapid proteolytic decay (92). Within the natural world, the functions of RNA extend to catalysis, for example, peptide bond formation in the ribosome (93). (see also the WE CB review on ribosome, chemistry of), the ribonuclease activity of RNase P, and splicing reactions (i.e., the cutting and ligation of RNA) are catalyzed by RNAs (94, 95) (see also the WE CB reviews on the Locked Nucleic Acids and Pre-mRNA splicing (96)).

The ability of RNA to bind other types of molecules with high affinity and specificity and to catalyze reactions is rooted in the complex 3-D structures it can form (96, 97). Although all RNAs are synthesized as a single-stranded transcript, stretches of complementary bases can pair to form a series of short, double-stranded helices that can generate several secondary structures. Examples include stem-loops, which are formed by transcripts looping back on themselves to form double helices (stems) at the base of the loops, and pseudoknots, where molecules loop back on themselves, but the looped segments are sufficiently long to contain secondary structures of their own. Such secondary structures are packed together with the assistance of metal ions to form higher-order, functional structures (98-100) (see also the WE CB reviews on crystallographic techniques: nucleic acids and NMR Tools for Nucleic Acids). Moreover, the strength of the interactions that form the secondary structure seems central to the thermostability of functional RNAs; strong correlations exist between both the GC-content and the predicted melting temperatures of RNA secondary structures and the optimal growth temperatures of the organisms in which they are found (101–103). In thermophiles, the functional conformation of RNAs may be stabilized also by nucleotide modifications, stronger tertiary interaction, macro-molecular crowding, and association with proteins.

Aptamers

The presentation of the four nucleotides of RNA within the context of a complex 3-D landscape permits RNA molecules to bind with specificity and high affinity to targets that range from small molecules to cells. This remarkable ability has been and continues to be demonstrated through an in vitro selection approach called the “systematic evolution of ligands by exponential enrichment” (SELEX) (104, 105). The details of the SELEX procedure are described in accompanying reviews (see WE CB reviews on the In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes) and aptamer-based therapeutics), so only a brief overview is provided here. The process begins with the synthesis of a DNA template that can be transcribed by T7 polymerase to produce single-stranded RNA. The template contains a central region typically of 20-40bp that is randomized at each position. Therefore, when the template is transcribed, it produces a population of RNA molecules each with a different sequence and thus a different structure at its core about 1015 different sequences can be produced from about 2 nmoles of template, the amount used in a typical transcription reaction. RNA molecules that bind a target are selected, a complementary DNA (cDNA) strand is produced by reverse transcription, and then large amounts of the template are regenerated by using the polymerase chain reaction (see WE CB review on the Polymerase Chain Reaction (PCR) in Chemical Biology). The process then is repeated until RNA molecules with the desired affinity and specificity are obtained. Finally, after using cloning techniques to isolate individual templates from the successful pool, the RNAs can be sequenced, characterized, and synthesized chemically using phosphoramidite chemistry. The products of selections are called “aptamers” (Latin aptus means “fitted”) and have found a wide variety of uses as biosensors (106, 107), probes for drug target validation (108), diagnostic indicators, and therapeutics (109-114) (see also the two WE CB reviews on the In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes) and on Combinatorial Libraries: Overview of Applications in Chemical Biology). Recently, Macugen (also called Pegaptanib), the first aptamer-based therapeutic agent, has entered clinical use for the treatment of age-related macular degeneration (AMD), which in the developed world is the leading cause of blindness in individuals over 50 years in age (115). Other aptamer-based agents are entering clinical trials (109-113) (see also the WE CB review on aptamer-based therapeutics).

Stabilizing RNA-Based Therapeutics

The body fluids of humans contain an array of ribonucleases (116-118); however, RNA molecules can be modified to greatly increase their stability in vivo. Modified RNAs differ from modified DNA in that body fluids of humans and other mammals are members of a large family of RNase A, a relatively small protein that was one of the earliest models for studying protein structure and function (119-121).

RNA Modifications

RNase A and its relatives cleave RNA molecules via a hydrolysis reaction that requires a hydroxyl at the 2’ position on the ribose of the nucleotide immediately on the 5’ side of a phosphodiester bond (Fig. 3). The 2’ OH is necessary for the transfer of an electron between two catalytic histidines (Fig. 3A) in order to form the transition state (122). Thus, protection can be provided by replacing the OH group at the 2’ position with a H, O-methyl, O-methoxymethyl, NH2 (aminoo), or F (fluoro) group, all of which cannot participate in the electron transfer (Fig. 3B). Protection also can be provided by ribose moieties that have an extra covalent “bridge” between the 2’ and 4’ positions (122). Not only does this bridge prevent electron transfer between the catalytic histidines, but also it “locks” the sugar in a conformation (C3’-endo pucker) that increases the affinity of the interaction between the modified RNA and its complementary target (123). Complementary base pairing is promoted similarly by 2 F substitutions. Replacement of oxygen with sulfur at
the 4’ position of the ribose and borophosphonate backbone (P = B) modifications also have been reported to confer increased resistance to nucleases (124). All modifications described above (Fig. 3B) can be introduced into phosphorodiester nucleosides, the building blocks for the solid-phase synthesis of nucleic acids (78, 79); indeed, most, if not all, are available currently from several commercial sources (e.g., Glen Research, Sterling, VA) (see also the WECB review on the solid-phase synthesis of biomolecules). The 2’ modifications also protect the 3’-5’ phosphodiester bonds of RNA against nonenzymatic alkaline hydrolysis, which requires a hydroxyl group at the 2’ position (125). This reaction is enhanced by magnesium, which stabilizes a cyclic 2’-3’ phosphate intermediate. Thus, unmodified RNAs should be stored at low temperatures under slightly alkaline conditions in the presence of a divalent-cation chelator, such as ethylenediaminetetraacetic acid (EDTA) or sodium citrate. The absence of 2’ OH groups in RNA explains the chemical stability of this macromolecule relative to RNA.

Assessing nuclease resistance

Usually the nuclease resistance of RNA is assessed first by using it as a source of nucleases body fluids such as urine (126), blood serum (127), or plasma (128); however, it is not uncommon for the source to be an extract of a tissue, such as pancreas (129), or purified nucleases (130). The range of methods that can be used to detect the continuing presence of synthetic RNAs is dependent on their length and mode of synthesis (see also the WECB review on labeling techniques: nucleic acids). Short oligonucleotides can be labeled readily during their solid-phase synthesis via a phosphoramidite that incorporates a fluorophore (131) (see also the WECB review on fluorescence spectroscopy: overview of applications in chemical biology) or post synthesis by adding the radioactive 32P to the 5’ end by using the enzyme polynucletidyl kinase (43). The latter also can be used to label RNA synthesized from a DNA template by using an RNA polymerase, provided the triphosphate group at the 5’ end of the synthesized RNA is removed first using a phosphatase (43).

Enzymatically synthesized RNA also can be labeled internally via the incorporation of fluorocarbon labeled nucleotides (131). Following incubation with a source of nucleases, the integrity of labeled RNAs then can be assessed by electrophoresing samples in a gel, most commonly for this purpose made of acrylamide, in a gel, most commonly for this purpose made of acrylamide, and visualizing the labeled species using the corresponding labeled RNAs then can be assessed by electrophoresing samples in a gel, most commonly for this purpose made of acrylamide, and visualizing the labeled species using the corresponding fluorophore. Specificity is achieved by careful calibration of the HPLC systems to determine the elution point of the RNA-based therapeutic. The stability of longer RNAs (L > 100 nt), which tend to be synthesized enzymatically, can be determined without labeling by using reverse transcription PCR (44). An amplicon will be produced no longer from an RNA once cleavage has occurred between or within either of the two segments recognized by the primers. This method is preferable to Northern blotting, which is less sensitive and more cumbersome to perform. The half-life of RNA in the body fluids of living animals can be determined by assaying samples taken at different time points after administration; however, the values obtained reflect not only degradation but also the rates of renal clearance and uptake into cells and tissue. In the development of M acugen, an HPLC-based assay was used to determine the plasma half-lives of aptamers in animal models (133).

Compatibility of stabilizing modifications with RNA interference

In the context of developing siRNAs for therapeutic applications, the RNA must resist intracellular nucleases as well as interact with cellular machinery to produce a complex that can silence (or at least dampen) the expression of the target gene(s). Fortunately, the effects of RNA-stabilizing modifications on silencing can be assessed relatively easily by targeting in cultured cells the mRNA of genes that encode readily assayable products; for example, the green fluorescent protein (129, 134). Through this type of study, it has been demonstrated that the resistance of siRNA to nucleases can be increased significantly without compromising silencing efficiency, at least ex vivo. Resistance has been achieved, for example, by modifying the RNA such that multiple nucleotides at positions in both strands have a 2’ fluoro group (130), particular positions have a 2’-O-methyl group (135), or nucleotides on both strands alternate between having a 2’ fluor and a 2’-O-methyl group (136, 137), or nucleotides of the sense strand have 2’-O-methoxyethyl (137) or 4’-thio modifications (138). It is likely that this list is far from exhaustive, given the number of potential modifications (e.g., the effect of the 2’-amino modification has yet to be studied within the context of an siRNA) and the multiplicity of positions within siRNAs that can be modified (79, 124, 139, 140). Recently, it has been found that the extent to which modified RNAs can bind the RNA-induced silencing complex (RISC) is a predictor of their potency in vivo (141). Whether this biochemical assay will be sufficient to replace cell-based assays remains to be determined at this time.

The final choice of stabilizing modifications is likely to be influenced by additional studies that will determine the influence of RNA modifications on renal clearance and cellular uptake and whether issues relating to cytotoxicity exist (124, 140). With regard to the latter, at least one study has indicated that siRNAs with phosphorothioate backbone modifications can have cytotoxic properties (142). That living organisms cannot metabolize all unnatural molecules without consequence on growth should not be surprising. Small interfering RNAs should not need modification to provide resistance to intracellular nucleases because the RNAs they mimic function effectively within cells.

The production of nuclease-resistant aptamers

In considering the therapeutic application of aptamers produced using SELEX, the contribution of the 2’ OH group to the 3-D structure of RNA needs to be taken into account. This group contributes to the physico-chemical properties of RNA, which in
RNA Stability, Chemistry, Measurement and Modulation of

Figure 3 Hydrolysis of RNA by members of the RNase A family. The mechanism is shown in Panel A. Modifications that stabilize RNA are shown in Panel B.

Turn influences the thermal stability of secondary structures, and contributes to bonding that can determine the arrangement of secondary structures in 3-D space (99). Thus, the introduction of modifications into an aptamer post selection can alter its interaction with the target. As such, aptamers that have increased resistance to nucleases usually are produced using “front-loaded” SELEX, in which the libraries are synthesized using nucleotides with modifications, such as 2′fluoro, O-methyl, or amino groups (111, 113). Indeed, Macugen was developed from a library made from pyrimidines with 2′fluoro modifications. Interestingly, the aptamers selected from this pool bound their target, which is an isoform of vascular endothelial growth factor (VEGF), with higher affinity than those derived from a library of equal complexity synthesized by using pyrimidines with 2′amino modifications (143). This occurrence may have been because a higher proportion of the 2′fluoro-modified aptamer population was in a conformation(s) that can interact with the target because of double-stranded segment(s) having greater thermal stability. Although front-loaded SELEX is used routinely as part of strategies to develop nuclease-resistant aptamers, once the minimal sequence required for binding to the target is defined, the RNA can be synthesized using solid-phase chemistry and additional modifications can be introduced. The introduction of 2′O-methyl substituted purines into aptamers raised against VEGF from the library made using pyrimidines with 2′F modifications increased the half-life by around eightfold (143). This result is not particularly surprising; although cleavage by many ribonucleases found in body fluids require a pyrimidine with a 2′OH immediately on the 5′side of a susceptible bond, some can cleave bonds that have an unmodified purine on the 5′side (116–118). 2′amino-modified or 2′fluoro-modified pyrimidines can be incorporated into RNAs using wild-type T7 RNA polymerase; however, the incorporation of 2′0-methyl pyrimidines required modification of the enzyme using directed evolution.
described recently for the chemical synthesis of 4-thio-modified nucleotides; moreover, these schemes are incorporated into RNA by wild-type T7 polymerase with an acceptable efficiency (145). From the above, it is clear that as described for the synthesis of siRNAs, several modifications can be introduced into aptamers to decrease their susceptibility to degradation by nucleases. RNA-stabilizing modifications can be introduced also into ribozymes and RNA-based antisense oligonucleotides (ASOs) that inhibit or alter gene expression independently of the RNA interference pathway (146-149) (see the WECB review on the oligonucleotide-directed inhibition of gene expression).

Stability also can be achieved by using aptamers that have nucleotides based on L- instead of D-ribose (150-152). Such aptamers are generated by a process that first selects aptamers composed of the naturally occurring sugar against an epitope (i.e., a complete mirror image of the intended target). Once an aptamer(s) with the required function is obtained, its L-ribose-based epitope, which is also called a Spiegelmer (from the German word Spiegel meaning “mirror”), is synthesized chemically. The RNA-Spiegelmer will be completely resistant to nucleases and will bind the intended target with the same characteristics that its D-ribose-based progenitor bound the epitope in the initial selection. The only limitation to this approach will be the ingenuity of chemists to synthesize the enantiomers of targets of interest. Schemes have been described already for the chemical synthesis of peptides composed of D-rather than L-amino acids (see WECB review on the Natural and Unnatural Amino Acids, Synthesis of).

Future Perspectives

In all likelihood molecules based on RNA will be used increasingly as diagnostic tools and therapeutic agents. As this article illustrates, the inherent instability of RNA can be overcome effectively via the introduction of an increasing number of modifications. Concomitantly, renal clearance has been reduced sufficiently to increase bioavailability via the encapsulation of RNA within lipoproteins or by attachment to the 5′ or 3′ end of bulky groups such as polyethylene glycol, cholesterol, or biotin–streptavidin (111, 113, 114, 124). Moreover, promising developments exist in the area of delivery to specific cell-surface receptors can be isolated, the approach of using a chimera was processed effectively by the silencing machinery. Binding to the receptor induced internalization, whereupon the siRNA as a chimera with an aptamer that binds the receptor.

A particular receptor. This delivery was achieved by producing the siRNA-resistant to nucleases and will bind the intended target with the same characteristics that its D-ribose-based progenitor bound the epitope in the initial selection. The only limitation to this approach will be the ingenuity of chemists to synthesize the enantiomers of targets of interest. Schemes have been described already for the chemical synthesis of peptides composed of D-rather than L-amino acids (see WECB review on the Natural and Unnatural Amino Acids, Synthesis of).

References


126. Green LS, Jellinek D, Bell C, Beebe LA, Feistner BD, Gill SC, Lipkin D, Talbert PT, Cohn M. The mechanism of the al-
125. Manoharan M. RNA interference and chemically modified small
124. Jepsen JS, Sorensen MD, Wengel J. Locked nucleic acid: a
123. Frieden M, Hansen HF, Koch T. Nuclease stability of LNA
122. Beintema JJ, Kleineidam RG. The ribonuclease A superfamily:
120. Ng EWM, Shima DT, Calias P, Cunningham ET, Guyer DR,
119. Lee JF, Stovall GM, Ellington AD. Aptamer therapeutics ad-
117. Nimjee SM, Rusconi CP, Sullenger BA. Aptamers: an emerging
116. Sorrentino S. Human extracellular ribonucleases: multiplicity,
114. Capodici J, Kariko K, Weissman D. Inhibition of HIV-1 infec-
111. Capodici J, Elbashir SM, Vandenburgh K, Manninga H, Scaringe
110. Nimjee SM, Rusconi CP, Sullenger BA. Aptamers: an emerging
108. Nimjee SM, Rusconi CP, Sullenger BA. Aptamers: an emerging
107. Tombelli S, Minunni A, Mascini A. Analytical applications of
106. Navani NK, Li YF. Nucleic acid aptamers and enzymes as
105. Gold L, Brown D, He YY, Shtatland T, Singer BS, Wu Y. From
104. Wilson DS, Szostak JW. In vitro selection of functional nucleic
103. Juneau K, Cech TR. In vitro selection of RNAs with increased
102. Juneau K, Cech TR. In vitro selection of RNAs with increased
101. Juneau K, Cech TR. In vitro selection of RNAs with increased
100. Juneau K, Cech TR. In vitro selection of RNAs with increased
99. Juneau K, Cech TR. In vitro selection of RNAs with increased
98. Juneau K, Cech TR. In vitro selection of RNAs with increased
97. Juneau K, Cech TR. In vitro selection of RNAs with increased
96. Juneau K, Cech TR. In vitro selection of RNAs with increased
95. Juneau K, Cech TR. In vitro selection of RNAs with increased
94. Juneau K, Cech TR. In vitro selection of RNAs with increased
93. Juneau K, Cech TR. In vitro selection of RNAs with increased
92. Juneau K, Cech TR. In vitro selection of RNAs with increased
91. Juneau K, Cech TR. In vitro selection of RNAs with increased
90. Juneau K, Cech TR. In vitro selection of RNAs with increased
89. Juneau K, Cech TR. In vitro selection of RNAs with increased
88. Juneau K, Cech TR. In vitro selection of RNAs with increased
87. Juneau K, Cech TR. In vitro selection of RNAs with increased
86. Juneau K, Cech TR. In vitro selection of RNAs with increased
85. Juneau K, Cech TR. In vitro selection of RNAs with increased
84. Juneau K, Cech TR. In vitro selection of RNAs with increased
83. Juneau K, Cech TR. In vitro selection of RNAs with increased
82. Juneau K, Cech TR. In vitro selection of RNAs with increased
81. Juneau K, Cech TR. In vitro selection of RNAs with increased
80. Juneau K, Cech TR. In vitro selection of RNAs with increased
79. Juneau K, Cech TR. In vitro selection of RNAs with increased
78. Juneau K, Cech TR. In vitro selection of RNAs with increased
77. Juneau K, Cech TR. In vitro selection of RNAs with increased
76. Juneau K, Cech TR. In vitro selection of RNAs with increased
75. Juneau K, Cech TR. In vitro selection of RNAs with increased
74. Juneau K, Cech TR. In vitro selection of RNAs with increased
73. Juneau K, Cech TR. In vitro selection of RNAs with increased
72. Juneau K, Cech TR. In vitro selection of RNAs with increased
71. Juneau K, Cech TR. In vitro selection of RNAs with increased
70. Juneau K, Cech TR. In vitro selection of RNAs with increased
69. Juneau K, Cech TR. In vitro selection of RNAs with increased
68. Juneau K, Cech TR. In vitro selection of RNAs with increased
67. Juneau K, Cech TR. In vitro selection of RNAs with increased
66. Juneau K, Cech TR. In vitro selection of RNAs with increased
65. Juneau K, Cech TR. In vitro selection of RNAs with increased
64. Juneau K, Cech TR. In vitro selection of RNAs with increased
63. Juneau K, Cech TR. In vitro selection of RNAs with increased
62. Juneau K, Cech TR. In vitro selection of RNAs with increased
61. Juneau K, Cech TR. In vitro selection of RNAs with increased
60. Juneau K, Cech TR. In vitro selection of RNAs with increased
59. Juneau K, Cech TR. In vitro selection of RNAs with increased
58. Juneau K, Cech TR. In vitro selection of RNAs with increased
57. Juneau K, Cech TR. In vitro selection of RNAs with increased
56. Juneau K, Cech TR. In vitro selection of RNAs with increased
55. Juneau K, Cech TR. In vitro selection of RNAs with increased
54. Juneau K, Cech TR. In vitro selection of RNAs with increased
53. Juneau K, Cech TR. In vitro selection of RNAs with increased
52. Juneau K, Cech TR. In vitro selection of RNAs with increased
51. Juneau K, Cech TR. In vitro selection of RNAs with increased
50. Juneau K, Cech TR. In vitro selection of RNAs with increased
49. Juneau K, Cech TR. In vitro selection of RNAs with increased
48. Juneau K, Cech TR. In vitro selection of RNAs with increased
47. Juneau K, Cech TR. In vitro selection of RNAs with increased
46. Juneau K, Cech TR. In vitro selection of RNAs with increased
45. Juneau K, Cech TR. In vitro selection of RNAs with increased
44. Juneau K, Cech TR. In vitro selection of RNAs with increased
43. Juneau K, Cech TR. In vitro selection of RNAs with increased
42. Juneau K, Cech TR. In vitro selection of RNAs with increased
41. Juneau K, Cech TR. In vitro selection of RNAs with increased
40. Juneau K, Cech TR. In vitro selection of RNAs with increased
39. Juneau K, Cech TR. In vitro selection of RNAs with increased
38. Juneau K, Cech TR. In vitro selection of RNAs with increased
37. Juneau K, Cech TR. In vitro selection of RNAs with increased
36. Juneau K, Cech TR. In vitro selection of RNAs with increased
35. Juneau K, Cech TR. In vitro selection of RNAs with increased
34. Juneau K, Cech TR. In vitro selection of RNAs with increased
33. Juneau K, Cech TR. In vitro selection of RNAs with increased
32. Juneau K, Cech TR. In vitro selection of RNAs with increased
31. Juneau K, Cech TR. In vitro selection of RNAs with increased
30. Juneau K, Cech TR. In vitro selection of RNAs with increased
29. Juneau K, Cech TR. In vitro selection of RNAs with increased
28. Juneau K, Cech TR. In vitro selection of RNAs with increased
27. Juneau K, Cech TR. In vitro selection of RNAs with increased
26. Juneau K, Cech TR. In vitro selection of RNAs with increased
25. Juneau K, Cech TR. In vitro selection of RNAs with increased
24. Juneau K, Cech TR. In vitro selection of RNAs with increased
23. Juneau K, Cech TR. In vitro selection of RNAs with increased
22. Juneau K, Cech TR. In vitro selection of RNAs with increased
21. Juneau K, Cech TR. In vitro selection of RNAs with increased
20. Juneau K, Cech TR. In vitro selection of RNAs with increased
19. Juneau K, Cech TR. In vitro selection of RNAs with increased
18. Juneau K, Cech TR. In vitro selection of RNAs with increased
17. Juneau K, Cech TR. In vitro selection of RNAs with increased
16. Juneau K, Cech TR. In vitro selection of RNAs with increased
15. Juneau K, Cech TR. In vitro selection of RNAs with increased
14. Juneau K, Cech TR. In vitro selection of RNAs with increased
13. Juneau K, Cech TR. In vitro selection of RNAs with increased
12. Juneau K, Cech TR. In vitro selection of RNAs with increased
11. Juneau K, Cech TR. In vitro selection of RNAs with increased
10. Juneau K, Cech TR. In vitro selection of RNAs with increased
9. Juneau K, Cech TR. In vitro selection of RNAs with increased
8. Juneau K, Cech TR. In vitro selection of RNAs with increased
7. Juneau K, Cech TR. In vitro selection of RNAs with increased
6. Juneau K, Cech TR. In vitro selection of RNAs with increased
5. Juneau K, Cech TR. In vitro selection of RNAs with increased
4. Juneau K, Cech TR. In vitro selection of RNAs with increased
3. Juneau K, Cech TR. In vitro selection of RNAs with increased
2. Juneau K, Cech TR. In vitro selection of RNAs with increased
1. Juneau K, Cech TR. In vitro selection of RNAs with increased
binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain. J. Biol. Chem. 1998;273:20556–20567.


Further Reading


See Also

Cross-references to other WECB articles that should contain material relevant to this article have been provided in italics within the main text.
Watson–Crick Base Pairs, Character and Recognition of

Wilma K. Olson, Andrew V. Colasanti and Xiang-Jun Lu,
Rutgers, the State University of New Jersey, New Brunswick
Victor B. Zhurkin,
National Institutes of Health, Bethesda, Maryland
doi: 10.1002/9780470048672.wecb452

The unique Watson–Crick arrangement of hydrogen-bonded bases in DNA accommodates two different, complementary purine–pyrimidine pairs, \( A \cdot T = T \cdot A \) and \( G \cdot C = C \cdot G \), in a common spatial setting. Nature takes advantage of these isomorphous structures, which store genetic information in terms of the proton donor and acceptor atoms that hold the bases in place. As outlined here, the Watson–Crick base pairs carry other chemical signals that are used to recognize and to process specific sequences of bases. The relative stabilities of \( G \cdot C \) versus \( A \cdot T \) pairs reflect their different electronic structures. The distributions of electronic charge on the exposed major-groove and minor-groove edges of the base pairs present unique motifs for direct sequence recognition, and the deformations of the paired bases from ideal, planar configurations provide subtle, indirect recognition elements. The biologic significance of the latter signals is not fully understood but is becoming clearer as more and more high-resolution structures of DNA and RNA are determined.

The simple, yet elegant structure of double-helical DNA—two sugar-phosphate strands wrapped along antiparallel right-handed pathways around a central core of stacked and hydrogen-bonded base pairs—provides the molecular basis to interpret the storage, duplication, and rearrangement of genetic information. The same type of base pairing persists in double-stranded RNA, DNA–RNA hybrid duplexes, and synthetic multi-stranded polymers, such as PNA (1), which allow the chemical message to be duplicated, transcribed, blocked, and so on. The information reported below draws on the three-dimensional spatial arrangements of Watson–Crick base pairs and bound ligands in the many DNA and RNA structures now stored in the Nucleic Acid Database (NDB) (2).

Complementarity

Classic Watson–Crick base pairs are formed by unique hydrogen-bonding interactions between the nitrogenous bases of DNA and RNA. The purine adenine associates specifically with the pyrimidine thymine in DNA (or the related unmethylated analog, uracil, in RNA), and the purine guanine interacts with the pyrimidine cytosine. These complementarity rules, \( A \cdot T \) or \( A \cdot U \) and \( G \cdot C \) pairs, make it possible to build regular double-stranded structures of arbitrary base sequence and to provide a mechanism to copy the genetic code. That is, if the sequence of one strand of DNA is known, the sequence of the other strand is determined automatically. Therefore, if the strands are separated and new DNA is synthesized, two double-stranded DNA molecules are obtained, each an exact copy of the original.

These complementarity rules owe their discovery to the chemical analysis of DNA by Chargaff and associates (3). The DNA from many different organisms shows the same patterns of base composition, namely A and T are always present in equal quantities, as are G and C. The immediate corollary of this observation, that a purine base (R) exists for every pyrimidine base (Y) and vice versa, led Watson and Crick to propose that two helical strands in DNA are held together by specific, intermolecular purine–pyrimidine (R·Y) interactions (4). In turn, this unique chemical complementarity of the double-helical structure, proved to be a major breakthrough to understand the self-recognition and self-reproduction of DNA and forms the cornerstone of structural biology as we know it today, more than half a century later.
By contrast, the proportion of G–C versus A–T base pairs is highly variable in the DNA from different organisms, with over-representation and under-representation of residues found at dimeric, i.e., adjacent base-pair step, and higher levels (5). Factors, which may underlie the observed compositional patterns, are not yet understood.

Isomorphous Base Pairs

The Watson–Crick postulate places complementary bases in similar spatial configurations so that the two-stranded molecule can adopt a regular structure. The bases of each purine-pyrimidine pair lie in a common plane, with the distance, between C1′ sugar atoms approximately constant (~3.05 Å) and the C1′···C1′ vector forming roughly equivalent (~55°) angles with the (purine C1···N9 and pyrimidine C1···N1) glycosidic bonds that join the bases to the sugar-phosphate backbone (Fig. 1). A pseudo-twofold symmetry axis, also referred to as the dyad axis, passes through the center of each base pair, which permits the exchange of complementary bases without change in the relative positions of either the attached sugar residues or the selected base-recognition elements (see below). In particular, an A·T base pair is converted into a T·U pair; and a G·C base pair is converted into a C·G pair by a 180° rotation about this axis. As a result of this isomorphous geometry, any base-pair combinations can be fitted into the same regular structural framework. As a first approximation, the “ideal” double-helical structure of Watson and Crick is sequence-independent. As noted below, this regularity breaks down in high-resolution crystal structures of DNA and is impossible without the R·Y base-pairing interactions.

Hydrogen-bond Recognition and Stability

In turn, the isomorphous structures of the Watson–Crick base pairs dictate specific hydrogen-bonding recognition patterns. The A·T pairs associate via two hydrogen bonds that involve N1(A)····H···N4(T), and N6(A)····H···O4(T), and the G·C base pairs are held in place by three such interactions: N1(G)····H···N3(C), N2(G)····H···O2(C), and O6(G)····H···N4(C) (Fig. 2). The extra hydrogen bond of the G·C pair apparently gives rise to the higher melting temperature of G·C versus AT-rich DNA (8). Direct in vacuo measurements of the binding energies support this idea. Isolated G·C base pairs are more stable than free A·T pairs (~21 vs. ~13 kcal mol⁻¹) (7,8). Furthermore, G·C pairs are typically less deformed from ideal planar geometry than A·T pairs (see Table 2 and discussion below).

Base pairing is substantially weaker in solution. For example, the G·C hydrogen-bonding energy drops to ~5.8 kcal mol⁻¹ in chloroform (9). The hydrogen-bonding energies in water, however, are uncertain because isolated planar bases prefer to associate in parallel stacked arrays rather than to pair with complementary bases (10). According to direct experimental solution measurements of the melting properties of oligonucleotides (which contain both base side groups and the sugar-phosphate backbone), hydrogen bonding adds 0.5–1.8 kcal mol⁻¹ stabilization per base pair depending on DNA sequence (11). Thus, base-pair formation in aqueous solution is governed by base-base hydrogen bonds only slightly more favorable than base-water hydrogen bonds.

The electrostatic character of hydrogen bonding brings the protons of one base and the (N and O) acceptors of the complementary base closer together (1.9–2.0 Å) than their characteristic (2.7–2.8 Å) van der Waals’s separation distance. Accordingly, the base-pairing interaction is stronger and more specific than ordinary van der Waals’s interactions. The hydrogens, which are “shared” in the hydrogen-bonding interactions, have partial positive charges because of their attachment to nitrogen donor atoms, whereas the (carbonyl) oxygen C=O or imidazole nitrogen N) acceptors on the complementary bases bear partial negative charges. To maximize the electrostatic attraction, the hydrogen usually approaches the donor atom along the direction of and in a plane with the lone-pair orbitals of O or N. The partial charges of the donor, hydrogen, and acceptor atoms of the common bases (Fig. 2) determine the overall character of the electrostatic potentials that guide the mutual recognition of base pairs and their interactions with other molecules (see below).

Interestingly, the pairing of guanine and cytosine depicted initially by Watson and Crick (12) entailed only the two hydrogen bonds, O6(G)····H···N4(C) and N1(G)····H···N3(C), congruent with the N6(A)····H···O4(T) and N1(A)····H···N3(T) hydrogen bonds holding adenine and thymine in place. Later, Pauling and Corey (13) showed that guanine and cytosine were joined by a third N2(G)····H···O2(C) hydrogen bond in the minor groove, and Crick (14) used the three G·C hydrogen bonds to account for the higher stability of the Watson–Crick pair compared with a “wobble” G·U pair. Subsequent crystallographic investigations have revealed the existence of “weak” C·H···O hydrogen bonds between nitrogenous bases, also somewhat shorter than the sum of the van der Waals’s distances (15), see Table 2. The geometry of the Watson–Crick A·T base pair naturally forces a third such C2(A)····H···O2(T) hydrogen bond (16), which emphasizes its similarity to the G·C base pair (Fig. 1). Alternatively, the direct contact between adenine and thymine in the minor groove can be interpreted as an electrostatic attraction (17). In any case, this additional interaction in the A·T pair is advantageous for selective recognition during replication and transcription.
Hybridization

In principle, the hydrogen atoms of the purine and pyrimidine bases can rearrange in different tautomeric or hybridized forms. The exocyclic nitrogen atoms attached to the adenine and cytosine rings usually are in the amino (NH₂) form rather than the imino (NH) configuration. Likewise, the exocyclic oxygen atoms attached to the carbons of guanine and thymine rings normally adopt the keto (C=O) form rather than the enol (C–OH) configuration. Watson and Crick suggested that keto to enol or amino to imino base tautomerism could be the origin of the point mutations that underlie evolution (12). Such rearrangements would allow adenine to associate with cytosine or guanine to bind to thymine in geometries close to those of the canonical base pairs. For example, the normal d:d:a pattern of hydrogen-bond donors (d) and acceptors (a) in guanine is converted by keto-enol tautomerism to the d:a:d motif, which complements the preferred a:d:a motif of thymine (Fig. 1). Errors like these destroy the perfect complementarity between opposing chains that gives DNA its capacity for self-replication.

In this regard, significant chemical modification is required to effect either keto-enol or amino-imino tautomerism of the nitrogenous bases, with the consequent formation of A·C and G·T mispairs that fit into the canonical double-helical structure. Notably, the N6-methoxy A*·C mispairs and the O6-methylated G*·T mispairs, which are observed in crystalline duplex structures, are isomorphous with standard A·T and G·C base pairs (Tables 1 and 2; NDB entries: bd0009, bd0126, and bd05819-21).

Fortunately, the imino forms of A and C and the enol forms of G and T occur rarely. Most A·C and G·T mispairs observed to date in high-resolution crystal structures (e.g., References 22 and 23) associate through a “wobble” configuration (14), with the bases “sheared” past one another relative to the Watson-Crick configuration (Tables 1 and 3). These structural perturbations alter the patterns of atomic charges and accessibility that are presumably required for protein recognition and enzymatic action (see discussion below).
Figure 2  Color-coded electrostatic surface of A · T (left) and G · C (right) base pairs produced with the GRASP software package (25, 26) from ideal planar atomic coordinates (27) and partial atomic charges of the CHARMM27 nucleic acid force field (28): (top) the major-groove edges showing the unique donor-acceptor patterns of the base pairs; (middle) the upper faces (24) of purines and the under sides of pyrimidines (see also Fig. 1); (bottom) minor-groove edges showing the common Watson–Crick donor atom “signature.” Conventional hydrogen-bond donor and acceptor atoms on base-pair edges are designated respectively by – and + symbols. The CH and CH₃ groups in the major and minor grooves are noted by (⊥) to emphasize their moderate positive charges and their tendency to be in close contact with O and N acceptor atoms (17, 29). Isopotential contours, in units of kT (numerical scale at top of figure), reveal the approximate electrostatic equivalence of the minor-groove edges of the base pairs. Molecular surfaces generated using a spherical probe with 1.4 Å radius. Bases “neutralized” by adjustment of the partial charges on C1′ atoms. Electrostatic potential omits counterions and incorporates the difference in dielectric between water and bases, i.e., 80 vs. 2. Essentially, the same results are found with other well-known sets of partial atomic charges (30, 31), including those determined with state-of-the-art quantum mechanical methods (32).

Antiparallel Strand Alignment

Because of the lack of symmetry in their chemical structures, individual nucleic acid bases have unique faces (24), which specify the directions of the DNA strands in the Watson–Crick model. The upper faces (tops) of complementary bases point in opposite directions, with the two attached sugar-phosphate backbones aligned in an antiparallel sense. The top of a base corresponds with the configuration that orients the C1′ → N glycosidic bond vector in a “northeast” heading with respect to its “north–south” Watson–Crick base-pairing edge (found between the associated bases in Figs. 1 and 2). If the base and sugar are attached by the normally preferred anti glycosidic linkage (with the six-membered purine ring or the pyrimidine O2 directed away from the sugar ring), the DNA backbone roughly runs perpendicular to the plane of the base, with the 3′-oxygen displaced above the top side and the 5′-oxygen below the bottom side of each base. The vector that connects the tops of consecutive bases coincides with the 5′ → 3′ direction of the sugar-phosphate chain (Fig. 3).

Duplex Grooves and Recognition

The attachment of the sugars to the same side of each Watson–Crick base pair introduces an asymmetry in base-pair accessibility inside the grooves formed by the DNA backbone. The edge of the Watson–Crick base pair that contains the pyrimidine O2 and the purine N3 atoms is called the minor groove, and the longer edge on the opposite side of the glycosidic bonds is termed the major groove (Fig. 1). The hydrogen-bond donors and acceptors that line each groove (Figs. 1, 2) serve as recognition motifs for interactions of DNA with proteins, drugs, and solvent molecules. The pseudo-symmetric positioning of the N3 acceptor atoms of A and G and the O2 acceptor atoms of T and C provides a common minor-groove recognition element for all four Watson–Crick base pairs (33) (shown by the pattern of red and blue atoms in Fig. 2).

The positioning of the N3/O2 acceptor atoms provides a simple and reliable mechanism to distinguish the Watson–Crick pairs from the “wobble” pair and other mismatches. As suggested initially by Brudov and Poltev (34), the fidelity of
nucleic-acid biosynthesis would be increased substantially if the recognition elements of a polymerase had NH or OH groups that interacted with the invariant N3/O2 atoms in the minor groove. The orientation of the amino proton attached to the N2 donor on G differentiates the G-C and C-G base pairs from each other as well as from the A-T and T-A pairs. The latter base pairs can be discriminated by small synthetic molecules that take advantage of both the asymmetric steric structure of the adenine C2-H and the capability of the thymine O2 (with two sets of lone pair electrons) to form an additional hydrogen bond not possible with the pseudo-symmetrically related adenine N3 (37, 38). It is not yet clear whether naturally occurring, DNA-binding proteins use similar principles to distinguish between A-T and T-A base pairs in the minor groove.

By contrast, all four bases present unique protein recognition patterns in the major groove (Figs. 1 and 2). The N7 acceptor atoms of A and G set the purines apart from the pyrimidines, whereas the pseudo-symmetrically placed N6 donor and O6 acceptor, respectively, separate A from G. The corresponding isomorphic interchange of the O4 acceptor and N4 donor on T and C discriminates the two pyrimidines. On the other hand, the latter motifs provide a common identification mechanism for C or A and G or T.

In addition to the classic donor-acceptor recognition mechanism described above, hydrophobic and electrostatic interactions facilitate the discrimination of the base pairs one from another. As shown in Figs. 1 and 2, the pyrimidines are hydrophobic in the vicinity of C5 and either have a positive charge (cytosine) or are of the opposite sign. The positioning of Watson–Crick pairs along the global axis of the double-stranded B-DNA structure enhances the accessibility of major-groove versus minor-groove atoms, thereby favoring the binding of proteins that recognize specific sequences. The narrow B-DNA minor groove is less receptive to proteins but easily accommodates long, crescent-shaped drugs (e.g., Ref. 37 and 38). The binding of proteins in the minor groove often necessitates a partial and sometimes a complete B → A conformational change in DNA, which displaces the base pairs
**Table 2** Average hydrogen-bonding geometry of Watson–Crick base pairs in high-resolution DNA structures

<table>
<thead>
<tr>
<th>Base pair</th>
<th>$d_{C1\cdot C1}$ (Å)</th>
<th>$\lambda^Y$ (deg)</th>
<th>$\lambda^R$ (deg)</th>
<th>Hydrogen-bond distances (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Watson–Crick base pairs†</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A·T (A DNA)</td>
<td>10.4 (0.2)</td>
<td>55.2 (2.3)</td>
<td>55.7 (3.3)</td>
<td>—</td>
</tr>
<tr>
<td>G·C</td>
<td>10.6 (0.2)</td>
<td>55.4 (2.3)</td>
<td>55.7 (3.3)</td>
<td>2.8 (0.2) 2.9 (0.2)</td>
</tr>
<tr>
<td>A·T (B DNA)</td>
<td>10.5 (0.2)</td>
<td>56.9 (2.3)</td>
<td>55.7 (3.3)</td>
<td>—</td>
</tr>
<tr>
<td>G·C</td>
<td>10.7 (0.1)</td>
<td>55.4 (2.2)</td>
<td>55.8 (2.2)</td>
<td>—</td>
</tr>
<tr>
<td><strong>Keto–enol tautomers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A·C</td>
<td>10.4</td>
<td>58.5</td>
<td>55.5</td>
<td>—</td>
</tr>
<tr>
<td>G·T</td>
<td>10.6</td>
<td>57.9</td>
<td>54.6</td>
<td>2.8</td>
</tr>
<tr>
<td><strong>'Wobble' base pairs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A·C</td>
<td>10.3</td>
<td>68.2</td>
<td>47.5</td>
<td>—</td>
</tr>
<tr>
<td>G·T</td>
<td>10.4</td>
<td>71.0</td>
<td>43.6</td>
<td>2.7</td>
</tr>
</tbody>
</table>

*See Table 1 for structures included in the survey.

$\lambda^Y = \angle C1'(Y)–N1(Y)\cdot C1'(R)$ and $\lambda^R = \angle C1'(R)–N9(R)\cdot C1'(Y)$ describe the pivoting of complementary bases in the base pair plane (Fig. 3).
pairs designate the origins of the base-pair frames. The gray arrows designate the positive signs of rotations for created with 3DNA (18) illustrate positive values of the designated right-handed B-DNA structure by enhancing stacking overlaps positive propeller is illustrated). This deformation stabilizes the negative propeller (see parameter definitions in Fig. 3 where positive propeller is illustrated). This deformation stabilizes the right-handed B-DNA structure by enhancing stacking overlaps with bases in adjacent residues (46). The degree of propeller twisting depends on both base-pair and conformational context. The A-T pairs in B-DNA are more distorted on average than the G-C pairs (–13° vs. –10°), and almost no propeller twisting of base pairs exists in the left-handed Z-DNA conformation, e.g., –3° in the d(CGCCCGG); duplex structures (NDB entry: zdf001) (47). Buckle, although fixed on average near zero, shows more pronounced variability than propeller and, for G-C base pairs, exhibits a notable dependence on helical conformation. The G-C base pairs tend to buckle in a positive sense in B-DNA duplexes (as shown in Fig. 3) and in a negative direction in A-DNA structures. The constraints of hydrogen-bond stretching and bending (Table 2) presumably lead to the more limited variations in opening and stretch (Table 3) compared with the other complementary base-pair angles and distances. On average, the A-T pairs are characterized by relatively small magnitudes of buckle, especially in the B form (Table 1), but they normally show a larger variability in propeller twisting compared with the G-C pairs. This observation is consistent with the different hydrogen bonding of A-T versus G-C. Both the buckle and the propeller angles enhance base stacking in B DNA. In G-C pairs, stabilized by three strong hydrogen bonds, excessive propeller twisting is expected to be unfavorable, as this would distort two NH-O bonds in the minor and major grooves (Fig. 3). Thus, propeller twist is expected to be less pronounced in G-C than in A-T pairs. To improve base stacking, apparently the G-C pairs “take advantage” of the other angular, degree of freedom, buckle, which is less prohibitive for the hydrogen bonds. In A-T pairs with only two strong hydrogen bonds (Fig. 3), a large propeller of 15–20° is acceptable. A s a result, there is no need for a large buckle, which remains –1° on average (Table 1, B-DNA).

### Nonplanar Geometry

The Watson-Crick double-stranded model of DNA has been confirmed abundantly and refined with fiber and single-crystal X-ray diffraction studies. The high resolution crystal structures accumulated to date, starting with the 0.84-Å resolution structure of the d(CCGGCGG)2 uridine miniduplex (43) and now including structures of oligonucleotide duplexes of comparable resolution (44, 45), show that complementary Watson-Crick base pairs are not perfectly coplanar. The bases in most solved structures are twisted with respect to each other like the blades of a propeller, with the C1'-atom on the sequence strand typically shifted below and that on the complementary strand displaced above the mean base-pair plane, i.e., negative propeller (see parameter definitions in Fig. 3 where positive propeller is illustrated). This deformation stabilizes the right-handed B-DNA structure by enhancing stacking overlaps away from the global axis and concomitantly exposes unpaired atoms on the minor-groove edges (40). On the other hand, small molecules bind specifically to the narrow major groove of double-stranded RNA (44), which adopts only A-type geometry.

### Protein-induced Base-pair Deformations

The associations of DNA with proteins and drugs introduce additional deformations of Watson-Crick geometry. Several examples, which are illustrated in Fig. 4, demonstrate the functional importance of the base-pair deformations discussed above.

#### Buckle

The partial intercalation of aromatic side groups of the yeast TATA-box binding protein between DNA base pairs (48) is accompanied by a pronounced (32°) buckling in one of the two adjacent A-T pairs, which apparently facilitates “penetration” of the phenylalanine rings into the minor groove (NDB entry: pdb012; Fig. 4a). A second example involves the integration host factor-DNA complex (49) (NDB entry: pdb040; Fig. 4b), where the partial minor-groove insertion of arginine 63 and close association of arginine 60 result in marked buckling in the opposite direction of the surrounding T-A-T and G-C-C base pairs (with respective buckle angles of –47° and –35°).
Watson–Crick Base Pairs: Character and Recognition of…

Figure 4  Protein-induced distortions of complementary base-pair parameters in representative crystal structures: (a) view from the major groove of the large buckling of the (upper) A₈T₂₂ base pair in the DNA complexed with the yeast TATA-box binding protein (48) brought about by the partial minor-groove visitation at the A₈₉₂₂ (major) step of phenylalanine 99 (wire-frame ring connected to the magenta (glutamic acid 93 and isoleucine 103) polypeptide ribbon; NDB entry: pdt022); (b) major-groove view of negatively buckled (upper) T₃₇A₃₇ and (lower) G₃₆C₃₆ base pairs, the hydrogen-bonded (arginine 60 (lower right) and arginine 63 (upper left) side groups, and a fragment (glutamine 59 and lysine 66) of the minor-groove-bound, extended β-sheet recognition element of integration host factor (49) (NDB entry: pdt040); (c) view of the upper face of the opened A₁₃T₁₉ base pair and the minor-groove contacting (serine 183 and asparagine 190) C-terminal ribbons of the Hin recombinase (50) (NDB entry: pde009); (d) major-groove view of the extreme opening and base-pair displacement of the chemically modified 4′-thio-2′-deoxycytidine in the C₅Me₄₀₇G₄₀₈C₄₀₉·G₄₂₈C₄₂₇G₄₂₆ trimer step (C₅Me₄₀₇·G₄₂₈ at top, C₄₀₉·G₄₂₆ at bottom) and closely associated amino acids (glycine 78 and lysine 91) in the complex of DNA with HhaI methyltransferase (51) (NDB entry: pde141); (e) minor-groove view of the staggered (lower) T₄₉₉A₅₄₁ base pair, the H6(G₁₄₁)···O6(C₁₄₄) inter-strand hydrogen bond (dashed line), and the major-groove recognition helix (serine 1151 and lysine 1165) of RXR-a in the complex with its idealized direct repeat DNA target (52) (NDB entry: pd0071). Images created with 3DNA (18) and Raster3D (53). Planes of bases colored as follows: A—red; T—blue; G—green; C—yellow.

Opening and stretching

A subtle (24°) base-pair opening of A₁₃T₁₉ is induced by contacts of Hin recombinase (50) with the minor-groove edge of T₁₉ (NDB entry: pde009; Fig. 4d). By contrast, the major-groove capture of cytosine by HhaI DNA cytosine-5-methyltransferase (51) introduces nearly maximal opening (178°) and extreme lateral base-pair displacement, i.e., stretch (8.5 Å), of the broken base pair (NDB entry: pde141; Fig. 4c).

Stagger

The close fit of the recognition helix of the 9-cis retinoic acid receptor, RXR, against the DNA major groove is responsible for the noticeable stagger (1.2 Å) of the T₁₄₉₉A₅₄₁ base pair.
the accompanying inter-strand N6(A1541)···O6(G1498) hydrogen bonding (52) (NDDB entry: pd0071; Fig. 4e).

Summary

The Watson–Crick base-pairing scheme is characterized by several unique structural properties, including the complementarity and perfect isomorphism that are used in replication and in transcription of the genetic code. These general features underlie the ability of DNA to incorporate any arbitrary sequence in a nearly regular duplex. On the other hand, the subtle, sequence-dependent variability of A·T versus G·C base-pairing geometry is used by the DNA-binding proteins involved in regulation. The ingenious base-pairing principle postulated more than half a century ago and subsequently confirmed in high-resolution crystal structures of DNA and RNA continues to surprise us by its beauty, simplicity, and complexity.

Acknowledgments

We are grateful to Dr. A.R. Srinivasan for generous help. Support of this work through U.S.P.H.S. Grant GM20861 is gratefully acknowledged. Computations were carried out at the Rutgers University Center for Computational Chemistry and through the facilities of the Nucleic Acid Database project.
Watson–Crick Base Pairs, Character and Recognition of

References


Watson–Crick Base Pairs, Character and Recognition of


See Also

DNA-Based Structures; DNA Recognition by Enzymes; Nucleic Acid Hydration; Nucleic Acid Recognition; Peptide Nucleic Acids; Protein-Nucleic Acid Interactions; Small Molecule-Nucleic Acid Interactions
Approaches to Enzyme Inhibition

Richard B. Silverman, Northwestern University, Evanston, Illinois
doi: 10.1002/9780470048672.wecb157

This article describes various approaches to inhibition of enzyme catalysis. Reversible inhibition includes competitive, uncompetitive, mixed inhibition, noncompetitive inhibition, transition state, and slow tight-binding inhibition. Irreversible inhibition approaches include affinity labeling and mechanism-based enzyme inhibition. The kinetics of the various inhibition approaches are summarized, and examples of each type of inhibition are presented.

Enzyme inhibitors are compounds that interact with an enzyme and slow down or prevent catalysis from occurring. Natural enzyme inhibitors are often present to control metabolism; synthetic inhibitors generally are used for the purpose of blocking enzyme-catalyzed reactions in the treatment of human and lower animal disease and in agriculture. Many biological systems are controlled with two molecules that have opposite activities, such as excitatory and inhibitory neurotransmitters or hypertensive and hypotensive peptide hormones. Specific enzymes are responsible for the biosynthesis of these molecules can be selectively inhibited. Diseases often develop from an excess or deficiency of a particular metabolite, from infestation of a foreign organism, or from aberrant cell growth. All of these disease etiologies, at least when related to an enzyme activity, can be addressed by specific enzyme inhibition. For example, if an excess of a particular metabolite exists, then inhibition of the enzyme that produces that metabolite would decrease its concentration; inhibition of an enzyme that degrades a particular metabolite would lead to an increase in that compound, thereby reversing a deficiency of that molecule. Inhibition of an essential enzyme for a foreign organism, particularly if humans do not have that enzyme, would be an ideal method for selective toxicity of the foreign organism. A similar approach can be taken to destroy insects, fungi, or weeds for agricultural purposes, namely by inhibiting an essential enzyme for those pests. Inhibition can occur either reversibly or irreversibly. In this article, the various common approaches of reversible and irreversible inhibition are summarized with relevant examples provided for each. The reversible inhibition methods include competitive, uncompetitive, mixed, noncompetitive, transition state analog, multisubstrate analog, slow-binding, and slow tight-binding inhibition. Irreversible inhibition includes affinity labeling and mechanism-based inhibition. Derivations of the kinetic equations are given in the books in the Further Reading section at the end of this review.

Reversible Enzyme Inhibition

The most common type of inhibition is reversible inhibition. As the name implies, a reversible inhibitor functions by binding reversibly (generally noncovalently) to the target enzyme. An inhibitor forms a complex with the enzyme (an E·I complex); the dissociation of that complex is measured by the Ki (the ratio of the rate of dissociation of the inhibitor to association of the inhibitor, koff/kon; the concentration of inhibitor that gives half the maximal inhibition). As it is a dissociation constant, the smaller the Ki value for an inhibitor, the tighter the binding, and the more potent is the inhibitor. How these dissociation constants for the various types of inhibition are determined experimentally is described in the respective sections below.

Reversible inhibition that produces complete loss of catalytic activity is referred to as linear inhibition because the plots of K_m/V or 1/V versus [I] are straight lines. When some catalytic activity remains, even at saturating amounts of inhibitor, it is referred to as hyperbolic inhibition because these plots are nonlinear (this case will not be considered here). Both of these types of reversible inhibition are further classified according to the various apparent Michaelis-Menten parameters that are affected by the inhibitor. The two limiting cases are competitive inhibition and uncompetitive inhibition; a third type is mixed inhibition, which includes as a special case noncompetitive inhibition.

The simplest method to determine whether an inhibitor is reversible after enzyme inhibition with excess inhibitor is to remove the inhibitor by dialysis or gel filtration. If full enzyme activity returns, then inhibition is reversible. However, it
Approaches to Enzyme Inhibition

Scheme 1

Competitive inhibition. does not answer the question of whether the compound is non-covalently bound or covalently bound with a weak, reversible covalent interaction.

Competitive Reversible Enzyme Inhibition

The most common type of reversible inhibition is when the inhibitor binds to the free enzyme in the substrate binding site (called the active site), which is known as competitive reversible inhibition. Competitive inhibition prevents the substrate from binding because of a competition between the substrate and inhibitor for binding to the active site (Scheme 1). Particularly in medicinal chemistry, another common expression of inhibition, in addition to the other approaches (i.e., non-competitive inhibition), is the IC$_{50}$ value, the concentration of an inhibitor that results in 50% inhibition of the enzyme in the presence of a specific concentration of substrate. The IC$_{50}$ and $K_i$ values for a competitive reversible inhibitor are roughly interconverted by the following expression (Eq. 1) (1–3)

$$IC_{50} = \frac{(1 + [S]/K_m)}{K_i}$$

Kinetics of Simple Competitive Inhibition

The rate of the reaction, which depends on [I], $K_i$, [S], and the $K_m$, is depicted in Equation 2.

$$v = \frac{V_{max}[S]}{[K_m(1 + [I]/K_i)] + [S]}$$

The reciprocal of this equation, known as the Lineweaver-Burk equation (4), is shown in Equation 3.

$$\frac{1}{v} = \frac{K_m}{V_{max}}\frac{1}{[S]}\frac{1}{[I] + [I]/K_i} + \frac{1}{V_{max}}$$

The Lineweaver-Burk expression is depicted graphically in Fig. 1. This graph correlates the rate as a function of [S], first in the absence of inhibitor (line a), then with added inhibitor concentrations (lines b–e). Note that as more inhibitor is added, the slope of the line increases, but the $V_{max}$ does not vary. As 1/[S] approaches zero (infinite [S]), the lines converge on the data in the absence of inhibitor (all of the inhibitor is displaced by substrate), namely, 1/$V_{max}$. At low [I], inhibitor competes effectively with substrate for the enzyme. When a competitive inhibitor binds to the free enzyme, there are fewer enzyme molecules for the substrate to bind to, so the rate decreases. According to Equations 2 and 3, the reciprocal of the rate is proportional to $K_m/V_{max}$, so the slopes of the lines vary with [I] (in the presence of inhibitor, it requires more substrate to attain the same $V_{max}$ as in the absence of inhibitor). Therefore, the $K_m$ appears to be larger in the presence of inhibitor (a larger $K_m/V_{max}$ means a larger slope). The apparent $K_m$ ($K_{m,app}$) at each [I] can be calculated from each negative x-axis intercept using Equation 4.

$$-\frac{1}{K_{m,app}} = -\frac{1}{K_m(1 + [I]/K_i)}$$

Other ways of displaying kinetic data are with the use of Dixon (5) and Cornish-Bowden (6) plots.

Example of Simple Competitive Inhibition

As the most common type of inhibition, myriad examples of competitive reversible inhibition exist in the literature. The following is a selection from Thomas Poulos and my laboratories related to the selective inhibition of the neuronal isozyme of

![Graphical depiction of Lineweaver–Burk expression for determination of competitive inhibitor kinetic constants.](image-url)
Approaches to Enzyme Inhibition

nitric oxide synthase (NOS) (7). Nitric oxide (NO) is a ubiquitous biological messenger involved in a variety of physiological processes that acts as a signal transducer but also exerts a variety of regulatory and cytostatic functions (8). It is produced by the enzyme NOS, which catalyzes the oxidation of L-arginine to L-citrulline and NO in a NADPH- and O2-dependent process (9, 10). There are at least three distinct isoforms of NOS. The constitutive endothelial isoform (eNOS) is involved in the regulation of smooth muscle relaxation and blood pressure and in the inhibition of platelet aggregation (11). A second constitutive isoform is neuronal NOS (nNOS), which is important for neurotransmission (12). A third isozyme is inducible NOS (iNOS), which is located in activated macrophage cells and acts as a cytotoxic agent in normal immune responses (13). All of the isoforms use NADPH, FAD, FMN, tetrahydrobiopterin, and heme as cofactors. The constitutive isoforms also require Ca2+ and calmodulin for activity, whereas the inducible isoform has tightly bound Ca2+ and calmodulin. They share only approximately 50% of primary sequence homology, suggesting that they may differ from each other in regulatory aspects. The use of NOS inhibitors in pathologically elevated synthesis of NO has great therapeutic potential (14, 15). NO overproduction by nNOS has been associated with neurodegeneration during stroke, spinal transmission of pain, migraine headaches, Parkinson’s disease, and Alzheimer’s disease. Thus nNOS represents an important therapeutic target for nNOS-selective inhibitors (16). Enhanced formation of NO after the induction of iNOS appears to be important in the tolerance to and dependence on morphine, development of colitis, cancer, and tissue damage and inflammation (17, 18). Selective inhibition of one isozyme over the others is essential because the three isoforms of NOS have unique roles in separate tissues (19).

Figure 2  (a) Replot of data from Lineweaver–Burk plots (Fig. 1) for determination of $K_i$ values for competitive inhibitors. (b) Alternative replot of data from Lineweaver–Burk plots (Fig. 1) for determination of $K_i$ values for competitive inhibitors.

Figure 3  (a) Compound 1 bound to active site of nNOS. (b) Compound 1 bound to active site of eNOS.
An analog of (L)-nitroarginine, L-nitroargininyl-L-2,4-diaminobutyramide (1), was shown to be a potent $K_i$, 130 nM and highly selective inhibitor of nNOS (1538-fold more potent than with eNOS and 192-fold more potent than with iNOS) (20). In addition to a kinetic analysis that demonstrated competitive inhibition, dialysis of the inhibited enzyme led to the return of enzyme activity, indicating that 1 is a reversible inhibitor. The competitive nature of the inhibition was firmly established by X-ray crystallography, showing 1 in the active site (7) of both nNOS (Fig. 3a) and eNOS (Fig. 3b). The nitroargininyl moiety binds at the guanidino binding site for L-arginine. However, in nNOS 1 adopts a somewhat curled conformation, whereas in eNOS it has an extended conformation. It is thought that this binding difference develops from two sources: a 1 amino acid difference in the active site, namely, Asp597 in nNOS is Asn368 in eNOS, which is responsible for an important electrostatic interaction in nNOS that is missing in eNOS, and the smaller Val106 side chain in eNOS near the active site entry, which allows inhibitors to adopt an extended conformation while the larger Met336 in nNOS encourages a curled conformation. If Asp597 in nNOS is mutated to Asn, then 1 adopts an extended conformation in D597 N nNOS as if it was binding to eNOS. Furthermore, the $K_i$ value for 1 with the D597 N mutant is 223 times larger (weaker binding) than wild-type nNOS, again simulating binding interactions like those with eNOS (21). Conversely, if Asn368 in eNOS is mutated to Asp, then 1 adopts a curled conformation in N368D eNOS, as if it was nNOS, and the $K_i$ value for 1 bound to N368D eNOS is 10 times lower (more potent) than in wild-type eNOS, again simulating binding to nNOS rather than to eNOS. A double mutant of eNOS, N368D/V106 M eNOS, was prepared to make the active site of eNOS appear even more nNOS-like, and the $K_i$ value for 1 was an additional fivefold lower than the single mutant; the crystal structure showed that 1 was still in the curled conformation, as in the case with wild-type nNOS. Likewise, 1 bound to the D597N/M336V double mutant of nNOS was in the extended conformation (Fig. 4) as if it was bound to eNOS, although there was little change in the $K_i$ value relative to that of the single mutant.

Uncompetitive Enzyme Inhibition

The other extreme is when a compound binds only to the $E\cdot S$ complex but not to the free enzyme, in which case uncompetitive inhibition occurs (Scheme 2). Although it is rare in single substrate reactions, it is common in multiple substrate systems. An inhibitor of a two-substrate enzyme that is competitive against one of the substrates often is found to give uncompetitive inhibition when the other substrate is varied. The inhibitor binds at the active site but only prevents the binding of one of the substrates.

Kinetics Of Uncompetitive Inhibition

The rate of this type of inhibition is described by Equation 5; the Lineweaver–Burk transformation of this rate equation is given
Approaches to Enzyme Inhibition

Figure 5 Lineweaver–Burk plots for determination of uncompetitive inhibitor kinetic constants.

\[ v = \left( \frac{V_{\text{max}}[S]}{1 + [I]/K_i} \right) / \left( \frac{K_m}{1 + [I]/K_i} + [S] \right) \]

(5)

\[ \frac{1}{v} = \left( \frac{K_m}{V_{\text{max}}} \right) \left( \frac{1}{[S]} \right) + \left( \frac{1}{V_{\text{max}}} \left( \frac{1}{1 + [I]/K_i} \right) \right) \]

(6)

As uncompetitive inhibitors do not bind to free enzyme, the inhibitor has no effect on the \( V_{\text{max}}/K_m \), and the slopes are independent of inhibitor concentration (Fig. 5). The \( K_i \) values can be obtained from replots of either \( 1/V_{\text{max,app}} \) (the y-axis intercepts) versus [I] (Fig. 6a) or \( 1/K_{m,app} \) (the x-axis intercepts) versus [I] (Fig. 6b).

The \( IC_{50} \) value for an uncompetitive inhibitor is related to the \( K_i \) by Equation 7 (22). When \([S] \gg K_m\), which is often the case, the \( IC_{50} = K_i \).

\[ IC_{50} = K_i (1 + K_m/[S]) \] (7)

Example of Uncompetitive Inhibition

Tuberculosis is responsible for 2,000,000 deaths worldwide annually (23). It is one of the leading opportunistic infections in AIDS patients (24) and is spreading because of multidrug-resistant strains of the organism that causes tuberculosis, namely Mycobacterium tuberculosis (25). The principal drug used for tuberculosis is isoniazid (\( 2, Scheme 3 \)), which inhibits the biosynthesis of mycolic acids, thereby disrupting the mycobacterium’s cell wall (26). The target for isoniazid is the enoyl reductase (known as InhA) in the type II fatty acid biosynthesis pathway (27). InhA is inhibited by the product of the mycobacterial catalase-peroxidase (KatG) (28) reaction of isoniazid and NADH (\( 3, Scheme 3 \)) (29–31). Resistance to isoniazid principally develops from mutations in KatG, which blocks the activation of isoniazid (32). Compounds that directly inhibit the ultimate target, InhA, without requirement of activation by KatG should be more promising drugs for resistant Mycobacterium tuberculosis.

![Figure 5](image-url)
A series of uncompetitive inhibitors of InhA were developed using structure-based design from the crystal structure of triclosan (4) bound to both E. coli enoyl reductase (called ecFabI) (33) and to InhA (Fig. 7) (34). Triclosan was shown by kinetic analysis to be an uncompetitive inhibitor of InhA (35). The structure of triclosan bound to InhA confirmed its uncompetitive inhibition properties; triclosan binds to InhA only in the presence of NAD\textsuperscript{+}. The phenol ring of triclosan π-stacks with the nicotinamide ring of NAD\textsuperscript{+} as well as forming a hydrogen bond to Tyr158 and the 2′-hydroxyl group of NAD\textsuperscript{+}. Based on this structure, a series of alkyl diphenyl ethers, which are uncompetitive inhibitors of InhA and do not require activation by KatG, was designed. The most potent analog (5) has a \(K_i\) value of 1 nM for InhA and MIC\textsubscript{99} (minimum inhibitory concentration to destroy 99% of the organisms) values of 2–3 \(\mu\)g/mL (6–10 \(\mu\)m) for drug-sensitive and drug-resistant strains of Mycobacterium tuberculosis.

**Mixed Enzyme Inhibition**

Sometimes an inhibitor can bind to both the free enzyme (E) and to the enzyme-substrate (E·S) complex, resulting in mixed inhibition (Scheme 4). This type of inhibition involves binding of the inhibitor to a site other than at the active site for binding to the E·S complex to occur. A special case of mixed inhibition when \(K_m\) and \(K_m'\) are equal is called noncompetitive inhibition.

**Kinetics of Mixed Inhibition**

If dissociation of the substrate from the E·S complex is the same as that from the E·S·I complex (i.e., \(K_m = K_m'\)), then pure noncompetitive inhibition occurs, and the rate is given as Equation 8. The Lineweaver-Burk equation from this is shown in Equation 9, which is depicted graphically in Fig. 8. Unlike competitive inhibitors, noncompetitive inhibitors affect \(V_{max}\) (interception on the y-axis) is different at different inhibitor concentrations but not \(K_m\) (all concentrations of inhibitor give lines with the same interception on the x-axis). A replot of these data either as the slope versus \(I\) (Fig. 9a) or \(1/V_{max}\) versus \(I\) (Fig. 9b) provides the \(K_i\) value, Dixon (5) and Cornish-Bowden (6) plots also can be constructed.

The situation above, where \(K_m = K_m'\), for dissociation of the substrate from both the E·S and the E·S·I complexes, is actually relatively rare compared with the case of different dissociation constants. If the E·I complex has a lower affinity for
Appropes to Enzyme Inhibition

Figure 8  Graphical depiction of Lineweaver–Burk expression for determination of pure noncompetitive inhibitor kinetic constants.

Figure 9  (a) Replot of data from Lineweaver–Burk plots (Fig. 8) for determination of $K_i$ values for pure noncompetitive inhibitors. (b) Alternative replot of data from Lineweaver–Burk plots (Fig. 8) for determination of $K_i$ values for pure noncompetitive inhibitors.

the substrate than free enzyme, then inhibition is a mixture of competitive and uncompetitive inhibition, referred to as mixed inhibition. In the presence of inhibitor, there will always be some ESI complex, which lowers the free enzyme concentration, and $V_{max}$ will be less than with just free enzyme (the ESI complex is nonproductive). As the EI complex has a lower affinity for substrate than free enzyme, the $K_m$ will be greater (lower affinity) for the EI complex than that with free enzyme. The Lineweaver–Burk equation for this case is shown in Equation 10.

$$\frac{1}{V} = \frac{\frac{K_m}{V_{max}}}{(1/S)(1 + [(I)/K_i])}$$

$$+ \frac{1}{V_{max}}(1 + [(I)/K_i]) \quad (10)$$

In this case, the interception of lines as a function of inhibitor concentration can occur either above the x-axis (Fig. 10a) or below the x-axis (Fig. 10b). $K_i$ values for mixed inhibition are obtained by replotting the data from the Lineweaver–Burk plot as the slope ($\frac{V_{max}}{V_{max}}[1 + [(I)/K_i]]$) versus $[I]$ (Fig. 11a) or as the y-axis intercept ($\frac{1}{V_{max}}$) versus $[I]$ (Fig. 11b). With Fig. 11a, the x-axis intercept is $-K_i$; with Fig. 11b, the x-axis intercept is $-K_i'$. The $IC_{50}$ value for a mixed inhibitor can be related to the $K_i$ by Equation 11 (36). For a noncompetitive inhibitor, in which $K_i = K_i'$, or when the $[S] \gg K_m$, $IC_{50} = K_i$.

$$IC_{50} = \frac{(K_m + [S])/(K_m/K_i) + ([S]/K_i')]}{(1 + [(I)/K_i'])} \quad (11)$$

Example of Mixed Inhibition

The mitogen-activated protein kinase (MAPK) signaling pathway is responsible for the regulation of cell growth and differentiation consisting of a MAPK (such as ERK1 and ERK2), a MAPK kinase (such as MEK1 and MEK2), and a MAPKK kinase (such as M KK1 and M KK2). The MAPK signaling pathways are important for the formation, progression, and survival of tumors (40) as well
as being intracellular mediators in many inflammatory processes (41). MEK1 and MEK2 phosphorylate ERK at specific tyrosine and threonine residues, which is a prerequisite for cell proliferation. MEK1 and MEK2 are equally competent to phosphorylate ERK (42), so both are targets for the design of antitumor agents. Compound 6 has been found to be a potent and highly selective inhibitor of MEK1 and MEK2; it was shown kinetically to be a mixed inhibitor with both Mg·ATP and MAPK and went into clinical trials for the treatment of colon cancer (43). Further evidence for the mixed inhibition properties of 6 was provided by crystal structures of an analog of 6 (compound 7) bound to an N-terminally truncated form of human MEK1 containing Mg·ATP (to 2.4 Å resolution) and compound 8 bound to an N-terminally truncated form of human MEK2 containing Mg·ATP (to 3.2 Å resolution) (44). Therefore, the inhibitors bind to the E-S complex. The Mg·ATP in MEK1 binds in a location similar to that in other protein kinases; 6, however, binds in a separate binding site adjacent to Mg·ATP (Fig. 12). The side chains of Lys97 and Met143 separate these two binding pockets. When Mg·ATP and 6 are bound to unphosphorylated MEK1, the two lobes adopt a closed conformation, causing conformational changes in the activation loop and helix C, which interferes with an important electrostatic interaction between a conserved glutamate residue (Glu114) and a conserved lysine residue (Lys97) in the ATP binding site, resulting in stabilization of an inactive conformation of the enzyme. A similar structure was obtained for 8 bound to MEK2 containing Mg·ATP. It is believed that the specificity of these mixed inhibitors derives from their binding in a region where sequence homology to
Approaches to Enzyme Inhibition

Figure 12: Crystal structure of MEK1 showing MgATP and mixed inhibitor 6 (referred in the paper as PD318088) in different binding pockets.

other protein kinases is very low and distinct from the highly homologous ATP binding site.

Transition State Analog Inhibition and Slow Tight-Binding Inhibition

A special type of reversible inhibition in which the inhibitor resembles the transition state structure of the enzyme-catalyzed reaction is known as transition state analog inhibition. It has long been believed that an enzyme catalyzes its reaction by an initial interaction of the substrate with the active site of the enzyme, triggering a conformational change in the enzyme, which induces the chemistry of the reaction (45, 46). As the reaction approaches the transition state, the binding interactions with the substrate increase until, at the transition state, maximal binding interactions result. These interactions produce transition state stabilization, which is responsible for lowering the activation energy and accelerating the rate of the reaction.

This process suggests that enhanced binding by an inhibitor to an enzyme would occur if the inhibitor structure resembled that of the transition state of the reaction rather than the ground state (i.e., the substrate structure) (47-50), which has been shown to be the case in many examples (51). Typically the kinetics of inhibition of transition state analog inhibitors are different from that of other types of reversible inhibitors, which exhibit instantaneous inhibition (millisecond time scale); transition state analog inhibitors generally produce time-dependent inhibition (t1/2 of seconds or even hours) as a result of a conformational change that is induced when these inhibitors bind to the enzyme, leading to a slow isomerization of the initial E·I complex to another complex (E·I*) in which the inhibitor is bound more tightly to the enzyme than in the E·I complex (Scheme 5). This process is known as slow-binding inhibition.

Kinetics of Slow-Binding Inhibition and Slow Tight-Binding Inhibition

Reversible inhibitors that attain equilibrium between enzyme, inhibitor, and the E·I complex slowly, as compared with the enzyme-catalyzed substrate reaction, are called slow-binding inhibitors (52). In some cases, the ratio of total inhibitor to total enzyme must be high, as in the case of the classic competitive inhibitors; but in other cases, the attainment of the equilibrium of E, I, and the E·I complex occurs when the inhibitor concentration is approximately the same as the enzyme concentration, in which case the inhibitors are called slow tight-binding inhibitors (53, 54). Slow-binding and slow tight-binding inhibition generally occur in competition with the substrate, but also can be a mixed type inhibition. The lifetimes of enzyme complexes with slow-binding inhibitors
Approaches to Enzyme Inhibition

or slow tight-binding inhibitors are relatively long and exhibit slow off rates \(k_{-2}\); the on rates \(k_{2}\) may be fast or slow. As the rate of release of the inhibitor from the E·I complex becomes exceedingly slow, inhibition approaches irreversible. The overall dissociation constant for the E·I* complex \(K_i^*\) is defined by Equation 12.

\[
K_i^* = \frac{[E][I]}{[E·I] + [E·I^*]} = \frac{k_{-2}}{k_{2} + k_{-2}}
\]

Equation 12

Progress curves for a slow-binding or slow tight-binding inhibitor are described by general Equation 13, where \(v_o\), \(v_s\), and \(k\) represent, respectively, the initial rate, the final steady-state rate, and the apparent first-order rate constant for establishment of the equilibrium between E and E·I*. The following equations are valid only under specific simplifying conditions (52). Equilibria to the formation of the E·S (from E + S) and E·I (from E + I) complexes must be attained rapidly. In addition to a slow establishment of equilibrium between E·I and E·I*, the reverse rate \(k_{-2}\) of this equilibrium must be much less than the forward rate \(k_{2}\). The initial rate is obtained from Equation 14, where \(V_{\text{max}}\) is the maximum rate, \([S]\) is the concentration of the substrate for which I is an inhibitory analog.

\[
v_o = \frac{V_{\text{max}}[S]}{K_m}\frac{1}{1 + [I]/K_i} + [S]
\]

Equation 14

The final steady-state rate is obtained from Equation 15, where \(K_i^*\) is the overall inhibition constant as defined in Equation 12. The apparent first-order rate constant \(k\) for the interconversion of E·I and E·I* in the presence of substrate S is expressed in Equation 16. Figure 13 gives an example of a progress curve for a slow-binding inhibitor. For each curve with inhibitor present, there is an initial burst followed by a slower steady-state rate.

\[
k = k_{-2} + k_{2}\left(\frac{[I]}{K_i} + \frac{[S]}{K_m}\right)
\]

Equation 16

Example of a Transition State Analog Inhibitor that Exhibits Slow Tight-Binding Inhibition

5′-Methylthioadenosine/S-adenosylhomocysteine nucleosidase (MTAN) catalyzes the hydrolysis of 5′-methylthioadenosine (MTA), a byproduct of polyamine synthesis, to adenine and 5-methylthio-D-ribose (MTR) (Scheme 6). MTA is a feedback inhibitor of polyamine synthesis, so inhibition of MTAN would inhibit polyamine synthesis and the salvage pathways for adenine and methionine, which would result in antibacterial (55) or antitumor activity (56). Kinetic isotope effects on this reaction were measured to determine the transition state structure (57). A large 1′-3H and small 1′-14C isotope effect suggest that the MTAN-catalyzed reaction proceeds by a dissociative (S_{+1}) mechanism with little involvement of the leaving group or attacking nucleophile at the transition state, which indicates that the transition state has significant ribooxacarbenium ion character. Based on this hypothetical transition state structure, two series of transition state analog inhibitors were designed (9 and 10) to mimic the transition state structure. The protonated
Approaches to Enzyme Inhibition

Scheme 6

\[ \text{5'-Methylthioadenosine/S-adenosylhomocysteine nucleosidase (MTAN)-catalyzed hydrolysis of 5'-methylthioadenosine (MTA) to adenine and 5-methylthio-D-ribose (MTR).} \]

1. Nitrogens in each series mimic the ribooxacarbenium ion intermediate (related to the transition state) and a methylene bridge was added between the ribooxacarbenium ion mimic and the 9-deazaadenine base to emulate the N-ribosidic bond distance of a fully dissociated transition state. A kinetic analysis shows that 9 and 10 are competitive inhibitors; progression curves demonstrated the time dependence of the inhibition. Application of Equations 14–16 to the 9 series resulted in dissociation constants in the picomolar range and for the 10 series in the femtomolar range. Compound 11 has a \( K_i = 2.6 \mu M \) and a \( K^* = 47 \mu M \) with a \( K_m/K_i^* \) value of 9.1 million and 91 million (11 binds this amount greater) relative to the substrates MTA and S-adenosylhomocysteine, respectively! These are some of the most potent noncovalent competitive enzyme inhibitors known. Furthermore, there is no MTAN in humans; 5'-methylthioadenosine phosphorylase (MTAP) is the only enzyme in humans capable of metabolizing MTA, and MTAN has a larger 5'-alkylthio binding site. Specificity factors in the thousands between inhibition of MTAN in bacteria over MTAP in humans were achieved, which allows for selective toxicity against microorganisms.

The crystal structures of 9 (\( R = CH_3 \)) and 10 (\( R = CH_3 \)) bound to MTAN at 2.2 Å resolution revealed a new ion pair between the positive charge on the ring and the nucleophilic water (hydroxide ion) (58). Based on the distances of these interacting atoms, 9 (\( R = CH_3 \)) appears to be a mimic of an early transition state, and 10 (\( R = CH_3 \)) mimics a highly dissociated transition state, for the MTAN-catalyzed reaction.

Example of a Multisubstrate Analog Inhibitor

Purine nucleoside phosphorylase (PNP), which catalyzes the reversible phosphorolysis of ribonucleosides to ribose 1-phosphate and the free base (Scheme 7), plays an important role in purine nucleoside metabolism and in T-cell development (59, 60). PNP inhibitors have therapeutic utility for selective destruction of T cells in T-cell leukemias and T-cell lymphomas as well as in the treatment of T-cell-mediated autoimmune diseases and for the suppression of the post-organ transplant T-cell response (61).

To mimic the long bond length in the transition state, compound 12 was synthesized as a multisubstrate inhibitor (62). This compound mimics both substrates of PNP, the purine nucleoside, and phosphate. The two-carbon bridge between the ribose ring and the phosphate group was incorporated to place the phosphonate moiety into the phosphate-binding site. A crystal structure of 12 bound in PNP at 1.7 Å resolution (Fig. 14) resembles that of a previously determined structure of bovine
Scheme 7  Purine nucleoside phosphorylase (PNP)-catalyzed phosphorolysis of ribonucleosides to ribose 1-phosphate and the free base.

Figure 14  Stereographic projection of 12 bound to active site of PNP at 1.7 Å resolution.

PNP with substrate analogs (inosine and sulfate) bound (63). The nucleoside part of 12 binds in the purine nucleoside binding pocket, and the phosphate binds in the phosphate binding pocket. The IC_{50} for 12 is 30 nM, one of the most potent PNP inhibitors reported.

Substrate and Product Inhibition

Michaelis-Menten kinetics predict that as the concentration of the substrate increases, the rate increases hyperbolically. However, some enzymes exist in which a maximum velocity is obtained at low substrate concentration, but further increases in the substrate concentration lead to a decrease in velocity. This effect is known as substrate inhibition and can eventually lead to complete enzyme inhibition or partial enzyme inhibition. It is thought that substrate inhibition occurs if two substrate molecules bind to the enzyme simultaneously in an incorrect orientation and produce an inactive E·S·S complex, analogous to that discussed for uncompetitive inhibition. The rate of the enzyme reaction that undergoes substrate inhibition is given by Equation 17, where $K_i$ represents the equilibrium constant for formation of the E·S·S complex from E·S and S and $K_{m}^{'}$ is a modified Michaelis constant. The reciprocal of Equation 17 is described by Equation 18, which is analogous to the Lineweaver-Burk equation and is depicted graphically in Fig. 15.

$$
\frac{1}{v} = \frac{1}{V_{max}[S]} + \frac{1}{V_{max} + (K_{m}^{'}[S]/V_{max})} 
$$  (18)

Another type of inhibition results from the product formed in an enzyme-catalyzed reaction (64). The rate of substrate turnover in the presence of preexisting product is less than the initial rate at the same substrate concentration in the absence of product. This process occurs for two reasons: Some of the enzyme is in an E·P complex so there is less enzyme available to which the substrate can bind, which increases the apparent $K_m$ for the substrate; some of the product is being converted back to substrate, which makes it appear as if less substrate is available.

Irreversible Enzyme Inhibition

Irreversible enzyme inhibition, also called enzyme inactivation (or active-site directed irreversible inhibition, because it is generally competitive with substrate), occurs when a compound blocks the enzyme activity for an extended period of time, generally via covalent bond formation. Therefore, even though some slow tight-binding inhibitors functionally block the enzyme activity irreversibly, they are still considered reversible.
inhibitors because they can dissociate from the enzyme, albeit very slowly. The most effective inactivators are ones whose structures are similar to those of the substrates or products of the target enzyme, but this is not essential when studying an isolated enzyme. Once the reaction has occurred, it is no longer necessary to sustain a concentration of the inactivator to retain the E-I complex because of the covalent nature of the bond formed. In some cases only a stoichiometric amount of inactivator is needed to produce complete irreversible inhibition. Irreversible inhibition is generally determined by dialysis or gel filtration of the inactivated enzyme; no enzyme recovery is expected. Two general types of irreversible enzyme inhibitors exist, affinity labeling agents and mechanism-based inactivators.

Affinity Labeling Agents

Affinity labeling agents are intrinsically reactive compounds that initially bind reversibly to the active site of the enzyme then undergo chemical reaction (generally an acylation or alkylation reaction) with a nucleophile on the enzyme (Scheme 8). To differentiate a reversible inhibitor from an irreversible one, often the dissociation constant is written with a capital I, K_I (65), instead of a small i, K_i, which is used for reversible inhibitors. The K_I denotes the concentration of an inactivator that produces half-maximal inactivation. Note that this kinetic Scheme is similar to that for substrate turnover except instead of the catalytic rate constant, k_cat, for product formation, k_inact is used to denote the maximal rate constant for inactivation.

Kinetics of Affinity Labeling Inactivation

When the equilibrium for reversible E-I complex formation (K_I) is fast relative to k_inact (the most common situation), then k_inact is the rate-determining step, and time-dependent loss of enzyme activity occurs. Under these conditions, when [I] >> [E], then Kitz and Wilson (65) described k_app by Equation 19. Two limiting situations can be considered. For [I] >> K_I, k_app = k_inact, and pseudo first-order kinetics are observed, as shown in Fig. 16a. For [I] < K_I, then k_app = k_inact/K_I, which gives simple bimolecular kinetics, in which case the straight line in Fig. 16b would pass through the origin rather than the +y-axis. To determine the K_I and k_inact values, first a plot of the log of the enzyme activity versus time is constructed (Fig. 16a). The rate of inactivation is proportional to low concentrations of the inactivator, but becomes independent at high concentrations. In these cases, the inactivator reaches enzyme saturation (just as substrate saturation occurs during catalytic turnover). Once all of the enzyme molecules are in the E-I complex, the addition of more inactivator does not affect the rate of the inactivation reaction. The half-lives for inactivation (t_{1/2}) at each inactivator concentration (lines a-e in Fig. 16a) are determined. The t_{1/2} at any inactivator concentration equals log 2/k_inact, in the limiting case of infinite inactivator concentration, t_{1/2} = 0.693/k_inact (log 2 = 0.693). A replot of these half-lives versus the inverse of the inactivator concentration, referred to as a K-I and Wilson replot, is constructed to obtain the K_I and k_inact values (Fig. 16b).

Example of an Affinity Labeling Agent

7-Aminocephalosporanic acid (15, Scheme 9) is an important intermediate in the production of many semisynthetic cephalosporin antibiotics (66, 67). However, direct deacylation of cephalosporin C (13) to 15 by cephalosporin C acylase is unfavorable, so an enzymatic process is used involving D-amino acid oxidase (DAAO) oxidation of 13 to D-glutaryl-7-aminocephalosporanic acid (14, GL-7-ACA) followed by deacylation to 15 and glutaric acid, catalyzed by GL-7-ACA acylase from Pseudomonas sp. 130 (Scheme 9) (68, 69). GL-7-ACA acylase underwent pseudo first-order time-dependent inactivation by 7-bromoacetyl aminocephalosporanic acid (16) (70). Dialysis did not regenerate enzyme activity, indicating irreversible inhibition. The rate of inactivation was lowered by the presence of either glutaric acid or 15.
Figure 16  (a) A plot showing time-dependent inactivation by affinity labeling agents and mechanism-based inactivators used for determination of kinetic constants. (b) Replot of the half-lives of inactivation from Fig. 16a versus the inverse of the inactivator concentration to determine the $K_I$ and $k_{inact}$ values for affinity labeling agents and mechanism-based inactivation.

Scheme 9  D-Amino acid oxidase (DAAO) oxidation of 13 to N-glutaryl-7-aminocephalosporanic acid (14, GL-7-ACA) followed by deacylation to 15 and glutaric acid catalyzed by GL-7-ACA acylase from Pseudomonas sp. 130. indicating that inactivation is active-site directed (competitive inhibition). The site of covalent attachment was identified as the indole ring of Trp-B4 by LC-MS peptide mapping, tandem MS analysis, and NMR spectroscopy. MALDI-TOF spectrometric analysis showed that the fourth tryptophan residue in the $\beta$-subunit (Trp-B4) was alkylated. $^{1}H-^{13}C$ HSQC NMR spectrometry of the acylase inactivated by [2-13C]-16 demonstrated that a C-C bond was formed, and it was the tryptophan residue that was alkylated. Based on other inactivations, it is reasonable that attachment is at C-2 of the indole ring of the tryptophan after rearrangement from the C-3 alkylation product (71) (17). As Trp-B4 is in the middle of an $\alpha\beta\beta\alpha$ sandwich structure, which does not have sufficient space for the inactivator to fit, it is believed that a conformational change has to occur to allow alkylation. Mutations at Trp-B4, except WB4Y, prevented covalent modification of the enzyme.
Mechanism-Based Enzyme Inactivators

Mechanism-based inactivators are unreactive compounds that bear a structural similarity to the substrate or product of a specific enzyme and are converted by the normal catalytic mechanism of that enzyme into a product that, before its release, inactivates the enzyme. These inactivators, when active, act initially as substrates for the target enzyme. Generally, the product undergoes a covalent bond reaction with the target enzyme, but it may just form a tight-binding complex that is functionally irreversible. The key features are an unreactive compound that is enzymatically converted to the actual inactivating species, which binds to the same enzyme before the inactivator’s release. Inactivation by these compounds requires knowledge of the mechanism of the enzyme. Alternatively, because these compounds act as substrates, they can be designed to test a hypothetical catalytic mechanism for an enzyme and determine if a particular type of intermediate is reasonable.

Kinetics of Mechanism-Based Inactivation

As activation of a mechanism-based inactivator requires a catalytic step, an additional intermediate (E-I) is required relative to affinity labeling agents (Scheme 10). If $k_4$ is a fast step and the equilibrium $k_{eq}$ is established rapidly, then $k_2$ is the inactivation rate constant ($k_{inact}$), which determines the maximal rate for the inactivation process leading to E-I'. As both affinity labeling agents and mechanism-based inactivators initially form reversible E-I complexes, which then lead to inactivation of the enzyme (affinity labeling agents by direct covalent reaction and mechanism-based inactivators by initial [generally] rate-determining modification to the activated species followed by rapid reaction), the kinetic constants ($K_i$ and $k_{inact}$) for both types of inactivators are determined in the same way (see Figs. 16a and 16b). The key features of mechanism-based inactivators that differentiate them from affinity labeling agents are their unreactivity and the requirement for an enzyme-catalyzed conversion to the activated species before inactivation. Often this activated species is quite reactive, so it acts as an affinity labeling agent already at the active site. However, inactivation ($k_2$ in Scheme 10) does not occur every time the inactivator is activated because the enzyme will attempt to eject it from the active site as a product ($k_3$ in Scheme 10). It would not be a favorable process if the activated species is highly reactive because it could attach to other parts of the enzyme or, if more than one enzyme was present, could attach to another enzyme. The ratio of the number of turnovers that leads to product release per inactivation event ($k_3/k_2$) is known as the partition ratio.

Example of a Mechanism-Based Inactivator

There are many examples of mechanism-based inactivators from which to choose, but I have selected an example from mine and Tilman Schirmer’s laboratories that relates to the inactivation of γ-aminobutyric acid aminotransferase (GABA-AT). This enzyme is a pyridoxal 5'-phosphate (PLP) dependent enzyme that catalyzes the conversion of the inhibitory neurotransmitter GABA to succinic semialdehyde with concomitant conversion of L-ketoglutarate to the excitatory neurotransmitter L-glutamate (Scheme 11). When the concentration of GABA in the brain diminishes, convulsions can occur. Inhibition of GABA-AT prevents the degradation of GABA, resulting in an increase in the brain GABA concentrations, which is one approach for the treatment of epilepsy. Based on the structure of 4-amino-5-fluoropentanoic acid (18), which was shown to be a very efficient mechanism-based inactivator of GABA-AT (72), (19, 35, 45)-3-amino-4-fluorocyclopentane-1-carboxylic acid (19) was designed as a potentially more lipophilic analog (73). Two inactivation mechanisms were envisioned for 19, one leading to 23 (pathway a) and one leading to 24 (pathway b) (Scheme 12). A crystal structure refined to 1.9 Å resolution of the enzyme inactivated by 19 (Fig. 17) showed explicitly that the more reasonable mechanism is that shown in pathway a to give 23 (74). Inactivator 19 is a classic example of a mechanism-based inactivator; it is an unreactive molecule that is converted by the normal catalytic mechanism of GABA-AT.

Scheme 10  Mechanism-based enzyme inactivation.
Approaches to Enzyme Inhibition

Scheme 11  Mechanism for GABA-AT.

to 20. Deprotonation leads to elimination of fluoride ion (probably via an E1cB mechanism into the PLP) to give a product (22), which after reaction with the active site lysine produces 23. Before its release from the active site, 22 undergoes an enamine reaction with Lys-bound PLP to give covalent adduct 23, the structure identified in the crystal structure. The partition ratio for this inactivator was measured by determination of the number of fluoride ions released per inactivation event to be 147; 148 molecules are turned over, one of which leads to inactivation. Therefore, the crystal structure was able to differentiate these two mechanistic possibilities. The use of crystallography to resolve mechanistic alternatives has been referred to as mechanistic crystallography.

References

1. Cheng YC, Prusoff WH. Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction. Biochem. Pharmacol. 1973;22:3099–3108.
Approaches to Enzyme Inhibition

Scheme 12 Two possible mechanisms for inactivation of GABA-AT by 19.

22. Cheng YC, Prusoff WH. Relationship between the inhibition constant (Ki) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. Biochem. Pharmacol. 1973;22:3099–3108.
Figure 17  Stereographic projection of 59 bound to G484-A7 at 1.9 Å resolution.

36. Cheng YC, Prudoff WH. Relationship between the inhibition constant (Kᵢ) and the concentration of inhibitor which causes 50% per cent inhibition (Iₕ) of an enzymatic reaction. Biochem. Pharmacol. 1971;22:3099-3108.


Further Reading


See Also

Enzyme Catalysis, Chemistry of Enzyme Kinetics, Enzyme Catalysis, Chemical Strategies for Enzyme Catalysis, Roles of Structural Dynamics in
The vitamin B₁₂ cofactor called coenzyme B₁₂ (adenosylcobalamin) assists enzymes called mutases and eliminases in the catalysis of molecular rearrangements. The mutases comprise reactions in which a substrate equilibrates with a product by migration of an amino group (as with β-lysine 5,6-aminomutase) or a carbon-based group (e.g., COSCoA in methylmalonyl-CoA mutase, which interconverts methylmalonyl-CoA with succinyl-CoA). The eliminases (e.g., propane-1,2-diol dehydratase) have either hydroxyl or amino as the migrating group, but they differ from the mutases by affording an intermediate that eliminates water or ammonia to give the observed product (e.g., propanal from propane-1,2-diol via propane-1,1-diol). Methylcobalamin is essentially an intermediary for synthetic reactions catalyzed by methyltransferases. These reactions depend on the “supernucleophility” of reduced vitamin B₁₂ (cob(II)alamin) and in humans provide for the synthesis of the amino acid methionine.

The pursuit of the “antipernicious anemia factor” seemingly ended with the publication in Science in 1948 by Karl Folkers (1906–1997) of a paper entitled “Crystalline vitamin B₁₂.” Just a few weeks later, Lester Smith (1904–1992), who had been guided by testing column fractions on pernicious anemia patients, independently obtained crystals of the vitamin [for a review of the early history of B₁₂ isolation and characterization, and references to the work described in this section see Lester Smith's monograph (1)]. He identified cobalt in B₁₂ and gave the crystals to Dorothy Hodgkin (1910–1994) that led to the structure of the isolated vitamin as cyanocobalamin (CN-Cbl: see Fig. 1 for cobalamin structures and other aspects of nomenclature of so-called corrinoids). In 1958, Horace A. (bet Barker (1907-2000) discovered coenzyme B₁₂ (adenosylcobalamin, AdoCbl); the structure determination of which, again by Dorothy Hodgkin, revealed the presence of a Co-C σ-bond. Finally, methylcobalamin (MeCbl) was recognized as another member of this very exclusive club of natural organometallic compounds. In this article, we review the fundamental chemistry of the B₁₂ cofactors in selected enzymatic reactions for which they are obligatory participants.

**Biological Context**

**Human cobalamins and dietary requirements**

Quantitative assays have shown that the plasma of nonsmoking, healthy adults contains MeCbl (250 pg cm⁻³), AdoCbl, and hydroxocobalamin (OH-Cbl) [AdoCbl + OH-Cbl (125 pg cm⁻³ with AdoCbl] of total cobalamin), which is derived from hydrogen cyanide in tobacco smoke. CN-Cbl can also develop from the consumption of foods (e.g., cassava) that release cyanide ions. In erythrocytes, the chief cobalamin is AdoCbl (>50% of total cobalamin) followed by OH-Cbl. The blood of smokers contains CN-Cbl (~2% of total cobalamin), whereas CN-Cbl has no established role in humans. The total human body store of the corrinoids described is ~5 mg, and the recommended daily requirement of corrinoids is ~2 µg; this nutrient is provided by a typical “Western” diet but may not be included in a vegan diet. Whether 2 µg per day is sufficient to maintain health and protect against disease, especially of a degenerative kind, is a subject of current debate in which some believe it wise to consume larger amounts of
ery?” We know that coenzyme B12 assists a group of enzymes whose mechanisms of action could begin to be understood; today one may (Eq. 1, 2). Corrinoids are sequestered from food sources by a glycoprotein of mass ∼45 kDa called “intrinsic factor,” which is secreted in the stomach and binds B12 derivatives very tightly to form a 1:1 complex (for CN-Cbl, K = 1.5 × 10^10 mol⁻¹ dm³). Several other proteins bind and transport B12 into cells (3). The disease pernicious anemia has been recognized since the early nineteenth century and linked to a deficiency of what William Castle called “extrinsic factor” (i.e., vitamin B12) in 1928 (1). This disease develops because of the failure of the patient to secrete sufficient intrinsic factor (i.e., hydroxocobalamin (OH-Cbl); R = CN denotes cyanocobalamin (CN-Cbl), whilst R = Me a methylcobalamin (MeCbl). By definition all cobalamins contain 5,6-dimethylbenzimidazole, which is the so-called 6th ligand to cobalt in an evolutionary origin, whereas the eliminases evolved separately. Possibly the mutases and methyltransferases have a common origin, whereas the eliminases evolved separately. N-methylcobalamin is completely different from adenosylcobalamin because it is essentially a conduit for synthetic reactions catalyzed by methyltransferases, illustrated in Scheme 2 for the case of methionine. These reactions depend on the “supernucleophilicity” of cob(I)alamin. In one case, this species removes a methyl group from N⁵-methyltetrahydrofolate with the formation of methylcobalamin, and then transfers this group to the acceptor homocysteine, which results in the synthesis of methionine (see Scheme 2). Examples of mutases, eliminases, and methyltransferases are discussed in detail below.

**Coenzyme B12: Nature’s Most Complex Cofactor**

Coenzyme B12 (adenosylcobalamin, Fig. 1) has the most complex structure of all of the cofactors used by nature to aid enzymatic catalysis. Although many unusual features are observed in the structure, it is the cobalt–carbon sigma (σ) bond that is vital, above all, for the mode of action of the coenzyme. The bond dissociation energy (BDE) of ∼130 kJ mol⁻¹ (6) is typical for a metal–carbon bond, although this value is much lower than for σ-bonds from carbon to hydrogen, nitrogen or oxygen, or another carbon (BDE ∼ 350 kJ mol⁻¹), the coenzyme is rather stable in water provided light is excluded. The

---

**Figure 1** In the above structure, R = CN denotes cyanocobalamin (CN-Cbl), whilst R = OH is hydroxocobalamin (OH-Cbl). R = Me denotes methylcobalamin (MeCbl). By definition all cobalamins contain 5,6-dimethylbenzimidazole, which is the so-called 6th ligand to cobalt in the above structure. Substances containing the corrin ligand, i.e. the planar 14 electron p-system embracing cobalt in the above structure, are also called corrinoids.

**Overview of cobalamin-dependent reactions**

With the identification of vitamin B12’s cofactors, their mechanisms of action could begin to be understood; today one may ask, “What do we know some 50 years after their discovery?” We know that coenzyme B12 assists a group of enzymes (Table 1, entries 1–9) in the catalysis of molecular rearrangements, which can all be described by Scheme 1. In Table 1, these enzymes are subdivided into mutases and eliminases. The mutases catalyze equilibrations and comprise reactions in which a carbon skeleton rearranges (entries 1–4, sometimes called Class I reactions; substituted carbon atom as migrating group) as well as the amino mutases (entries 5 and 6, Class II reactions; amino as migrating group) (4). The eliminases (entries 7–9, Class II reactions (4)) have either hydroxyl or amino as the migrating group, but they differ from the mutases in that the rearrangement affords an intermediate that eliminates water or ammonia to give the observed product. Coenzyme B12-dependent ribonucleotide reductase (entry 10) is included in Table 1 because its catalytic reaction exhibits some features of the diol and glycerol dehydratase reactions (i.e., elimination of OH from a 1,2-diol moiety), although the overall reaction is a redox process and not a rearrangement. How the reactions of Table 1 occur remains in part a puzzle, but this review intends to enlighten the reader. The coenzyme B12-dependent mutases and eliminases are a distinctive class of so-called “radical enzymes” (Fig. 2) (5), which are unique among enzymes because their catalytic pathways function via species with an unpaired electron (i.e., radicals). These species are normally highly reactive, but containment within a protein “stranglehold” enables their reactivity to be tamed and harnessed without destructive side reactions. The term “negative catalysis” was coined by János Rétyi in 1990 to describe this phenomenon.

An important structural difference between eliminases and mutases concerns the axial base coordinated to the cobalt, perpendicular to the plane of the corrin ring. Whereas in all the eliminases, the axial base is the dimethylbenzimidazole of the coenzyme itself (Fig. 1), the mutases use a conserved histidine residue of the enzyme for this purpose. On binding of the coenzyme to the apo-enzyme, the axial base is replaced by the histidine and moves into a distinct pocket of the protein. Possibly the mutases and methyltransferases have a common origin, whereas the eliminases evolved separately.

Methyltransferases (see below) also use a protein histidine as axial base, whose reactivity is fine tuned by protonation.
### Table 1 Classification of Coenzyme B₁₂-dependent Radical Enzymes

<table>
<thead>
<tr>
<th>Type</th>
<th>Enzyme</th>
<th>Reaction Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon skeleton mutases</td>
<td></td>
<td>Reversible reactions that rearrange molecular skeletons (X is a carbon-based group in the carbon skeleton mutases and NH₂ in the amino mutases; X is shown in red below)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Glutamate mutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Isobutyryl-CoA mutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. 2-Methylene-glutarate mutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Methyl-malonyl-CoA mutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino mutases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. β-Lysine 5,6-aminomutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Ornithine 4,5-aminomutase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eliminases (&quot;base-on&quot;)</td>
<td></td>
<td>Irreversible reactions that eliminate water (or ammonia) [X is shown in red below]</td>
</tr>
</tbody>
</table>

**Carbon skeleton mutases ("base-off, his-on")**

Reversible reactions that rearrange molecular skeletons (X is a carbon-based group in the carbon skeleton mutases and NH₂ in the amino mutases; X is shown in red below)

**Amino mutases**

Irreversible reactions that eliminate water (or ammonia) [X is shown in red below]
rate constant for decomposition of coenzyme B12 at 30°C in water was determined to be $1.16 \times 10^{-8} \text{ s}^{-1}$, which corresponds to a half-life of 1.9 years (4). The power of the primary organic radical $5'$-deoxyadenosyl that develops from homolysis of the coenzyme's Co-C bond is only unleashed when the coenzyme is bound to a partner enzyme in the presence of a substrate molecule. The $5'$-deoxyadenosyl radical initiates the group of reactions summarized by Scheme 1 (4). In these reactions, a protein-bound radical is initially formed from a substrate molecule. The basic pathway requires at least two participating radicals, with one of these ($S^\bullet$) being structurally related to the substrate (SH) and the other ($P^\bullet$) structurally related to the product (PH). In addition, there may be an intermediate radical ($I^\bullet$) in some, if not all, cases. Homolysis of coenzyme B12 also releases cob(II)alamin, a paramagnetic ($d^7$) species, which is easily detected by electron spin resonance (EPR) spectroscopy of reaction mixtures created by incubating enzyme, coenzyme, and substrate molecules for a few seconds and then freezing in liquid nitrogen.

Coenzyme B12-Dependent Mutases

A common feature of the six known mutases is that a hydrogen atom has to be reversibly abstracted by the $5'$-deoxyadenosyl radical from a nonactivated methyl group to yield a methylene radical, which is $S^\bullet$ or $P^\bullet$ depending on the identity of the substrate and product and is not stabilized by any adjacent group. However, the partner radical is always stabilized by a neighboring group, which is a carboxylate or CoA-ester for entries 1-5 and methyl for entry 6.

Glutamate mutase: the first coenzyme B12-dependent mutase

Glutamate mutase was discovered by Barker, who showed that the enzyme catalyzes the equilibration of (S)-glutamate with (2S,3S)-3-methylaspartate (Entry 1, Table 1; $\Delta G^{\circ} = +6.3$ kJ mol$^{-1}$; $K = 0.095$) (for a review of glutamate mutase see Reference 7). This reaction is one of three distinct methods that nature uses to ferment glutamate to butyrate (5). Surprisingly, they all proceed through intermediate radicals. The coenzyme B12-dependent fermentation is the only one of the three that

\[ \text{Coenzyme B12-dependent enzyme-catalyzed rearrangements (for examples of a, b, X, and Y see Table 1).} \]
The reason is because the C-3 protons of glutamate have an hydrogen atom of glutamate to give 5′-deoxyadenosine to afford 3-methylaspartate (cf. pathways for 2-methyleneglutarate mutase, Scheme 5). The most convincing experimental evidence for this “fragmentation-recombination” mechanism was the isolation of similar amounts of acrylate and glycine (−0.06 mol of each per mol of enzyme) when the working enzyme was interrupted by addition of trifluoroacetic acid (9).

To achieve its catalytic reaction, glutamate mutase faces the problem of how to surmount the two relatively high transition-state energy barriers that lead from the 4-glutamyl radical to acrylate and the glycyl radical and then the recombination of these radicals to the 3-methyl-aspartate radical (Scheme 4). These barriers were computed as ∆H°+ = +59.9 and +66.5 kJ mol−1 (10), respectively, and the 3-methylene-aspartate radical was found to be significantly less stable than the resonance-stabilized 4-glutamyl radical (∆H°+ = +20.3 kJ mol−1). Likewise, with methylmalonyl-CoA mutase and 2-methyleneglutarate mutase the substrate and product radicals are interconnected via transition states of relatively high energy. In each case, the intermediate methylene radical (P• for glutamate and 2-methyleneglutarate mutase, S• for methylmalonyl-CoA mutase) is much less stable than the partner methine radical, which is stabilized by a carboxylate or CoA-ester group. It was recently suggested that stabilization of the methylene radicals by cob(II)alamin in the role of “conductor” might be required to lower the activation energy for the process of radical exchange (11). However, this proposal lacks experimental support.

2-Methyleneglutarate mutase: glutamate mutase in disguise?

The equilibration catalyzed by 2-methyleneglutarate mutase is similar to the glutamate mutase reaction but with a methylene group (−CH2) in 2-methyleneglutarate/3-methylitaconate in place of the amino center of glutamate/3-methylaspartate. However, this feature enables a different mechanism (addition-elimination, Scheme 5, path a) to be followed, although an analogous mechanism (fragmentation-recombination, Scheme 5, path b) to that of glutamate mutase is also possible in principle.
In all known mutases, but not in the
below), which was taken to suggest the participation of cob(II)
which enters the Krebs cycle. Methylmalonyl-CoA mutase was
ular orbital calculations invoked a “partial protonation” of the
mechanism for the catalytic reaction based on
framework was demonstrated by model studies (15).
The partial protonation was supposed to be provided by the hy-
radicals is ∼ 6 Å (cf. ∼ 11 Å in eliminases—see
In mutases, the separation between cob(II)alamin and the in-
In all known mutases, but not in the
eliminases, a methylene radical may need to be stabilized, pos-
sibly by interaction of the d12-orbital of cob(II)alamin with the
disappearing p-orbital of the 5-deoxyadenosyl radical and the
The mechanism of path a and possibly that of path b are consis-
tent with the results of a study in which 2-methylene glutarate
was shown to catalyze the equilibration of (Z)-3-methyl(1Z,2H1)
itaconate with an equal quantity of its (E)-isomer, as well as with
2.1 mixture of (E) and (Z) 2-methylene 2- H1 glutarate (12). Newcomb and Miranda (13) performed model studies in
which radicals that correspond to S•, P•, and I• (Scheme 5)
were generated, and their rates of interconversion were mea-
sured. It was concluded that these rates were too slow by a factor
of ∼ 103 relative to the known enzymatic kcat (30 s−1) for
either of the mechanisms of Scheme 5 to be kinetically plausi-
bile. To explain this discrepancy, there could be a role for cob(II)
alamin in the rearrangement as mentioned above and discussed
below.

Methylmalonyl-CoA mutase
This enzyme’s role in humans is to assist the detoxification
of propionate derived from the degradation of the amino acids
methionine, threonine, valine, and isoleucine. Propionyl-CoA is
carboxylated to (S)-methylmalonyl-CoA, which is epimerized to
the (R)-isomer. Coenzyme B12-dependent methylmalonyl-
CoA mutase isomerizes the latter to succinyl-CoA (Fig. 2),
which enters the Krebs cycle. Methylmalonyl-CoA mutase was
the first coenzyme B12-dependent enzyme to be characterized
crystallographically (by Philip Evans and Peter Leddlay). A
mechanism for the catalytic reaction based on ab initio molec-
ular orbital calculations invoked a “partial protonation” of the
oxygen atom of the substrate thioester carbonyl group that fa-
cilitated formation of an oxycyclopropyl intermediate, which
connects the substrate-derived and product-related radicals (14).
The partial protonation was supposed to be provided by the hy-
drogen bonding of this carbonyl to His 244, which was inferred
from the crystal structure of the protein. The ability of the sub-
strate and product radicals to interconvert even in the absence
of the enzyme was demonstrated by model studies (15).
The role of Cob(II)alamin in mutases
In mutases, the separation between cob(II)alamin and the in-
termediate radicals is ∼ 6 Å (cf. ∼ 11 Å in eliminases—see
below), which was taken to suggest the participation of cob(II)
alamin during turnover. In all known mutases, but not in the
Unusually high isotope effects have been observed for several
coaenzyme B12-dependent reactions. For methylmalonyl-CoA
mutase acting on (CD3)methylmalonyl-CoA, a primary deu-
terium isotope effect of 35.6 at 20 °C was recorded (16).
This effect was ascribed to quantum tunneling in the transi-
tion state for hydrogen (or deuterium) atom abstraction by the
5-deoxyadenosyl radical. To probe the nature of the hydro-
gen abstraction steps with glutamate, mutase experiments were
performed using labeled coenzyme, substrate, or product (17).
The results were interpreted as providing evidence for hydrogen
tunneling and a coupled motion of the hydrogen atoms at the
adenosyl C-5’ with the hydrogen atom being transferred from
substrate to product.

Coenzyme B12-Dependent Eliminases
Diol and glycerol dehydratase
Soon after the discovery of glutamate mutase, Robert Abeles
(1926-2000) recognized that the apparently simple conver-
sion of glycerol to 3-hydroxypropionaldehyde catalyzed by
diol dehydratase, was a coenzyme B12-dependent process (18).
It had been long been known that acrolein could develop
from the fermentation of glycerol. During the production of
whiskey, infection of the broth may lead to the production of
3-hydroxypropionaldehyde and hence to acrolein on distillation.
Isotopic labeling studies with glycerol dehydratase showed a re-
markable control of the movement of the oxygen atoms in the
rearrangement of 1,2-ethanediol to acetaldehyde and glycerol
of cob(II)alamin during turnover. In all known mutases, but not in the
below), which was taken to suggest the participation of cob(II)
alamin in the rearrangement as mentioned above and discussed
below. To explain this discrepancy, there could be a role for cob(II)
alamin in the rearrangement as mentioned above and discussed
below.

Methylmalonyl-CoA mutase
This enzyme’s role in humans is to assist the detoxification
of propionate derived from the degradation of the amino acids
methionine, threonine, valine, and isoleucine. Propionyl-CoA is
carboxylated to (S)-methylmalonyl-CoA, which is epimerized to
the (R)-isomer. Coenzyme B12-dependent methylmalonyl-
CoA mutase isomerizes the latter to succinyl-CoA (Fig. 2),
which enters the Krebs cycle. Methylmalonyl-CoA mutase was
the first coenzyme B12-dependent enzyme to be characterized
crystallographically (by Philip Evans and Peter Leddlay). A
mechanism for the catalytic reaction based on ab initio molec-
ular orbital calculations invoked a “partial protonation” of the
oxygen atom of the substrate thioester carbonyl group that fa-
cilitated formation of an oxycyclopropyl intermediate, which
connects the substrate-derived and product-related radicals (14).
The partial protonation was supposed to be provided by the hy-
drogen bonding of this carbonyl to His 244, which was inferred
from the crystal structure of the protein. The ability of the sub-
strate and product radicals to interconvert even in the absence
of the enzyme was demonstrated by model studies (15).
The role of Cob(II)alamin in mutases
In mutases, the separation between cob(II)alamin and the in-
termediate radicals is ∼ 6 Å (cf. ∼ 11 Å in eliminases—see
below), which was taken to suggest the participation of cob(II)
alamin during turnover. In all known mutases, but not in the

Unusually high isotope effects have been observed for several
coaenzyme B12-dependent reactions. For methylmalonyl-CoA
mutase acting on (CD3)methylmalonyl-CoA, a primary deu-
terium isotope effect of 35.6 at 20 °C was recorded (16).
This effect was ascribed to quantum tunneling in the transi-
tion state for hydrogen (or deuterium) atom abstraction by the
5-deoxyadenosyl radical. To probe the nature of the hydro-
gen abstraction steps with glutamate, mutase experiments were
performed using labeled coenzyme, substrate, or product (17).
The results were interpreted as providing evidence for hydrogen
tunneling and a coupled motion of the hydrogen atoms at the
adenosyl C-5’ with the hydrogen atom being transferred from
substrate to product.

Coenzyme B12-Dependent Eliminases
Diol and glycerol dehydratase
Soon after the discovery of glutamate mutase, Robert Abeles
(1926-2000) recognized that the apparently simple conver-
sion of glycerol to 3-hydroxypropionaldehyde catalyzed by
diol dehydratase, was a coenzyme B12-dependent process (18).
It had been long been known that acrolein could develop
from the fermentation of glycerol. During the production of
whiskey, infection of the broth may lead to the production of
3-hydroxypropionaldehyde and hence to acrolein on distillation.
Isotopic labeling studies with glycerol dehydratase showed a re-
markable control of the movement of the oxygen atoms in the
rearrangement of 1,2-ethanediol to acetaldehyde and glycerol
of cob(II)alamin during turnover. In all known mutases, but not in the

contains a glycyl radical formed by the action of the anaerobic bacterium Clostridium butyricum, the substrate-derived radical intermediate(s).

Recently, a glycerol dehydratase was discovered in the man’s B12 by Barker. Hence, this glycerol dehydratase per...
Methionine synthase

The crucial steps in the pathway catalyzed by human cytosolic enzyme methionine synthase (MetH) are the transfer of a methyl group, first from N1-methyltetrahydrofolate to the cobalt of cob(I)alamin to give methylcobalamin (MeCbl), and then from cobalt to the sulfur of homocysteine to give methionine (Scheme 2). Our knowledge of MetH is largely from the incisive experiments of Liptak et al. (23), with crystallographic characterization of the enzyme by Martha Ludwig and Catherine Drennan underpinning all else. One large polypeptide (1227 amino acids, 136 kDa) comprises all functions of MetH, which has four modules. The substrates N5-methyltetrahydrofolate and homocysteine bind to a module each. The cobalamin cofactor binds to a third module with its dimethylbenzimidazole replaced by His759. The function of the fourth module is to reactivate the oxidized cobalamin to MeCbl using SAM as source of the methyl group and flavodoxin as reductant (see below). Profound conformational changes bring the reacting components together in turn.

The key intermediate in the catalytic pathway is the “supernucleophile” cob(I)alamin, which attacks N5-methylthetrahydrofolate, generating tetrahydrofolate and MeCbl. Then homocysteine (probably as its thiolate) attacks MeCbl to regenerate cob(I)alamin and methylcobalamin (Scheme 2). The demethylation of N5-methyltetrahydrofolate is not trivial, even for the “supernucleophilic” cob(I)alamin, and considerable efforts have been invested into understanding this reaction, dubbed “improbable” by Duilio A. Togni. The obvious mode of activation is by proton transfer to N5 of N5-methyltetrahydrofolate, but as this is weakly basic (pKa 5.1) the nature of the proton source and mode of transfer has been difficult to pin down. Recent research from the Matthews group has shown that the reactivities of cob(I)alamin and methylcobalamin are modulated by the ligand trans to the lone pair of cob(I)alamin and methyl group of methylcobalamin (21).

It is interesting that E. coli contains two genes that code for methionine synthase: metH for the cobalamin-dependent enzyme and metE for a cobalamin-independent enzyme that depends on an active site Zn2+ to stabilize deprotonated homocysteine (24). This thiolate species demethylates N5-methyltetrahydrofolate, which is activated by proton transfer to N5 of N5-methyltetrahydrofolate, and so in the absence of E. coli it produces much more MetE to compensate for the lack of MetH.

Cobalamin-dependent methionine synthase contains a built-in repair mechanism. If accidental oxidation of cob(I)alamin leads to inactive cob(I)alamin, then the enzyme employs SAM and reduced flavodoxin to regenerate cob(I)alamin. Although the redox equilibrium below lies mainly on the left side, any cob(I)alamin formed is trapped by SAM-dependent methylation to yield methylcobalamin.

\[
\text{Cob(I)alamin} + \text{flavodoxin semiquinone} \rightarrow \text{Cob(I)alamin} + \text{flavodoxin hydroquinone} + \text{methylcobalamin}
\]

The anodic nitrous oxide (N2O) oxidizes E. coli by reacting with cob(I)alamin, which probably yields reactive hydroxyl radicals that damage the MetH protein (25). Model studies have shown that N2O reacts with cob(I)alamin, but not cob(II)alamin or cob(III)alamin complexes, which affords dinitrogen (N2) and hydroxocobalamin. Repeated anesthesia with N2O over a few days can be life threatening because the production of methionine is suppressed, and more seriously, turnover of folate cofactors stops because folates are trapped as N5-methyltetrahydrofolate (“methyl trap hypothesis”). This mechanism leads to inhibition of DNA and protein synthesis, and hence cell death.

Other methyltransferases

Coralidoxanthins participate in the global C1 carbon cycle through the synthesis of methane and acetyl-CoA. Methane formation by methanogenic archaea exhibits a cobalamin-dependent step in which an overall methyl transfer occurs from N5-methyltetrahydrodromethanopterin to coenzyme M (26). This methionine synthase-like process is exergonic (\(\Delta G° = -30 \text{ kJ mol}^{-1}\)) and is catalyzed by a multienzyme complex that comprises eight different subunits. The subunit MtrA binds a cob(II)amide cofactor, which reacts with N5-methyltetrahydrodromethanopterin during the catalytic cycle to give a methylcobamide. This reaction undergoes an Na+-dependent demethylation by coenzyme M. The methylcobamide has a formal Cob(III) center with the cobinamide in a base-off/His-on state, whereas the cob(I)amide is in the four-coordinate His-off state. The conformational change between these two states may drive the Na+-pump of a Na+-translocating, membrane-associated process.

The synthesis of acetyl-CoA by the Jorgensen-Wood path- way of autotrophic carbon fixation in diverse bacteria and archaea is catalyzed by a Co- and Fe-containing corrinoid iron-sulfur protein (CofFeSP). This protein participates in the transfer of a methyl group from N5-methyltetrahydrofolate to the cob(II)amide of CofFeSP to give a methylcob(III)amide, of which an overall methyl transfer occurs from CofFeSP to coenzyme M (27).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

The astonishing reactivity of cob(I)alamin is exploited in several other demethylation reactions, for example, the transfer of the methyl group of methanol to 2-mercaptoethanesulfonic acid (coenzyme M) catalyzed by methanol-2-mercaptoethanesulfonic acid methyltransferase. In this case, protein crystallography has shown that the methanol molecule is activated for nucleophilic attack by cob(I)alamin by coordination of its hydroxy group to a Zn2+ (28). Even more remarkable is the ability of cob(I)alamin to affect the dechlorination of vinyl halides, for example, tetrachloroethene to tetrahydrofuran, generating tetrahydrofuran and MeCbl. Then yields methionine and regenerates cob(I)alamin (29).

References


2. Ragsdale SW. Metals and their scaffolds to promote difficult enzymatic... 2007. Taylor & Francis, Boca Raton, FL.


Further Reading

See Also
Amino Acids Enzyme Catalysis, Chemical Strategies for Enzyme Catalysis, Roles of Structural Dynamics in Enzymatic Cofactors
Vitamins.
Biosynthesis of Nonribosomal Peptides

Georg Schoenafinger, Philipps-University of Marburg, FB Chemie/Biochemie, Marburg, Germany

Mohamed A Marahiel, Philipps-University of Marburg, FB Chemie/Biochemie, Marburg, Germany
doi: 10.1002/9780470048672.wecb398

Many microorganisms have evolved an unusual way of producing secondary peptide metabolites. Large multidomain enzymatic machineries, the so-called nonribosomal peptide synthetases (NRPSs), are responsible for the production of this structurally diverse class of peptides with various functions, such as cytostatic, immunosuppressive, antibacterial, or antitumor properties. These secondary metabolites differ from peptides of ribosomal origin in several ways. Their length is limited to a mere 20 building blocks, roughly, and mostly a circular or branched cyclic connectivity is found. Furthermore, aside from the proteinogenic amino acids, a larger variety of chemical groups is found in these bioactive compounds: D-configured amino acids, fatty acids, methylated, oxidized, halogenated, and glycosylated building blocks. These functional and structural features are known to be important for bioactivity, and often natural defense mechanisms are thus evaded. In this article, we describe the enzymatic machineries of NRPSs, the chemical reactions catalyzed by their subunits, and the potential of redesigning or using these machineries to give rise to new nonribosomal peptide antibiotics.
Biologic Background

Nonribosomal peptides are produced by a large number of bacteria, fungi, and lower eucaryotes. For most of these compounds, their biologic role is unknown. One might suspect that these secreted molecules are used for unknown forms of communication or simply to critically increase the chance of survival for the producing cell in its habitat, because the metabolic cost of their production is enormous. However, the function of some nonribosomal compounds has been identified: The well-studied penicillin produced by Penicillium notatum, for instance, is a weapon against nutrient competitors, and the siderophore bacillibactin helps its producer, Bacillus subtilis, acquire iron and thereby prevent iron starvation. For us, natural products produced by microorganisms attract considerable attention because they observed bioactivities range from antibiotic to immunosuppressive and from cytotoxic to antitumor. Not only have these secondary metabolites been optimized for their dedicated function over millions of years of evolution, but they also represent promising scaffolds for the development of novel drugs with improved or altered activities.

Catalytic Domains of Nonribosomal Peptide Synthetases

The catalytically active entities that NRPSs are composed of can be classified as being essential to all NRPSs or being responsible for special modifications. Only when a set of domains correctly acts in appropriate order, the designated product can be synthesized (3) (Fig. 3). The function, chemistry, and interactions of these domains are discussed in the following section (Fig. 4).

Essential domains of NRPSs

Before any peptide formation can occur, the amino acids or, generally speaking, the building blocks to be condensed need to be recognized and activated (12). The adenylation (A) domains are responsible for special modifications. Only when a set of domains correctly acts in appropriate order, the designated product can be synthesized (3) (Fig. 3). The function, chemistry, and interactions of these domains are discussed in the following section (Fig. 4).

(a) Gene → mRNA: Codons → Ribosomal Protein → Nonribosomal Peptide
(b) Gene → NRPS: Modules → Enzymatic Machinery → Peptide

Figure 1 Comparison of ribosomal and nonribosomal peptide synthesis. (a) In the ribosomal information pathway, the sequence of codons in the mRNA determines the sequence of amino acids in the protein, whereas (b) the sequence of modules in the nonribosomal peptide synthetase intrinsically determines the primary sequence of the peptide product.
Figure 2. A selection of nonribosomal peptides. Chemical and structural features that contribute to the vast diversity of this class of metabolites are highlighted: Heterocycle (bacitracin), lactone (surfactin, daptomycin), ornithine and lactam (Tyrocidine), sugar, chlorinated aromatics, C–C crosslink (Vancomycin), N-formyl groups (Coelichelin and linear gramicidin), fatty acid (daptomycin), dihydroxybenzoate and trimeric organization (bacillibactin), dimeric organization (gramicidin S), and ethanolamine (linear gramicidin).
translocation model implies that the biosynthesis is linear—altogether dependent on delicate, situationally changing affinities that guarantee correct timing for each reaction and that prevents side reactions (Fig. 4). Even though this model successfully puts the biosynthetic enzymes in relation with their products for most known NRPS systems, some exceptions are known: The structures of syringomycin (16) or coelichelin (17) cannot be sufficiently explained by merely deciphering the build-up of their NRPSs when using this model. Obviously, other regulatory mechanisms and forms of inter-domain communication are not yet fully understood.

When the last condensation reaction has occurred, the linear precursor needs to be released from the enzyme. For this important last step, several mechanisms are known: Simple hydrolysis of the thioester (bacitracin, vancomycin), intramolecular cyclization leading to a lactam (tyrocidine, bacitracin) or a lactone (surfactin), or even reductive thioester cleavage (linear gramicidin). In some cases, the linear precursor is dimerized (gramicidin S) or even trimerized (bacillibactin, enterobactin) before cyclization (Fig. 2). Even though these reactions are critical for the compound’s bioactivity, the catalytic domains responsible for the release are not found in all NRPS systems and will therefore be called “modifying” domains.

Modifying domains of NRPSs

Apart from the essential domains in NRPSs, several so-called modifying domains are not found in every NRPS system. Nevertheless, they are required for proper processing of their designated substrate within their synthetase. Deletion or inactivation of these modifying domains usually results in the production of compounds with bioactivities severely reduced or altogether abolished.

Most nonribosomal peptides have a cyclic connectivity. In these cases, a C-terminal so-called thioesterase (TE) domain, is often found in the synthetase. These TE domains all share an invariant serine residue belonging to a catalytic triad (Asp-His-Ser), which is known to be acylated with the linear peptide before cyclization (18). Once the substrate is translocated from the PCP domain onto the TE domain, the regiospecific and stereospecific intramolecular attack of a nucleophile onto the C-terminal carbonyl group of the substrate is directed by the enzyme. This nucleophile can be the N-terminal α-amino group of the linear peptide (tyrocidine, gramicidin S), a side-chain amino (bacitracin) or hydroxyl group (surfactin). Since the ester bond between the substrate and the TE domain is cleaved by these cyclization reactions, the resulting lactams or lactones are released from the synthetic machinery by this step. In a few cases, the modular arrangement of NRPSs suggests that only one half (gramicidin S) or one third (bacillibactin, enterobactin) of the extracted peptide product can be produced by one assembly line-like synthesis (Fig. 2). These synthetases are considered iterative (19) because they have to complete more than one linear peptide synthesis before one molecule of the secondary metabolite can be released. According to a proposed model, the first precursor is translocated onto the TE domain, the second monomer is then produced and transferred to the TE domain-bound first monomer leading to a dimer. An analogous trimerization occurs—if applicable—and finally the product is released by cyclization.

Another modifying reaction that is commonly found in NRPSs is the epimerization (E) of an amino acid (5). E domains that are always situated directly downstream of a PCP domain catalyse these reactions. The most C-terminal amino acid of the reaction intermediate is racemized by an E domain, no matter whether the substrate is an aminoacyl group alone or a peptidyl group. The mechanism of these E domains is so far unclear, even though a catalysis that involves one or more catalytic bases to deprotonate the α-carbon atom as a first step seems likely. The resulting planar double-bond species then needs to be reprototated from the other side to invert the absolute configuration of the building block. This result can be accomplished by a nearby protonated catalytic base in the enzyme or water, which is positioned opposite of the first catalytic base. Nevertheless, a mixture of both stereoisomers always can be detected when the substrate bound to the enzyme is analyzed, which is indicative for either a non stereospecific or a reversible reaction. Once the epimerized substrate undergoes the subsequent condensation reaction, only the species with an inverted stereocenter is found.
Biosynthesis of Nonribosomal Peptides

Figure 4 Single reactions in NRPSs and their timing. (a) After ribosomal synthesis of the apo-enzymes, the PCP domains are postsynthetically modified with 4′-Phosphopantetheine cofactors by a 4′Ppan transferase, e.g., Sfp. (b) In a second step, the A domains bind their cognate substrates as well as ATP and form the corresponding acyl-adenylate intermediates. These are transferred onto the cofactor of the neighboring PCP domains. (c) The C domains catalyse the condensation of two building blocks. The specificities of C domains and the affinities of aminoacyl-/peptidyl-PCP domains ensure that no internal start reactions occur. (d) Only after the first condensation domain has acted does the second C domain seem to process the intermediate. During synthesis, the growing product chain is continuously translocated toward the C-terminal end of the enzyme.

Techniques for the Production of Novel Nonribosomal Peptides

With a constantly growing number of pathogenic bacterial strains resistant to the known antibiotics, the demand for novel antibiotics or, more generally speaking, therapeutic agents is evident. Because many NRPS products already have such activities and their chemical and structural diversity is so huge, efforts have been made to use NRPSs to broaden the known spectrum of therapeutics. In this section, the possibilities of using NRPS machineries or parts of them to produce new bioactive compounds are addressed.
Module or domain exchange

When considering the modular buildup of NRPSs, the possibility of altering the peptide product by insertion, deletion, or exchange of modules seems to be an obvious approach for the production of new compounds. Because the A domains determine the specificity of each module, even an exchange of fractions smaller than whole modules in a synthetase could lead to an altered product. In the past, various attempts have succeeded using these strategies (24, 25). For instance, the exchange of an A domain in the surfactin NRPS with other A domains of both bacterial and fungal origin lead to the formation of the expected variants of surfactin (26). However, in all of these early studies, the apparent turnover rates were significantly lower than in the wild-type systems. Accordingly, two explanations for the drastically slowed down synthetic process can be given. First, the borders chosen to dissect and to fuse the catalytic domains might have been unsuitable. Even though the reoccurring, highly variable so-called linker regions between each pair of catalytic domains seem not to exhibit secondary structures, their sequence and length might be critical for proper inter-domain communication. So far, no structure of any enzyme consisting of two or more NRPS domains has been published, which makes it difficult to define the right domain border when preparing a cloning strategy for fusion or for dissection. Second, the specificity of the C domains might result in a reduced product turnover. Even though a relaxed specificity for the donor substrate has been reported, the acceptor site seems to be highly specific, which discriminates against artificial substrates (27). Both the mode of catalytic action and the molecular and structural basis for the selectivity are not fully understood for C domains so that a straightforward and approach for overcoming these low turnover rates currently cannot be given.

Changing the specificity code for A domains

Sequence alignments of A domains have revealed that domains activating the same type of building block share a set of conserved residues in the primary protein sequence (13). With the A domain’s crystal structure, one can find that these residues form the substrate binding pocket (28). These residues are therefore referred to as the “selectivity-conferring code” of NRPS (13). One can now rationally exchange these sets of residues and can obtain fully functional A domains with altered substrate recognition. For example, this process has been done for the first module of the surfactin synthetase srfA (31). A major advantage of using these strategies is their regiospecificity, stereospecificity, and chemoselectivity (29). Thus, no protective groups are needed during these reactions is their regiospecificity, stereospecificity, and chemo- selectivity. In many different cases, the peptide product is produced very quickly. The reason why one can use such peptides as substrates lies in the relaxed substrate specificity of many TE domains. The TE domain of the Tyrocidine synthetase, which carries out a head-to-tail cyclization of the decapeptide DPhe-Pro-Phe-DPhe-Asn-Gln-Tyr-Val-Orn-Leu, for instance, only recognizes the two N-terminal and C-terminal amino acids of the natural substrate. The side chains of the amino acids at other positions are not recognized by the enzyme, and experiments with substrates that carry substitutions to alarinate these positions still lead to analogous cyclodecamers (31). The major advantage of using TE domains for cyclization reactions is the regiospecificity, stereospecificity, and chemoselectivity. Thus, no protective groups are needed during these enzymatic cyclization reactions, and undesired side product formation is minimized. Additionally, these reactions are carried out under mild aqueous conditions, usually pH 7-8.

To follow this chemoenzymatic approach, the synthetic substrates must be transferred onto the catalytically active serine residue of the TE domain. This transfer can either be done directly or with the help of a PCP domain. In the natural system, translocation is realized by the interaction between the PCP and the TE domain. The substrate, which is bound to the 4Ppan cofactor of the PCP domain as a thioester, acylates the hydroxy group of the serine. Chemically speaking, the acylation of the TE is a result of a trans-esterification. When using TE domains, the substrate provided in trans must also have an appropriate acylation potential. Several key techniques have been developed to covalently attach synthetic substrates to PCP and TE domains. In the first method, the relaxed substrate specificity of the 4Ppan transferase Sfp is used to load aryl moieties are predicted to be very low. As discussed, the C domains that have to process the artificial substrates are predicted to discriminate against non-natural substrates, which kinetically impede product formation drastically. On the other hand, this method is limited to the building blocks that other known A domains activate. Yet, the vast diversity and bioactivity of nonribosomal peptides mainly arises from their unusual connectivities and a large number of post-synthetic modifications, which one cannot address when merely changing the A domains’ specificities.

Chemoenzymatic approaches

A very powerful method for producing novel antibiotics is the chemoenzymatic approach (30). The idea behind this strategy is to leave out the enzymatic buildup of the linear peptide scaffolds and replace it by solid-phase peptide synthesis (SPPS) (Fig. 5). Once the desired peptide is produced, its C terminus needs to be synthetically activated (usually as a thioester) before the substrate can be subjected to enzymatic cyclization using a TE domain. The advantages of solid-phase synthesis (SPPS) are obvious. Virtually any oligo-peptide can be made in a short time and in large quantities. Even though this is true for most oligo-peptides, some amino acid sequences seem very difficult to synthesize, and the popular Fmoc protective group strategy always imposes the risk of racemization. Many different building blocks (already modified with protective groups necessary for SPPS) can be purchased, and by automated parallel peptide synthesis whole libraries can be produced very quickly. The reason why one can use such peptides as substrates lies in the relaxed substrate specificity of many TE domains. The TE domain of the Tyrocidine synthetase, which carries out a head-to-tail cyclization of the decapeptide DPhe-Pro-Phe-DPhe-Asn-Gln-Tyr-Val-Orn-Leu, for instance, only recognizes the two N-terminal and C-terminal amino acids of the natural substrate. The side chains of the amino acids at other positions are not recognized by the enzyme, and experiments with substrates that carry substitutions to alarinate these positions still lead to analogous cyclodecamers (31). The major advantage of using TE domains for cyclization reactions is their regiospecificity, stereospecificity, and chemo-selectivity. Thus, no protective groups are needed during these enzymatic cyclization reactions, and undesired side product formation is minimized. Additionally, these reactions are carried out under mild aqueous conditions, usually pH 7-8.

To follow this chemoenzymatic approach, the synthetic substrates must be transferred onto the catalytically active serine residue of the TE domain. This transfer can either be done directly or with the help of a PCP domain. In the natural system, translocation is realized by the interaction between the PCP and the TE domain. The substrate, which is bound to the 4Ppan cofactor of the PCP domain as a thioester, acylates the hydroxy group of the serine. Chemically speaking, the acylation of the TE is a result of a trans-esterification. When using TE domains, the substrate provided in trans must also have an appropriate acylation potential. Several key techniques have been developed to covalently attach synthetic substrates to PCP and TE domains. In the first method, the relaxed substrate specificity of the 4Ppan transferase Sfp is used to load aryl moieties...
onto PCP domains enzymatically. Just like in the natural priming reaction in which the 4Ppn part of CoA is transferred onto the conserved serine residue of the apo-PCP domain, Sfp does analogously attach S-acylated 4Ppns, which originates from 5-acylated CoA substrates (32). These CoA substrates can readily be obtained by a coupling reaction directly after solid phase synthesis. With this technique, virtually any substrate can be brought to a desired position in recombinant NRPS enzymes containing an apo-PCP domain. This result is of great value when elucidating the catalytic properties and substrate specificities of other domains. When investigating TE domains, for example, the corresponding apo-PCP-TE would be the starting point for screening the cyclization abilities of the TE domain with a synthetic substrate library. However, the major disadvantage of this method becomes evident when looking closely at the enzyme after the release of the product. The PCP domain is now in its holo-state, and therefore, subsequent enzymatic loading of substrates with Sfp is impossible. Because product formation is limited to a single turnover, other methods have been developed to allow for multiple turnover.

For multiple turnover reactions, the TE domains must be supplied with substrates in trans; yet the acylation potential must be sufficient and the compound must be recognized by the enzyme. The first approach made was inspired by the natural system where the substrate is activated as a thioester bound to the Ppan cofactor. The idea was to minimize the 4Ppn moiety by replacing it with N-acetyl-cysteamine (SNAc) (33). This method works fairly well; however, it seems in later studies that the acylation potential is of greater importance than the similarity to the natural situation and a variety of peptidyl-thioesters with better leaving groups than SNAc was tested. The fastest turnover rates were found when thiophenol-esters were used (30, 34). Thiophenol has several advantages: it does not have any functional groups other than the thiol group; it is inexpensive, and it can easily be separated from the product. This method has been successfully used to shed light on the promiscuity of the TE domain of the daptomycin NRPS (35).

Even though these techniques allow for the production of new potentially bioactive compounds, they are usually closely related to known substances, which basically implies an analogous mode of action, and the deviations normally alter quantitative...
parameters such as solubility and affinity. Nevertheless, hundreds of nonribosomal systems still need to be explored, and the discoveries of new ones are frequently reported. The basic understanding of the catalytic functions of nonribosomal domains and modules that we have today is a good starting point for additional exploration and use of systems that are not yet understood fully.

References


Further Reading


See Also

Antibacterial Drugs, Design of

Antibiotics, From Microorganisms

Antibiotics, Synthesis of

Antibiotics, Mechanism of Action

Bacterial Resistance to Antibiotics

Coupling Methods: Peptide Synthesis

Natural and Unnatural Amino Acids, Synthesis of

Natural Products as Anticancer Agents

Natural Products in Microbes, Chemical Diversity of Peptides, Chemistry of

Peptide Combinatorial Libraries

Pharmaceuticals, Natural Products and Natural Product Models of Polyketides as Drugs

Polyketide Biosynthesis
Chemical Properties of Amino Acids

Donard S. Dwyer, Louisiana State University Health Sciences Center, Shreveport

doi: 10.1002/9780470048672.wecb007

A complete understanding of protein folding, protein recognition, enzyme catalysis, and allosteric regulation will require intimate knowledge about the fundamental physico-chemical properties of amino acids. This article will summarize the chemistry of the 20 naturally occurring amino acids with special emphasis on the relationship to protein folding and biologic activity. The major features of the amino acids are classified into three categories, which are as follows: 1) physico-chemical properties, 2) biologic properties, and 3) electronic properties. Cross-correlations among more than 34 different parameters revealed that these measures segregate into four main properties that are largely independent, as follows: 1) steric effects (polarizability), 2) hydrophilicity (Hp index), 3) inductive effects, and 4) field effects. Advances in quantum chemistry, nuclear magnetic resonance (NMR) analysis, and theoretic physics encourage efforts to derive all-electronic expressions for the fundamental properties of amino acids that will provide mechanistic insights into protein structure and function.

Proteins are complex polymers composed mainly of the 20 naturally occurring amino acids arranged in a series of peptide linkages. The precise sequence of amino acids determines the folding of a polypeptide chain and its ultimate 3-dimensional (3-D) structure. In addition, the amino acid sequence and the geometry of amino acid side chains specify protein binding sites and functional activities. Based on a comparison of heavy atoms, the 20 amino acids are on average greater than 50% identical in terms of their composition. Yet, it is the differences (sometimes very subtle) among individual amino acids that cause the variety of proteins in nature: from heat-stable DNA polymerases of thermal vent bacteria to antifreeze proteins of arctic fish, and from highly conserved structural proteins of microtubules to highly diverse signaling molecules of the G-protein-coupled receptor family. These differences are the focus of this review. The article is divided into three main sections that highlight the overall biologic relevance, the specific chemical properties of amino acid side chains, and methods for additional characterization of these properties.

Biologic Background

What is the biologic significance of diversity in amino acid side chains? The amino acid sequence is the blueprint for protein structure. Consequently, the complexity of protein structures is a function of the variety and the length of the sequences of polypeptide chains. In fact, multiple amino acid sequence alignments have improved the accuracy of secondary structure prediction and homology modeling greatly. Nevertheless, it is not known exactly how the properties of a single amino acid or a short stretch of amino acids determine the probability of that residue or sequence assuming a particular secondary structure. This issue is complicated additionally by the fact that identical sequences of five or more amino acids assume different secondary structures in proteins depending on the context (1, 2). A deeper appreciation of the chemical properties of amino acid side chains may improve modeling efforts and the prediction of secondary structure.

Whereas the 3-D structure of a protein is specified by its amino acid sequence, the active sites of receptors and enzymes are determined largely by the topological arrangement of non-contiguous amino acid side chains. Typically, ligand binding and enzyme catalysis require a precise geometry of functional groups of side chains to achieve specificity and catalytic activity. Moreover, amino acid side chains with distinct chemical properties are well suited for specialized tasks. For example, the presence of a cysteine residue in the active site defines a family of proteases (the caspases) that cleave at aspartic acid motifs in substrate proteins involved in apoptosis.
Chemical Properties of Amino Acids

Chemistry

Before discussing the chemistry of amino acid side chains, it is worthwhile to consider briefly the unique structure of amino acids and the functional implications. During evolution, amino acids likely were among the first chemical compounds to emerge on primitive earth (6, 7). Early experiments on the origins of life sought to recreate primordial atmospheric conditions with hydrogen, methane, ammonia, and water, and then to introduce a source of energy (e.g., electric discharges to mimic lightning strikes) or ultraviolet light to catalyze chemical reactions. Over time, these reactions yielded amino acids and other simple organic molecules that served as building blocks for the eventual synthesis of proteins, polynucleotides, and complex carbohydrates (6). Amino acids seem to have been formed from an initial reaction between aldehyde and ammonia and from additional convention in the Strecker synthesis (7). From a chemical standpoint, it is interesting that the first two building blocks of amino acids have opposite properties with respect to electron affinity. The amide group of ammonia releases electrons, whereas the aldehyde group tends to withdraw electrons when acting as a substituent group. This point is important because the bipolar construction of amino acids confers two of their most significant features. First, amino acids in aqueous solutions at physiologic pH are zwitterions (i.e., they carry a positive charge at the amino group and a negative charge at the carboxyl group). A single C₂ carbon separates these two oppositely charged species. Second, amino acids combine readily in condensation reactions to form polymers. This feature enabled the modular assembly of proteins from a diverse collection of interchangeable units that differ only in the side chain attached to the C₆ atom.

Electron delocalization in peptides

The bipolar nature of amino acids has additional ramifications. Amino acids are about 1000 times stronger than comparable aliphatic carboxylic acids because of electron withdrawal by the charged amino group. Thus, significant electronic effects (inductive and field effects) exist among the main chain atoms of an amino acid. Several findings support the existence of electron delocalization along the main chain of proteins. The peptide bond in proteins is planar with the partial double-bond character that reflects short-range electron delocalization clearly. Measurement of the pKₐ of dipeptides, tripeptides, and tetrapeptides (8) and nuclear magnetic resonance (NMR) studies of inductive effects (9) demonstrate that the delocalization is not restricted to the peptide bond, but it extends over a span of 3–4 residues. In addition, electron tunneling in proteins enables charge migration over very long distances and proceeds more efficiently through bonds than through space (10). Charges can migrate across the peptide bond (11), and, in fact, proteins behave as semiconductors under appropriate conditions (12). Finally, the bipolar nature of amino acids generates a distinct dipole in α-helical segments of proteins that is powerful enough to stabilize the binding of cofactors and ligands of opposite charge (13).

The C₆ atom is located in a unique position along the main chain because of electron delocalization between the amino and carboxyl groups. As discussed elsewhere, amino acid side chains can be considered as substituents along the peptide backbone that affect resonance and electron density at main chain atoms (14, 15). In turn, this reaction will affect bond lengths and rotational flexibility—the ultimate determinants of secondary structure. The chemical features of the side chains modulate the properties of localized segments of a protein in the same way that different substituent groups affect the reactivity and orientation of reactions that involve substituted molecules in classic chemistry. The idea that amino acid side chains affect the electron density along a polymer composed of repetitive units is consistent with observations of the effects of side chain composition on the conduction of semiconducting materials (16). We will return to this important notion of side chains as substituent groups along the peptide backbone during discussion of the electronic properties of amino acids. The next three sections will summarize the physico-chemical, biologic, and electronic properties of amino acid side chains.

Physico-chemical properties

For the purpose of this article, the various properties of amino acid side chains have been classified into three separate categories. The physico-chemical properties are represented by...
values that can be measured directly for each amino acid or that can be calculated directly from the behavior of component atoms or chemical groups. The biologic properties reflect indirect measures or context-dependent behavior of the amino acids (e.g., their preference for coil or helical conformations in proteins, and their partitioning into different solvents (hydrophobicity scales)). Finally, the electronic properties refer to a mixture of measured and calculated parameters that attempt to describe fundamental electronic effects of amino acid side chains. The electronic properties of an amino acid determine ultimately its physico-chemical properties; however, these two categories are discussed separately here to highlight the need for better characterization of these electronic effects. Summaries of these various properties are presented in Tables 1–3.

The physico-chemical properties of amino acids are summarized in Table 1. This includes a wide array of measures, from refractivity and melting point to the pKa at the amino group. Some of the parameters span a narrow range of values (e.g., the molecular weights and melting points). Other parameters differ by a factor of 100-fold or more, which include the pKas at the amino group and solubility. The AAindex database compiled by Kawashima et al. (21) is an excellent source of additional information that concerns the physico-chemical properties of amino acids.

Table 1: Physico-chemical properties of amino acids

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>M.W.</th>
<th>Refractivity*</th>
<th>t_m†</th>
<th>Heat of Formation‡</th>
<th>Solubility§</th>
<th>ΔG of Solution¶</th>
<th>pK_a (NH)∥</th>
<th>pK_a (s.c.)‡‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>89.09</td>
<td>4.34</td>
<td>297</td>
<td>-604.0</td>
<td>165.0</td>
<td>-350</td>
<td>9.69</td>
<td>—</td>
</tr>
<tr>
<td>Arg</td>
<td>174.20</td>
<td>26.6</td>
<td>244</td>
<td>-623.5</td>
<td>182.6</td>
<td>9.04</td>
<td>12.10</td>
<td>—</td>
</tr>
<tr>
<td>Asn</td>
<td>132.12</td>
<td>13.28</td>
<td>235</td>
<td>-789.4</td>
<td>25.1</td>
<td>950</td>
<td>8.80</td>
<td>—</td>
</tr>
<tr>
<td>Asp</td>
<td>133.10</td>
<td>12.00</td>
<td>270</td>
<td>-973.3</td>
<td>4.9</td>
<td>2550</td>
<td>9.60</td>
<td>3.71</td>
</tr>
<tr>
<td>Cys</td>
<td>121.16</td>
<td>35.77</td>
<td>240</td>
<td>-534.1</td>
<td>v.s.</td>
<td>10.23</td>
<td>8.14</td>
<td>—</td>
</tr>
<tr>
<td>Glu</td>
<td>147.13</td>
<td>17.26</td>
<td>160</td>
<td>-1009.7</td>
<td>8.6</td>
<td>2300</td>
<td>9.67</td>
<td>4.15</td>
</tr>
<tr>
<td>Gln</td>
<td>146.15</td>
<td>17.56</td>
<td>185</td>
<td>-826.4</td>
<td>42.0</td>
<td>9.13</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Gly</td>
<td>75.07</td>
<td>0</td>
<td>290</td>
<td>-528.5</td>
<td>259.9</td>
<td>-650</td>
<td>9.60</td>
<td>—</td>
</tr>
<tr>
<td>His</td>
<td>155.16</td>
<td>21.81</td>
<td>287</td>
<td>-466.7</td>
<td>43.5</td>
<td>9.17</td>
<td>6.04</td>
<td>—</td>
</tr>
<tr>
<td>Ile</td>
<td>131.17</td>
<td>19.06</td>
<td>284</td>
<td>-637.8</td>
<td>34.2</td>
<td>700</td>
<td>9.68</td>
<td>—</td>
</tr>
<tr>
<td>Leu</td>
<td>121.17</td>
<td>18.78</td>
<td>293</td>
<td>-637.4</td>
<td>22.0</td>
<td>950</td>
<td>9.60</td>
<td>—</td>
</tr>
<tr>
<td>Lys</td>
<td>146.19</td>
<td>21.29</td>
<td>224</td>
<td>-678.7</td>
<td>5.8</td>
<td>8.95</td>
<td>10.67</td>
<td>—</td>
</tr>
<tr>
<td>Met</td>
<td>149.21</td>
<td>21.64</td>
<td>281</td>
<td>-577.5</td>
<td>56.0</td>
<td>850*</td>
<td>9.21</td>
<td>—</td>
</tr>
<tr>
<td>Phe</td>
<td>165.19</td>
<td>29.40</td>
<td>283</td>
<td>-466.9</td>
<td>27.9</td>
<td>1000</td>
<td>9.13</td>
<td>—</td>
</tr>
<tr>
<td>Pro</td>
<td>115.13</td>
<td>10.93</td>
<td>221</td>
<td>-512.5</td>
<td>282.0</td>
<td>1050</td>
<td>9.60</td>
<td>—</td>
</tr>
<tr>
<td>Ser</td>
<td>105.09</td>
<td>6.35</td>
<td>228</td>
<td>-732.7</td>
<td>50.2</td>
<td>450*</td>
<td>9.15</td>
<td>—</td>
</tr>
<tr>
<td>Thr</td>
<td>119.12</td>
<td>11.01</td>
<td>256</td>
<td>-807.2</td>
<td>98.1</td>
<td>9.10</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Trp</td>
<td>204.23</td>
<td>42.53</td>
<td>289</td>
<td>-415.3</td>
<td>13.2</td>
<td>1750</td>
<td>9.39</td>
<td>—</td>
</tr>
<tr>
<td>Tyr</td>
<td>181.19</td>
<td>31.53</td>
<td>343</td>
<td>-685.1</td>
<td>0.5</td>
<td>3550</td>
<td>9.11</td>
<td>10.10</td>
</tr>
<tr>
<td>Val</td>
<td>117.15</td>
<td>13.92</td>
<td>315</td>
<td>-617.9</td>
<td>88.5</td>
<td>200</td>
<td>9.62</td>
<td>—</td>
</tr>
</tbody>
</table>

*The refractivity data are from Jones (37).
†The melting point (t_m), heat of formation, solubility, and pKa of the side chain (s.c.) data were obtained from the CRC Handbook of Chemistry and Physics (18).
‡Free energy of solution (ΔG) values were obtained from Greenstein and Winitz (19).
¶The pKa at the amino group were published by Edsall (20).
### Table 2: Biologic properties of amino acids

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>P&lt;sub&gt;α&lt;/sub&gt;</th>
<th>P&lt;sub&gt;β&lt;/sub&gt;</th>
<th>P&lt;sub&gt;γ&lt;/sub&gt;</th>
<th>C-F</th>
<th>C-F</th>
<th>C-F</th>
<th>Bulk&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Gyration&lt;sup&gt;1&lt;/sup&gt;</th>
<th>Polarizability&lt;sup&gt;§&lt;/sup&gt;</th>
<th>K-D&lt;sup&gt;¶&lt;/sup&gt;</th>
<th>RF&lt;sup&gt;**&lt;/sup&gt;</th>
<th>H&lt;sub&gt;p&lt;/sub&gt;&lt;sup&gt;††&lt;/sup&gt;</th>
<th>M-P&lt;sup&gt;P-P&lt;/sup&gt;</th>
<th>Hydrophobicity&lt;sup&gt;††&lt;/sup&gt;</th>
<th>IIMED</th>
<th>IILONG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala A</td>
<td>1.44</td>
<td>0.76</td>
<td>0.87</td>
<td>1.45</td>
<td>0.97</td>
<td>0.66</td>
<td>-1.67</td>
<td>0.77</td>
<td>1.1</td>
<td>1.0</td>
<td>-3.42</td>
<td>12.97</td>
<td>0.05</td>
<td>2.11</td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td>Arg R</td>
<td>1.25</td>
<td>0.92</td>
<td>1.01</td>
<td>0.79</td>
<td>0.90</td>
<td>1.20</td>
<td>1.27</td>
<td>3.28</td>
<td>8.5</td>
<td>6.46</td>
<td>0.14</td>
<td>3.72</td>
<td>-10.5</td>
<td>1.94</td>
<td>3.78</td>
<td></td>
</tr>
<tr>
<td>Asn N</td>
<td>1.15</td>
<td>0.75</td>
<td>1.77</td>
<td>0.63</td>
<td>0.33</td>
<td>0.51</td>
<td>-0.07</td>
<td>1.45</td>
<td>3.7</td>
<td>-3.34</td>
<td>0.15</td>
<td>11.42</td>
<td>-0.74</td>
<td>1.84</td>
<td>3.64</td>
<td></td>
</tr>
<tr>
<td>Asp D</td>
<td>1.20</td>
<td>0.66</td>
<td>1.87</td>
<td>0.80</td>
<td>0.69</td>
<td>0.12</td>
<td>-0.89</td>
<td>0.67</td>
<td>3.4</td>
<td>-0.42</td>
<td>-0.42</td>
<td>10.85</td>
<td>-1.04</td>
<td>1.80</td>
<td>2.85</td>
<td></td>
</tr>
<tr>
<td>Cys C</td>
<td>0.79</td>
<td>0.77</td>
<td>1.00</td>
<td>1.07</td>
<td>1.07</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>2.7</td>
<td>2.5</td>
<td>-1.34</td>
<td>14.63</td>
<td>0.06</td>
<td>1.88</td>
<td>5.55</td>
<td></td>
</tr>
<tr>
<td>Glu E</td>
<td>1.46</td>
<td>0.61</td>
<td>1.06</td>
<td>0.87</td>
<td>0.19</td>
<td>0.77</td>
<td>0.41</td>
<td>0.69</td>
<td>-0.35</td>
<td>0.41</td>
<td>-2.65</td>
<td>11.89</td>
<td>-0.99</td>
<td>2.09</td>
<td>2.72</td>
<td></td>
</tr>
<tr>
<td>Gln Q</td>
<td>0.85</td>
<td>0.88</td>
<td>0.88</td>
<td>1.13</td>
<td>0.24</td>
<td>0.97</td>
<td>0.48</td>
<td>0.63</td>
<td>-0.35</td>
<td>0.97</td>
<td>11.76</td>
<td>-0.73</td>
<td>2.03</td>
<td>3.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly G</td>
<td>0.54</td>
<td>0.66</td>
<td>1.02</td>
<td>0.53</td>
<td>0.81</td>
<td>1.42</td>
<td>-0.16</td>
<td>0.00</td>
<td>0.03</td>
<td>0.46</td>
<td>-1.02</td>
<td>12.43</td>
<td>-0.26</td>
<td>1.53</td>
<td>4.31</td>
<td></td>
</tr>
<tr>
<td>His H</td>
<td>0.83</td>
<td>0.67</td>
<td>1.08</td>
<td>1.24</td>
<td>0.92</td>
<td>0.52</td>
<td>-0.16</td>
<td>2.88</td>
<td>3.6</td>
<td>-3.82</td>
<td>0.84</td>
<td>12.16</td>
<td>-0.21</td>
<td>1.98</td>
<td>3.77</td>
<td></td>
</tr>
<tr>
<td>Ile I</td>
<td>1.30</td>
<td>0.85</td>
<td>0.70</td>
<td>1.00</td>
<td>0.60</td>
<td>0.16</td>
<td>-0.16</td>
<td>1.56</td>
<td>4.3</td>
<td>4.5</td>
<td>0.91</td>
<td>10.33</td>
<td>1.83</td>
<td>1.77</td>
<td>5.58</td>
<td></td>
</tr>
<tr>
<td>Leu L</td>
<td>1.38</td>
<td>1.04</td>
<td>0.55</td>
<td>1.34</td>
<td>1.22</td>
<td>0.66</td>
<td>1.09</td>
<td>1.54</td>
<td>4.2</td>
<td>3.8</td>
<td>17.6</td>
<td>14.90</td>
<td>0.99</td>
<td>2.19</td>
<td>4.59</td>
<td></td>
</tr>
<tr>
<td>Lys K</td>
<td>0.66</td>
<td>0.81</td>
<td>0.05</td>
<td>1.07</td>
<td>0.74</td>
<td>0.02</td>
<td>2.08</td>
<td>5.04</td>
<td>5.2</td>
<td>-3.35</td>
<td>4.70</td>
<td>11.36</td>
<td>-1.15</td>
<td>1.96</td>
<td>2.79</td>
<td></td>
</tr>
<tr>
<td>Met M</td>
<td>1.13</td>
<td>1.18</td>
<td>0.40</td>
<td>1.20</td>
<td>1.67</td>
<td>0.61</td>
<td>0.18</td>
<td>1.80</td>
<td>5.1</td>
<td>1.9</td>
<td>14.9</td>
<td>14.39</td>
<td>0.70</td>
<td>2.27</td>
<td>14.14</td>
<td></td>
</tr>
<tr>
<td>Phe F</td>
<td>0.92</td>
<td>1.54</td>
<td>0.76</td>
<td>1.12</td>
<td>1.28</td>
<td>0.81</td>
<td>0.28</td>
<td>1.90</td>
<td>8.0</td>
<td>2.8</td>
<td>18.8</td>
<td>-7.60</td>
<td>14.00</td>
<td>1.19</td>
<td>4.53</td>
<td></td>
</tr>
<tr>
<td>Pro P</td>
<td>0.45</td>
<td>0.65</td>
<td>1.28</td>
<td>0.59</td>
<td>0.62</td>
<td>1.45</td>
<td>-0.33</td>
<td>1.29</td>
<td>4.3</td>
<td>-1.6</td>
<td>14.8</td>
<td>-6.22</td>
<td>11.37</td>
<td>0.17</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>Ser S</td>
<td>1.24</td>
<td>0.74</td>
<td>1.21</td>
<td>0.79</td>
<td>0.72</td>
<td>1.27</td>
<td>1.08</td>
<td>1.08</td>
<td>1.6</td>
<td>-0.08</td>
<td>0.32</td>
<td>11.23</td>
<td>-0.43</td>
<td>1.57</td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td>Thr T</td>
<td>0.76</td>
<td>1.29</td>
<td>0.88</td>
<td>0.82</td>
<td>1.20</td>
<td>1.05</td>
<td>0.70</td>
<td>1.24</td>
<td>2.7</td>
<td>-0.76</td>
<td>-2.63</td>
<td>11.69</td>
<td>-0.30</td>
<td>1.57</td>
<td>4.09</td>
<td></td>
</tr>
<tr>
<td>Trp W</td>
<td>1.05</td>
<td>1.25</td>
<td>0.82</td>
<td>1.14</td>
<td>1.19</td>
<td>0.82</td>
<td>1.23</td>
<td>2.21</td>
<td>2.12</td>
<td>0.09</td>
<td>17.1</td>
<td>-5.95</td>
<td>13.93</td>
<td>1.13</td>
<td>4.89</td>
<td></td>
</tr>
<tr>
<td>Tyr Y</td>
<td>1.09</td>
<td>1.32</td>
<td>0.82</td>
<td>0.61</td>
<td>1.29</td>
<td>1.19</td>
<td>1.48</td>
<td>2.13</td>
<td>8.8</td>
<td>-1.3</td>
<td>4.55</td>
<td>13.42</td>
<td>0.44</td>
<td>1.67</td>
<td>4.93</td>
<td></td>
</tr>
<tr>
<td>Val V</td>
<td>0.48</td>
<td>1.04</td>
<td>0.66</td>
<td>1.04</td>
<td>1.65</td>
<td>0.66</td>
<td>-0.71</td>
<td>0.79</td>
<td>3.2</td>
<td>-4.2</td>
<td>14.3</td>
<td>-8.02</td>
<td>15.71</td>
<td>0.78</td>
<td>1.63</td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>The preferences for α-helix, β-strands, and coil structures, P<sub>α</sub>, P<sub>β</sub>, and P<sub>coil</sub>, respectively, were taken from Dwyer (14).

<sup>2</sup>Secondary structural preference data were obtained from Chou and Fasman (C-F) (26).

<sup>3</sup>The bulk measure was derived by Kidera et al. (22).

<sup>4</sup>Levitt (23) calculated the gyration index.

<sup>5</sup>Polarizability and hydrophilicity (Hp) measures were derived from QM calculations by Dwyer (15, 32). Polarizability refers to the α-component calculated with the PM3 QM method.

<sup>6</sup>The hydropathy scores from Kyte and Doolittle (K-D index) (30) are presented.

<sup>7</sup>The RF data were obtained from Zimmerman et al. (31). These values were based on the average mobility of the amino acids in a series of solvents determined by paper chromatography.

<sup>8</sup>The hydrophobicity index of Manavalan and Ponnuswamy (M-P) (33) was calculated from the average surrounding hydrophobicity based on 3-D structures of proteins.

<sup>9</sup>The hydrophobicity index of Palliari and Parry (P-P) calculated a hydrophobicity index based on the average normalized values of 127 scales (35).

<sup>10</sup>Gromiha and Selveraj (34) determined the average number of medium and long range inter-residue interactions (IIMED and IILONG, respectively) for each of the 20 amino acids based on crystallographic data.
Other significant correlations were observed among the following: 1) melting point and dipole moment, average hydrophobicity (P-P), and long-range inter-residue interactions (IILONG), 2) solubility and pKa, and 3) pKa at the amino group and the z2 score of Hellberg et al. (25). The last relationship likely reflects the fact that pKa was one of the multiple components used to derive the z2-score, which is a composite of different electronic effects. An inverse relationship exists between free energy of solution and field effects (Table 4). This relationship may result from the polarity of the amino acid side chains.

**Biologic properties**

The three major groupings in the biologic properties correspond to preference for secondary structure (P-P, P-H, and P_a), steric or bulk effects (bulk, gyration, and polarizability), and hydrophobicity (K-D, R_H, H_p, M-P, P-P, and IILONG) (see Table 3).

Preferences for secondary structure (P-P) were derived by Dwyer (14) and Chou and Fasman (26) from statistical analysis of large databases of nonredundant protein structures. These data are in close agreement with similar statistical analyses of structural propensities of amino acids, for example that of Williams et al. (27). A second method to evaluate the preference of amino acids for secondary structure is host-guest analysis of short synthetic peptides (13). In these studies, amino acids are substituted into peptides that assume α-helical or β-strand structures and the effects of the substitution on structural stability are assessed. Several limitations to this approach exist; for instance, an overemphasis on preference at central positions within a segment leads to underestimates of the structural propensity of amino acids that are found commonly at the ends of secondary structures (e.g., asparagine and aspartic acid), which are excellent N-cap residues in α-helices (28). Nevertheless, the α-helix preferences summarized in Table 3 show significant correlation (r = 0.7) with indices derived from host-guest analysis, such as that of O’Neil and DeGrado (29). Finally, secondary structural preferences show significant correlations with electronic properties, including dipole, 〈focal charge〉, and NMR shift (Table 4).

Some amino acids show a clear preference for a particular secondary structure. For example, glutamic acid and alanine show a very high propensity for α-helices and are found at much lower frequencies in other structures. Similarly, valine, phenylalanine, cysteine, and threonine mainly prefer β-strands, whereas glycine, proline, and serine favor coil or turn conformations strongly. Other amino acids such as arginine, glutamine, and lysine, do not show an overwhelming preference to which a residue is buried in the native protein (related to hydrophobicity) and the number of inter-residue interactions it forms (32). The fact that amino acids prefer certain secondary structures seems to be valid and reflect fundamental properties of the side chain.

The fact that amino acids prefer certain secondary structures does not address the question of why these structural preferences are observed. This issue will be discussed in greater detail in a later section. Here, the role of steric effects and hydrophobicity will be considered. The composite bulk scale of Kidera et al. (22) and the gyration scale of Levitt (23) represent faithfully the bulk or the steric effects of amino acids. Therefore, it was interesting to observe such a striking correlation between these scales and polarizability, which is based on the calculation of a single, defined electronic feature of a molecule. Data in Table 4 suggest that the scales for steric effects are reasonably pure, and show no correlation with secondary structure. That is not to say that steric effects are unimportant in protein folding. Rather, it seems safe to conclude that they are not a primary driving force for the formation of secondary structure.

Two of the hydrophilicity scales in Table 2 were derived from experimental measures of the behavior of amino acids in various solvents, namely partitioning coefficients (K-D index of Kyte and Doolittle (30) or mobility in paper chromatography (R_f index of Zimmerman et al. (31))). By contrast, the H_p index was obtained from quantum mechanics (QM) calculations of electron densities of side chain atoms in comparison with water (32). The H_p index is correlated highly with these two established hydrophobicity scales (Table 4). Therefore, like the polarizability index, it is possible to represent fundamental chemical properties of amino acids (hydrophobicity, H_p) with parameters derived from ab initio calculations of electronic properties. However, in contrast to polarizability (steric effects), hydrophilicity shows significant correlation with preference for secondary structure. Thus, hydrophobic amino acids prefer β-strands (and β-sheet conformations) and typically are buried in protein structures, whereas hydrophilic residues are found commonly in turns (coil structure) at the protein surface.

Several scales have been derived to quantify the degree to which a side chain is exposed to the solvent (34). The z3-score, which is a composite of different physical and chemical properties, including dipole, 〈focal charge〉, local and non-local, and NMR shift, shows significant correlation (r = 0.70–0.89) with hydrophilicity scales and with strand versus coil conformations (Table 4). NMR studies reveal that the amide proton is shielded to a greater extent in coil conformations as compared with extended (g) structures (37); increased electron density exists at this atom in the coil conformation. Taken together, the data suggest strong interactions between hydrophilicity and electronic parameters in folding and provide support for additional refinement of the H_p index.
shift data for the amide proton, the $\gamma$ electronic effects. An empirical data set (40), and they represent a composite of coefficients, and to some degree the dipole index are related to some degree the dipole index are related to some degree.

Additional analysis of these electronic parameters reveals that they fall into two major categories. The NMR chemical shift data for the amide proton, the $\delta_{\text{LOCAL}}$ and $\delta_{\text{NON-LOCAL}}$ coefficients, and to some degree the dipole index are related closely to the hydrophilicity of the amino acid side chains. These findings are shown in Table 4, and they include significant correlations with K-D, H-P, M-P, and P-P indices. Given the extent of cross-correlation, it is probably best to consider the first three electronic scales in particular as surrogate indicators of hydrophilicity. The parameters that remain in Table 4 bear some relationship to each other and are considered, for the purpose of this article, to represent conventional electronic effects of amino acid side chains. This includes $\delta_{\text{LOCAL}}$, inductive and field effects, $z_3$, and VHSE5 parameters. The correlation between VHSE5 and inductive field and resonance effects is remarkable. Whereas VHSE5 is a composite from principal components analysis, the inductive and resonance effects are derived directly from simple QM calculations (15). Thus, the cross-correlations between empirically derived electronic parameters (VHSE5) and QM-derived calculations suggest that the QM measures are valid expressions of the electronic properties of amino acids.

Closer examination of the electronic properties of individual amino acids reveals good agreement with expectations. Thus, arginine and lysine with positively charged ammonium groups are the strongest electron withdrawing side chains (inductive effects), whereas the side chains of aspartic acid and glutamic acid are the strongest electron donors. Conversely, field effects (with respect to the amide proton) are opposite in direction for the

---

**Table 3** Electronic properties of amino acids

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Dipole $^*$</th>
<th>$\delta_{\text{LOCAL}}$ $^*$</th>
<th>$\delta_{\text{NON-LOCAL}}$</th>
<th>NMR Shift $^\dagger$</th>
<th>VHSE5 $^\ddagger$</th>
<th>$z_3$ $^\S$</th>
<th>Inductive $^|$</th>
<th>Resonance $^\S$</th>
<th>Field $^\S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala A</td>
<td>0</td>
<td>0.163</td>
<td>0.236</td>
<td>8.12</td>
<td>0.02</td>
<td>0.09</td>
<td>4.5978</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Arg R</td>
<td>5.78</td>
<td>0.220</td>
<td>0.233</td>
<td>8.23</td>
<td>1.55</td>
<td>-3.44</td>
<td>4.5381</td>
<td>-0.26</td>
<td>-0.49</td>
</tr>
<tr>
<td>Asn N</td>
<td>4.06</td>
<td>0.324</td>
<td>0.389</td>
<td>8.33</td>
<td>-0.55</td>
<td>0.84</td>
<td>4.5431</td>
<td>-0.14</td>
<td>-0.06</td>
</tr>
<tr>
<td>Asp D</td>
<td>4.33</td>
<td>0.212</td>
<td>0.168</td>
<td>8.38</td>
<td>-2.68</td>
<td>2.36</td>
<td>4.3934</td>
<td>0.51</td>
<td>1.29</td>
</tr>
<tr>
<td>Cys C</td>
<td>1.78</td>
<td>0.316</td>
<td>0.259</td>
<td>8.18</td>
<td>0</td>
<td>4.13</td>
<td>4.6375</td>
<td>-0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Glu E</td>
<td>6.13</td>
<td>0.212</td>
<td>0.306</td>
<td>8.40</td>
<td>-2.16</td>
<td>-0.07</td>
<td>4.4447</td>
<td>0.68</td>
<td>0.57</td>
</tr>
<tr>
<td>Gln Q</td>
<td>3.89</td>
<td>0.274</td>
<td>0.314</td>
<td>8.19</td>
<td>0.09</td>
<td>-1.14</td>
<td>4.6050</td>
<td>-0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>Gly G</td>
<td>0.00</td>
<td>0.080</td>
<td>-0.170</td>
<td>8.36</td>
<td>-0.53</td>
<td>0.30</td>
<td>4.7053</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>His H</td>
<td>4.04</td>
<td>0.315</td>
<td>0.256</td>
<td>8.36</td>
<td>0.51</td>
<td>1.11</td>
<td>4.5323</td>
<td>-0.03</td>
<td>0.22</td>
</tr>
<tr>
<td>Ile I</td>
<td>0.07</td>
<td>0.474</td>
<td>0.391</td>
<td>7.99</td>
<td>0.30</td>
<td>-1.03</td>
<td>4.4995</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>Leu L</td>
<td>0.09</td>
<td>0.315</td>
<td>0.293</td>
<td>7.99</td>
<td>0.22</td>
<td>-0.98</td>
<td>4.5929</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>Lys K</td>
<td>9.07</td>
<td>0.255</td>
<td>0.233</td>
<td>8.29</td>
<td>1.64</td>
<td>-3.14</td>
<td>4.5119</td>
<td>-0.16</td>
<td>0.95</td>
</tr>
<tr>
<td>Met M</td>
<td>1.80</td>
<td>0.356</td>
<td>0.367</td>
<td>8.12</td>
<td>0.23</td>
<td>-0.41</td>
<td>4.6201</td>
<td>0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>Phe F</td>
<td>0.29</td>
<td>0.410</td>
<td>0.328</td>
<td>7.93</td>
<td>0.25</td>
<td>0.45</td>
<td>4.5783</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Thr T</td>
<td>1.79</td>
<td>0.325</td>
<td>0.397</td>
<td>8.03</td>
<td>0.75</td>
<td>0.85</td>
<td>4.5755</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>Tyr Y</td>
<td>1.79</td>
<td>0.325</td>
<td>0.397</td>
<td>8.03</td>
<td>0.75</td>
<td>0.85</td>
<td>4.5755</td>
<td>0.06</td>
<td>0.09</td>
</tr>
<tr>
<td>Val V</td>
<td>0.06</td>
<td>0.515</td>
<td>0.436</td>
<td>8.08</td>
<td>0.22</td>
<td>1.29</td>
<td>4.6039</td>
<td>0.01</td>
<td>0.08</td>
</tr>
</tbody>
</table>

$^*$The dipole data were derived by Chipot et al. (38) from QM calculations.
$^\dagger$The NMR data refer to the chemical shifts measured by NMR for the amide proton in the crystal conformation. (37)
$^\ddagger$The VHSE5 electronic effects were derived from principal components analysis of 50 different physico-chemical variables. (15)
$^\S$The z3 electronic index was taken from Hellberg et al. (25). This scale was derived from principal components analysis of 29 variables. The z3 index is completely independent from hydrophobic (z1 scale) and steric (z2 scale) effects.
$^\|$The dipole data were derived by Chipot et al. (38) from QM calculations.
$^\|$The NMR shift data refer to the chemical shifts measured by NMR for the amide proton in the crystal conformation. (37)
$^\|$The z3 electronic index was taken from Hellberg et al. (25). This scale was derived from principal components analysis of 29 variables. The z3 index is completely independent from hydrophobic (z1 scale) and steric (z2 scale) effects.

Electronic properties

The electronic properties of amino acid side chains are summarized in Table 3, and they represent a wide spectrum of measures. The NMR data are derived experimentally (37). The dipole (38), $\delta_{\text{LOCAL}}$, inductive, field, and resonance effects were derived from QM calculations (15). The VHSE5 (39) and z3 (25) scales were developed for use in quantitative structure-activity relationship analysis of the biologic activity of natural and synthetic peptides. Both were derived from principal components analysis of assorted physico-chemical properties, which included NMR chemical shift data, electron-ion interaction potentials, charges, and isoelectric points. Therefore, these scales are composites rather than primary measures of electronic effects. The validity of these measures is indicated by their lack of overlap with hydrophobicity and steric parameters and by their ability to predict biologic activity of synthetic peptide analogs (25, 39). Finally, coefficients of electrostatic screening by amino acid side chains ($\delta_{\text{LOCAL}}$ and $\delta_{\text{NON-LOCAL}}$) were derived from an empirical data set (40), and they represent a composite of electronic effects.

Additional analysis of these electronic parameters reveals that they fall into two major categories. The NMR chemical shift data for the amide proton, the $\delta_{\text{LOCAL}}$ and $\delta_{\text{NON-LOCAL}}$ coefficients, and to some degree the dipole index are related closely to the hydrophilicity of the amino acid side chains. These findings are shown in Table 4, and they include significant correlations with K-D, H-P, M-P, and P-P indices. Given the extent of cross-correlation, it is probably best to consider the first three electronic scales in particular as surrogate indicators of hydrophilicity. The parameters that remain in Table 4 bear some relationship to each other and are considered, for the purpose of this article, to represent conventional electronic effects of amino acid side chains. This includes $\delta_{\text{LOCAL}}$, inductive and field effects, $z_3$, and VHSE5 parameters. The correlation between VHSE5 and inductive field and resonance effects is remarkable. Whereas VHSE5 is a composite from principal components analysis, the inductive and resonance effects are derived directly from simple QM calculations (15). Thus, the cross-correlations between empirically derived electronic parameters (VHSE5) and QM-derived calculations suggest that the QM measures are valid expressions of the electronic properties of amino acids.

Closer examination of the electronic properties of individual amino acids reveals good agreement with expectations. Thus, arginine and lysine with positively charged ammonium groups are the strongest electron withdrawing side chains (inductive effects), whereas the side chains of aspartic acid and glutamic acid are the strongest electron donors. Conversely, field effects (with respect to the amide proton) are opposite in direction for the...
Table 4: Correlations among the various properties of amino acids

<table>
<thead>
<tr>
<th>Property</th>
<th>M.W.</th>
<th>Refractivity</th>
<th>∆G solution</th>
<th>bulk</th>
<th>gyration</th>
<th>polarizability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.85</td>
<td>0.71</td>
<td>0.98</td>
<td>0.93</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>Refractivity</td>
<td>0.59</td>
<td>0.80</td>
<td>0.75</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>∆G solution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat of formation</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solubility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat of formation</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solubility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat of formation</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solubility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heat of formation</td>
<td>0.65</td>
<td>0.61</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solubility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Biologic Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Pα</th>
<th>C-F Pα</th>
<th>Hmed</th>
<th>0.66</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pα</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-F Pα</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hmed</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polarizability</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-D</td>
<td>0.64</td>
<td>-0.73</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>H2</td>
<td>0.80</td>
<td>-0.83</td>
<td>0.58</td>
<td>-0.75</td>
</tr>
</tbody>
</table>

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
### Table 4 (Continued)

<table>
<thead>
<tr>
<th>Dipole</th>
<th>γ_{LOCAL}</th>
<th>NMR shift</th>
<th>γ_{LOCAL}</th>
<th>NMR shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_p</td>
<td>-0.83</td>
<td>0.65</td>
<td>0.83</td>
<td>-0.73</td>
</tr>
<tr>
<td>H_p</td>
<td>-0.75</td>
<td>-0.36</td>
<td>-0.69</td>
<td>-0.80</td>
</tr>
<tr>
<td>M-P</td>
<td>-0.89</td>
<td>0.96</td>
<td>-0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>P-P</td>
<td>0.85</td>
<td>-0.81</td>
<td>0.64</td>
<td>-0.86</td>
</tr>
<tr>
<td>I_{LONG}</td>
<td>dipole</td>
<td>γ_{LOCAL}</td>
<td>NMR shift</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.74</td>
<td>0.59</td>
<td>-0.70</td>
<td></td>
</tr>
</tbody>
</table>

### Electronic Properties

<table>
<thead>
<tr>
<th>Dipole</th>
<th>γ_{LOCAL}</th>
<th>NMR shift</th>
<th>γ_{LOCAL}</th>
<th>NMR shift</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.55</td>
<td>0.74</td>
<td>-0.76</td>
<td>-0.67</td>
</tr>
</tbody>
</table>

*Linear regression analysis was performed to compare the various scales listed in Tables 1–3. The r values shown here are only for the significant correlations where p < 0.01 or p < 0.05 as noted by an asterisk. The (–) signs indicate the slope of the regression line.
Substituent effects of amino acid side chains

The concept of amino acid side chains as substituent groups along the peptide backbone has been developed previously (14, 15, 32). A corollary to this idea, the amino acids side chains affect the electron density at main chain atoms as a function of their physico-chemical properties (e.g., the degree to which they donate or withdraw electrons from the peptide backbone). The side chains modulate local electron density, and thus the bond lengths and rotation along the main chain (15, 32). Summed together, the various electronic effects (e.g., inductive effects and \( H_{\beta} \)) determine the preference of a protein segment for a particular secondary structure. Of course, this preference is modified by solvent effects, electrostatic screening, and ultimately by interactions between residues that are only brought into contact through the folding process. These long-range interactions determine both the folding rate and the stability of the folded protein (33, 34, 41).

Theoretic and experimental studies of charge migration in proteins support the notion of gating effects of amino acid side chains (8–12). Thus, movement of charge between adjacent residues through the peptide bond depends on the molecular motion and orientation of the side chains (11). Moreover, ab initio analysis of the electronic features of amino acids reveals that electronic effects are conformation sensitive (42). Therefore, different side chain rotamers will produce distinct electronic effects at the main chain, although the rotational preference of a side chain is also a function of its fundamental physico-chemical (electronic) properties. Quantum effects in molecular electronic devices reveal that side chain groups affect electron density and current flow through main chain (nonpeptide) atoms such that current transmission is blocked at eigenvalues of the side chains (16). The effects of various side chains are additive in this system. Therefore, main chain structure and rotational flexibility (i.e., folding) is linked inextricably to the electronic properties of amino acid side chains and to the propagation of electronic effects along the peptide backbone.

Nearest-neighbor effects

Nearest-neighbor effects refer to the reciprocal influence of adjacent amino acids on protein folding. In some cases, this term also refers to amino acids that are close in the 3-D structure of the protein (\( \sim 10 \) Å away), but distant in the sequence. Early studies found a nonrandom assorting of amino acids in secondary structures by pair-wise analysis of protein sequences (43). More recently, it was reported that the preference of pairs of amino acids for secondary structure was determined, in part, by the electronic properties of the neighboring residues (14, 32). Thus, adjacent pairs of amino acids that act as strong electron donors preferred \( \alpha \)-helical conformations, whereas adjacent residues with ambivalent electron affinity preferred strongly coil conformations. The existence of nearest-neighbor electronic effects in proteins is confirmed by NMR studies (9, 37), \( \varepsilon\alpha \) measurements (8), and QM analysis of electron densities in dipeptides (32). Finally, nearest-neighbor interactions in the final folded state contribute to the stability of a protein (44).

The electronic properties of neighboring amino acids can affect protein folding in complex ways. Theoretically, the electronic effects of adjacent residues may be additive, opposing, or neutral. Some effects extend over 3-4 residues in a peptide or protein (8, 9), which corresponds roughly to a loop, short \( \beta \)-strand, or the first turn of an \( \alpha \)-helix. The summation of these various nearest-neighbor effects will then determine the electron density along the peptide backbone, bond lengths, and rotational flexibility. Consequently, segments of a protein where strong electronic effects are exerted on the main chain atoms will tend to form different secondary structures than segments where the electronic effects are weak.

Chemical Tools and Refinements

A compelling case can be made for replacing empirically-derived scales of amino acid properties with parameters either measured directly (e.g., chemical shift data or infrared spectra) or calculated from basic principles. The goal would be to develop all-electronic expressions for the physico-chemical properties of amino acids based on computational methods that include QM calculations. A start in this direction was provided by the successful description of steric effects in terms of polarizability and hydrophilicity as a function of electron density (32). Application of more sophisticated computational approaches will speed progress toward this objective.

As discussed here and elsewhere, NMR chemical shift data reveal details about the secondary structure environment of amino acids in proteins and thus are measures of protein folding. The pioneering work of Oldfield and colleagues (45) and other groups demonstrated that QM calculations on model peptides estimated NMR chemical shifts accurately. Furthermore, chemical shifts at the amide proton are excellent indicators of \( \beta \)-strand and coil conformation (Table 4). Therefore, the fact that QM calculations can be used to derive chemical shifts in peptides suggests that these calculations also provide insight into protein folding. For some of the QM analysis of amino acids, older semiempirical methods have been used (15). These methods are sufficiently accurate for the relative assessment of electronic properties (i.e., for comparisons between amino acids). However, higher-level approaches will be needed to obtain more precise quantitative values for various electronic parameters, for example, through the application of density functional theory with correction for electron correlation effects (46). Alternatively, perturbation methods such as Miller-Plesset have also proven useful to derive the electronic properties of amino acids (46).
Several scales presented in Table 3 show promise as measures of fundamental electronic properties of amino acids as does the $H_p$ index of hydrophilicity. Nevertheless, additional improvements are desirable. The polarizability index of electronic derivatives for the component atoms of amino acid side chains. A more integrative approach with higher-order theory is likely to refine this measurement additionally.

Of the other electronic parameters described here, inductive effects and field effects seem to represent distinct properties as calculated from QM analysis and equations of localized substituent effects. However, inductive and resonance effects need to be better isolated from each other. QM calculation of these effects in a series of different host molecules may improve discrimination between these two parameters. Finally, it will be important to characterize the electronic properties of amino acid side chains in short dipeptides and tripeptides, which will lead to a better understanding of nearest-neighbor and context effects in proteins.

Conclusions

A better understanding of the chemical biology of amino acids is key to clarifying issues related to protein structure and function. Interactions that involve amino acid side chains contribute to the rate of protein folding, the stability of the protein fold, molecular recognition (e.g., ligand binding), and catalysis. Participation of amino acids in these processes is a function of the properties of the side chain groups. For example, the hydrophilicity of a side chain determines largely whether it is found at the protein surface or buried in the interior. Similarly, the propensity of an amino acid for medium to long-range inter-residue interactions is related to preference for secondary structure. Electronic properties of side chain groups contribute to folding preferences and create electric fields involved in recognition and catalysis. Many empirical measures of the properties of amino acids are now available. An important goal for the future will be to replace these empirical measures with fundamental parameters derived from QM calculations. Finally, the conceptualization of amino acid side chains as substituent groups that affect electron density along the main chain through gating effects may provide insight into how the amino acid sequence specifies the 3-D structure of a protein.

References

28. Richard et al., Principles and patterns of protein conformation. In Prediction of Protein Structure and the Principles...
Chemical Properties of Amino Acids


32. Dwyer DS. Nearest-neighbor effects and structural preferences in dipeptides are a function of the electronic properties of amino acid side-chains. Proteins 2006;63:939–948.


Further Reading

See Also

Chemistry and Chemical Reactivity of Proteins, Matthew Francis
Energetics of Protein Folding, Robert Baldwin
NMR to Study Proteins, Angela Gronenborn
Physical Chemistry in Biology, Allan Cooper
Synthetic Peptides and Proteins to Elucidate Biological Functions, Roger Goody
Chemical Strategies for Enzyme Catalysis

Timothy D.H. Bugg, Department of Chemistry, University of Warwick, Coventry, United Kingdom
doi: 10.1002/9780470048672.wecb154

This article will describe the different chemical strategies used by enzymes to achieve rate acceleration in the reactions that they catalyze. The concept of transition state stabilization applies to all types of catalysts. Because enzyme-catalyzed reactions are contained within an active site of a protein, proximity effects caused by the high effective concentrations of reactive groups are important for enzyme-catalyzed reactions, and, depending on how solvent-exposed the active site is, substrate desolvation may be important also. Examples of acid–base catalysis and covalent (nucleophilic) catalysis will be illustrated as well as examples of “strain” or substrate destabilization, which is a type of catalysis observed rarely in chemical catalysis. Some more advanced topics then will be mentioned briefly: the stabilization of reactive intermediates in enzyme active sites and the possible involvement of protein dynamics and hydrogen tunneling in enzyme catalysis.

Introduction

Enzymes are biologic catalysts that speed up chemical and biochemical reactions. Three particular characteristics of enzyme catalysis that distinguish enzymes from most man-made catalysts are speed, selectivity, and specificity. Enzymes are capable of rate accelerations of $10^7$–$10^{17}$ compared with the uncatalyzed reaction. Enzymes are highly selective catalysts; they can recognize and bind a single enantiomer of racemic mixture because the active site of the enzyme is chiral. The specificity of enzyme-catalyzed reactions is very high also; in reactions that generate a new chiral center (e.g., reduction of a ketone to an alcohol), only one enantiomer of the product is obtained. Again, this result is because of the chirality of the enzyme active site and the selective binding of the substrate by the enzyme.

This article will describe the different chemical strategies used by enzymes to achieve rate acceleration in the reactions that they catalyze. The concept of transition state stabilization applies to all types of catalysts. Because enzyme-catalyzed reactions are contained within an active site of a protein, proximity effects caused by the high effective concentrations of reactive groups are important for enzyme-catalyzed reactions, and, depending on how solvent-exposed the active site is, substrate desolvation may be important also. Examples of acid–base catalysis and covalent (nucleophilic) catalysis will be illustrated as well as examples of “strain” or substrate destabilization, which is a type of catalysis observed rarely in chemical catalysis. Some more advanced topics then will be mentioned briefly: the stabilization of reactive intermediates in enzyme active sites and the possible involvement of protein dynamics and hydrogen tunneling in enzyme catalysis.

Transition State Theory for Enzyme Catalysis

The rate of a chemical reaction of substrate S to product P is governed by the activation energy ($E_{act}$), which is the difference in free energy between the reactant(s) and the transition state for the reaction. The relationship between rate and $E_{act}$ is described by the Arrhenius equation:

$$k = A \cdot e^{-E_{act}/RT}$$

The transition state theory for enzyme catalysis predicts that catalysis is achieved by reducing the activation energy for the catalyzed reaction. This reduction in activation energy can be achieved either by stabilization (and hence reduction in free energy) of the transition state by the catalyst or by the catalyst finding some other lower energy pathway for the reaction. To illustrate the above equation, if a catalyst can provide 10kJ mol$^{-1}$ of transition stabilization energy for a reaction at 25°C, then a 55-fold rate acceleration will result, whereas a 20kJ mol$^{-1}$ stabilization will give a 3000-fold acceleration and a 40kJ mol$^{-1}$ stabilization a 10$^7$-fold acceleration.
Chemical Strategies for Enzyme Catalysis

Figure 1 illustrates the free energy profile of a typical acid-catalyzed chemical reaction that converts a substrate $S$ to a product $P$. In this case, an intermediate chemical species $SH^+$ is formed upon protonation of $S$. If the activation energy for conversion of $SH^+$ to $PH^+$ is lower than for the conversion of $S$ to $P$, then the reaction will go faster. It is important at this point to define the difference between an intermediate and a transition state: An intermediate is a stable (or semistable) chemical species formed during the reaction and therefore is a local energy minimum, whereas a transition state, by definition, is a local energy maximum.

An enzyme-catalyzed reaction can be analyzed thermodynamically in the same way as the acid-catalyzed example, except that enzyme-catalyzed reactions are multistep sequences that involve several intermediates. An enzyme-substrate intermediate $ES$ is formed during binding of the substrate, which then is converted to the enzyme-product complex $EP$ either directly or via one or more additional intermediates.

Under saturating substrate concentrations, the rate of the enzyme-catalyzed reaction will be governed by the activation energy for the conversion of the $ES$ complex to the $EP$ complex.

Exactly how do enzymes achieve transition state stabilization? I will illustrate three examples: the first two involve hydrogen-bonding interactions, and the third involves electrostatic interactions. The first example is the well-studied serine protease enzyme, $\alpha$-chymotrypsin, which cleaves polypeptide substrates via hydrolysis of the peptide bond adjacent to phenylalanine or tyrosine residues. The active site of $\alpha$-chymotrypsin contains a catalytic triad comprising Ser-195, His-57, and Asp-102. Active site base His-57 deprotonates the hydroxyl side chain of Ser-195, which attacks the amide carbonyl of the substrate, as shown in Fig. 2a [1]. The oxyanion intermediate that is generated is stabilized via hydrogen bonding to two backbone amide N–H groups, those of Ser-195 and Gly-193, in the "oxyanion hole," which is illustrated in Fig. 2b. These hydrogen bonds are formed only with the oxyanion, not with the bound substrate, and hence provide transition state stabilization.

The second example is a sialyltransferase enzyme from Campylobacter jejuni, which catalyzes the transfer of a 9-carbon sialic acid sugar from CMP–Neu5Ac to an acceptor substrate [2]. Analysis of the crystal structure of this enzyme revealed that the phosphate-leaving group of CMP interacted with two active site tyrosine residues, Tyr-156 and Tyr-162. Replacement of each tyrosine residue by phenylalanine gave 100-fold and 300-fold reduction in $k_{cat}$, respectively, which indicates that these residues are important in catalysis. As C–O cleavage takes place, the lengthening of the C–O bond and the additional negative charge on the phosphate group lead to much stronger hydrogen bonds with Tyr-156 and Tyr-162 in the transition state, as shown in Fig. 3 and hence transition state stabilization.

The third example is a phosphoryl transfer enzyme, alkaline phosphatase. The active site of alkaline phosphatase contains two Zn$^{2+}$ ions, with a separation of 3.9 Å. One zinc center is used to bind the phosphate monoester substrate, the other to activate Ser-102 for nucleophilic attack on the phosphate group of the substrate via an associative mechanism, as shown

---

Figure 1: Free energy profiles for (a) an acid-catalyzed reaction and (b) an enzyme-catalyzed reaction.

(a) Free energy profile for an acid-catalyzed reaction with transition states $E_{act(uncat)}$ and $E_{act(cat)}$.

(b) Free energy profile for an enzyme-catalyzed reaction with transition state $E_{act(uncat)}$ and $E_{act(cat)}$. 

---

...
Chemical Strategies for Enzyme Catalysis

Figure 2 (a) Mechanism of TS stabilization in α-chymotrypsin. (b) Structure of oxyanion hole in active site.

Figure 3 Sialyltransferase.

in Fig 4(3). Situated between the Zn$$^{2+}$$ ions is the guanidinium side chain of Arg-166, which provides electrostatic stabilization for the additional negative charge in the transition state of this phospho-transfer reaction.

Proximity Effects

Intramolecular reactions in organic chemistry generally proceed much more rapidly and under much milder reaction conditions than intermolecular reactions because the two reactive groups are “in close proximity” to one another. This effect also operates in enzyme-catalyzed reactions because of the binding of substrate(s) close to the catalytic groups at the enzyme active site.

One example is illustrated in Fig 5. Acid-catalyzed hydrolysis of sugar glycosides occurs via initial protonation of the departing hydroxyl group. An intramolecular example of glycoside hydrolysis is shown, in which a carboxylic acid group is positioned close to the departing hydroxyl group. The close proximity of the carboxylic group increases the local concentration of H$$^+$$, which increases the probability of the desired reaction.
Chemical Strategies for Enzyme Catalysis

![Figure 4](image_url) Alkaline phosphatase.

![Figure 5](image_url) Rate accelerations in glycoside hydrolysis.

A reaction and leads to a 10^4-fold increase in rate of glycoside hydrolysis. In the enzyme β-galactosidase, the substrate is bound in close proximity to a catalytic glutamic acid residue, which protonates the leaving group. The turnover number (k_cat) of the enzyme-catalyzed reaction is 10^3-fold higher than the intramolecular reaction, which indicates that the enzyme can employ additional strategies to achieve rate acceleration.

Another way of bringing reactants into close proximity, which is encountered commonly in transition metal chemistry, is through metal ion complexation. The coordination of a reactant to a metal ion complex often activates its reactivity and can bring the reactants into close proximity with a second reactant or with a catalytic group. One example, shown in Fig. 6, is a zinc (II) complex of 1,5,9-triazacyclononane, as a model for the enzyme carbonic anhydrase, which contains a zinc (II) cofactor in its active site (4). In the aqua complex, the bound water molecule has a dramatically reduced pK_a value of 7.3, which is similar to the pK_a of the active site nucleophilic water. The corresponding cobalt (III) complex catalyzed ester hydrolysis at twice the rate because Co(III) can coordinate both the hydroxide nucleophile and the ester carbonyl via a 5-coordinate intermediate, which is shown in Fig. 6. Many enzymes use metal ion cofactors, which can bind cosubstrates at a metal ion center, close to active site catalytic residues.

**Acid/Base Catalysis In Enzymatic Reactions**

All enzyme-catalyzed reactions that involve proton transfer use acid or base catalysis, so most enzyme active sites contain acidic or basic amino acid side chains that participate in catalysis. Because enzyme-catalyzed reactions take place close to pH 7, only fairly weak acids and bases are available, as shown in Fig. 7. Amino acid side chains with pK_a values below 7, such as aspartic acid or glutamic acid, will be deprotonated at pH 7 and, therefore, will be used normally in general base catalysis. Amino acid side chains with pK_a values above 7, such as lysine or tyrosine, will be protonated at pH 7 and, therefore, will be used normally in general acid catalysis. The imidazole side chain of histidine has a pK_a value of 6–8 and, therefore, might
Chemical Strategies for Enzyme Catalysis

Figure 6. Zinc (II) and cobalt (III) complexes of 1,5,9-triazacyclododecane as mimics of carbonic anhydrase.

Figure 7. Amino acid side chains involved in acid/base chemistry.

be either protonated or deprotonated (or a proportion of each) at pH 7; thus, histidine can be employed as either acid or base in enzyme catalysis and is found widely as an acid/base group in enzymes.

The actual pK_a value of an active site catalytic group will be influenced by the particular microenvironment of the active site, which could raise or lower the pK_a. For example, the enzyme acetoacetate decarboxylase contains an active site lysine residue that forms an imine link with its substrate; its pK_a value was found to be 5.9, which is much less than the expected value of 9. Adjacent to this residue in the active site is a second lysine residue, which in protonated form destabilizes the protonated amine and, therefore, reduces the pK_a. Conversely, aspartic acid or glutamic acid residues that are positioned in hydrophobic active sites can have increased pK_a values near 7 because the anionic form of the side chain is destabilized.

Enzyme active sites frequently have a pair of acid/base groups, one of which deprotonates one part of the substrate while the other protonates another part of the molecule; this dual action is known as bifunctional catalysis. One example is the enzyme ketosteroid isomerase, whose active site contains two catalytic residues; aspartate-38 acts as a catalytic base, and tyrosine-14 acts as an acidic group (5). The mechanism involves the formation of a dienol intermediate via a concerted step that involves simultaneous deprotonation of the substrate by Asp-14 and protonation of the substrate carbonyl by Tyr-14, as shown in Fig. 8.

Finally, enzymes that bind metal cofactors such as Zn^{2+} and Mg^{2+} can use their properties as Lewis acids, for example, electron pair acceptors. An example is the enzyme thermolysin, whose mechanism is illustrated in Fig. 9. In this enzyme, glutamate-143 acts as an active site base to deprotonate water for attack on the amide carbonyl, which is at the same time polarized by coordination by an active site Zn^{2+} ion (6). The protonated glutamic acid then probably acts as an acidic group for the protonation of the departing amine.
Nucleophilic Catalysis in Enzymatic Reactions

Several enzyme-catalyzed reactions involve the nucleophilic attack of an active site amino acid side chain on the substrate, which results in an immediate reaction that is attached covalently to the enzyme. This type of catalysis is known as nucleophilic (or covalent) catalysis.

Several different amino acid side chains can act as nucleophiles in enzyme catalysis. The most powerful nucleophile is the thiol side chain of cysteine, which can be deprotonated to form the even more nucleophilic thiolate anion. One example in which cysteine is used as a nucleophile is the enzyme glyceraldehyde 3-phosphate dehydrogenase, which uses the redox coenzyme NAD$^+$. As shown in Fig. 10, the aldehyde substrate is attacked by an active site cysteine, Cys-149, to form a hemi-thioketal intermediate, which transfers hydride to NAD$^+$ to form an oxidized thioester intermediate (7). Attack of phosphate anion generates an energy-rich intermediate 3-phosphoglycerate.

The ε-amino group of lysine is used in several enzymes to form imine links with ketone groups; for example, it is used in acetoacetate decarboxylase, shown in Fig. 11. Treatment of this enzyme with substrate and sodium borohydride leads to irreversible enzyme inactivation via in situ reduction of the enzyme-bound imine intermediate by borohydride, which indicates that a covalent link is formed. As mentioned, the $pK_a$ of this lysine group is abnormally low at 5.9, which is sufficiently low for it to act as a nucleophile at pH 7 (8).

The other nitrogen nucleophile available to enzymes is the versatile imidazole ring of histidine. This group is used more often for acid/base chemistry, but it is used occasionally as a nucleophile, for example, phosphotransfer reactions. The serine proteases, such as α-chymotrypsin, which is illustrated in Figs. 2a and 2b, are classic examples of the participation of serine as a nucleophile. Additional examples exist of nucleophilic mechanisms that employ the hydroxyl groups of threonine and tyrosine and the carboxylate groups of aspartate and glutamate.
Substrate Desolvation

Enzyme active sites sometimes are hydrophobic, buried sites that are excluded largely from water molecules. In these cases, the substrate(s) and enzyme catalytic groups are likely to be “desolvated.” In aqueous solution, a charged nucleophile is surrounded by several layers of water molecules, which greatly reduces its polarity and reactivity. However, a desolvated nucleophile at a water-excluded active site will be a much more potent nucleophile than its counterpart in solution. This effect particularly benefits nucleophilic substitution reactions, found in nucleophilic catalysis.

One recently discovered example is a “fluorinase” enzyme that is involved in the biosynthesis of fluoroacetic acid in Streptomyces cattleya. This enzyme catalyzes the nucleophilic...
the reaction will be accelerated (bound conformation and the transition state will be reduced and formation. In these cases, the difference in energy between the reactions, the enzyme binds the substrate in a strained conformation, which is closer to the transition state than the ground state conformation. In these cases, the difference in energy between the bound conformation and the transition state will be reduced and the reaction will be accelerated (Fig. 13).

One example of this is the enzyme lysozyme, which catalyzes the hydrolysis of the polysaccharide chain of the peptidoglycan layer of bacterial cell walls and, therefore, protects organs such as the human eye from bacterial infection. When the X-ray crystal structure of this enzyme was solved, attempts to model the substrate into the active site were possible only if a “kink” in the substrate was made at the point at which hydrolysis took place (10). The consequence of this kink was a flattening of the sugar ring that was cleaved; this flattened conformation is closer in structure to the oxonium ion formed during C-O bond cleavage, as shown in Fig. 14. Energetically, the enzyme compensated for the local flattening of this ring by more favorable binding interactions elsewhere in the substrate so that substrate binding still was a favorable process, but the activation energy for the bound substrate was reduced by this effect.

The Use of Strain Energy in Enzyme Catalysis

It is thought that in a small number of enzyme-catalyzed reactions, the enzyme binds the substrate in a strained conformation, which is closer to the transition state than the ground state conformation. In these cases, the difference in energy between the bound conformation and the transition state will be reduced and the reaction will be accelerated (Fig. 13).

One example of this is the enzyme lysozyme, which catalyzes the hydrolysis of the polysaccharide chain of the peptidoglycan layer of bacterial cell walls and, therefore, protects organs such as the human eye from bacterial infection. When the X-ray crystal structure of this enzyme was solved, attempts to model the substrate into the active site were possible only if a “kink” in the substrate was made at the point at which hydrolysis took place (10). The consequence of this kink was a flattening of the sugar ring that was cleaved; this flattened conformation is closer in structure to the oxonium ion formed during C-O bond cleavage, as shown in Fig. 14. Energetically, the enzyme compensated for the local flattening of this ring by more favorable binding interactions elsewhere in the substrate so that substrate binding still was a favorable process, but the activation energy for the bound substrate was reduced by this effect.

Stabilization of Reactive Intermediates

One remarkable feature of enzyme-catalyzed reactions is the ability to generate at enzyme sites reactive intermediates, such as carbanions, carbocations, and radicals, that normally would require strong reaction conditions to generate in chemical reactions and would be very unstable in aqueous solution.

Terpene cyclase enzymes catalyze the cyclization of allylic pyrophosphate substrates to form carbocyclic products via carbocation reaction intermediates. One well-studied example is pentalenene synthase (11,12), which catalyzes the cyclization of farnesyl pyrophosphate to give pentalenene, whose reaction mechanism is shown in Fig. 15. Cyclization of farnesyl pyrophosphate is proposed to form an 11-membered intermediate, humulene, which is followed by a five-membered ring closure to form a bicyclic tertiary carbocation. 1,2-Hydride migration followed by an additional five-membered ring closure gives a tricyclic carbocation, which gives pentalenene, at elimination.

How do these cyclase enzymes control the precise regiochemistry and stereochemistry of these multistep cyclizations? The active site of pentalenene synthase consists of a hydrophobic cleft, which is lined with aromatic and nonpolar residues. It is thought that the carbocation intermediates might be stabilized by the formation of π-cation interactions, with aromatic residues such as phenylalanine, tyrosine, and tryptophan. In pentalenene synthase, replacement of Phe-76 or Phe-77 by Ala gave >10-fold reduction in activity, which suggests that they may stabilize carbocationic intermediates through π-cation interactions.

One example of an enzyme-catalyzed reaction involving a radical intermediate is the enzyme ribonucleotide reductase, which catalyzes the conversion of ribonucleotides (used for RNA biosynthesis) to 2'-deoxyribonucleotides (used for DNA biosynthesis), as illustrated in Fig. 16. Spectroscopic studies of the R2 subunit of Escherichia coli ribonucleotide reductase have shown that it can form a stable, long-lived, tyrosyl radical species—the first protein radical to be discovered (13).

Single electron transfers within the protein lead to the formation of a cysteine radical on Cys-439 in the enzyme active site, which abstracts, then, the C-3 hydrogen to initiate a radical mechanism, as shown in Fig. 17. Protonation of the C-2-
Figure 14  Lysozyme mechanism.

Figure 15  Mechanism of pentalenene synthase.

Figure 16  Ribonucleotide reductase reaction; structure of tyrosyl radical.
The Involvement of Protein Dynamics in Enzyme Catalysis

Examination of protein structure in solution by nuclear magnetic resonance spectroscopy has revealed that a significant amount of internal motion exists in a protein on a timescale of 1 to 10 ns. Such internal motion could transmit kinetic energy from a distant part of the protein to the active site to assist in catalysis. It has been proposed that dynamic fluctuations in the protein structure are used by enzymes to organize the enzyme-substrate complex into a reactive conformation.

One example is in the enzyme dihydrofolate reductase, where replacement of Gly-120 to valine disrupts the internal motion of a protein loop, as illustrated in Fig. 18, which leads to a 500-fold reduction in rate; this reduction in rate implies that internal motion is involved in catalysis (14).

In conclusion, enzymes use a variety of strategies to achieve high rates of catalysis. Transition state stabilization seems to be the dominant factor in catalysis, but in some enzymes, the more sophisticated strategies such as substrate destabilization and protein dynamics seem to play an important role.

References
Chemistry of CoA-Dependent Enzymes

Dale G. Drueckhammer, Department of Chemistry, State University of New York at Stony Brook
doi: 10.1002/9780470048672.wecb093

Coenzyme A (CoA) and its thioesters play a diverse array of roles in biologic systems. The enzymes that catalyze reactions of CoA are of interest for a variety of reasons, including, but not limited to, their potential as drug targets. The enzymology of CoA biosynthesis is now well understood, and the genes for all of the enzymes involved have been identified. Thioesters are inherently reactive toward acyl transfer reactions and toward reactions involving deprotonation of the α-carbon, and these are the primary reactivity patterns in enzyme-catalyzed reactions of CoA thioesters. CoA utilizing enzymes have been widely studied mechanistically and structurally. Analogs of the natural CoA thioester substrates have been widely used in these studies, including thioesters of unnatural or uncommon acyl groups as well as a large number of analogs in which the thioester is replaced with alternative functionality. Recent technical applications of CoA have included the tagging of carrier protein domains and carrier protein fusions with tagged phosphopantetheine derivatives transferred enzymatically from the corresponding tagged CoA derivatives using promiscuous phosphopantetheinyl transferases.

Coenzyme A (CoA) is a cofactor that has been estimated to be used by about 4% of all enzymes, although more recent analysis of the BRENDA database (http://www.brenda.uni-koeln.de/) suggests the number may be closer to 9% (1). The biochemical pathways and processes involving CoA thioesters are diverse and widespread, whereas the kinds of reactions involved primarily follow the inherent reactivity of the thioester functionality. This article provides a brief overview of CoA biosynthesis and a summary of the common types of reactions of CoA thioesters. Also presented is a brief introduction to structural studies and a more extensive description of some types of analogs of natural CoA thioesters that have been employed as mechanistic probes for CoA using enzymes. The application of CoA derivatives and CoA biosynthetic enzymes for the tagging of carrier proteins and carrier protein fusions is also described.

Biologic Background

Coenzyme A (abbreviated CoA or CoASH, \( \text{G} \)) was discovered by Lipmann in the 1940s, and its structure was first reported in 1953 (2, 3). The structure of CoA consists of 3′-phosphoadenosine and pantetheine, linked by a phospho-phosphate group (Fig 1). The pantetheine moiety is derived from pantetheine 5′-diphosphate (4), also known as vitamin B5, CoA, and its thioester derivatives are involved in a wide range of biologic processes, including the TCA or Krebs cycle, fatty acid biosynthesis and degradation, antibiotic resistance mechanisms, gene expression, hormone biosynthesis and regulation, and nerve impulse conductance. Numerous CoA utilizing enzymes are either established or potential drug targets. The enzymes HMG-CoA reductase and acyl-CoA cholesterol acyltransferase (ACAT), which are involved in cholesterol biosynthesis and metabolism, are targets of drugs for cholesterol management. Inhibitors of malonyl-CoA decarboxylase may act as cardioprotective agents (4), whereas fatty acid synthase inhibitors have been explored as antimicrobial agents (5). CoA thioesters may also be involved in fatty acid-induced insulin resistance (6). Acetyl-CoA carboxylase and other acyl-CoA carboxylases are potential targets for anticaner and antiobesity agents as well as are herbicidal targets (7). Thioesters of CoA serve as the building blocks for the polyketide synthases that make a wide variety of natural bioactive compounds and are also involved as building blocks and intermediates in the biosynthesis of natural bacterial polyesters (8, 9). The interesting organometallic mechanism of the nickel-containing enzyme carbon monoxide dehydrogenase, which forms acetyl-CoA from carbon monoxide, CoA, and a methyl group donor, has been the target of extensive fundamental mechanistic studies (10). CoA thioesters have also been studied in ribozyme-catalyzed reactions (11). Thus, the roles...
Chemistry of CoA-Dependent Enzymes

of CoA utilizing enzymes in biology and the reasons for their interest are widespread and diverse.

Chemistry and Reactivity of CoA

CoA is involved in the activation and transfer of acyl groups in a wide variety of enzymatic reactions. Whereas many coenzymes function as co-catalysts that remain bound to a single enzyme molecule, CoA acts as a diffusible carrier of acyl groups between different enzymes.

CoA biosynthesis

The biosynthesis of CoA from pantothenic acid is shown in Fig. 1. Extensive coverage of this topic and references to the original literature can be found in recent reviews (1, 12). Pantothenic acid 1 is phosphorylated by pantothenate kinase to form 2. At least three distinct types of pantothenate kinases from different organisms have been observed (1). The cysteamine moiety is introduced by initially coupling with L-cysteine catalyzed by the phosphopantothenylcysteine synthetase activity to form 3 followed by a decarboxylation reaction to form phosphopantetheine 4. These two steps are catalyzed by a single bifunctional enzyme in most prokaryotes but by distinct enzymes in higher organisms. The bacterial phosphopantothenylcysteine synthetase is CTP-dependent, whereas the plant and mammalian enzymes prefer ATP. 3 is not inherently prone to decarboxylation, but this reaction has recently been shown to occur by temporary oxidation of the thiol to the thioaldehyde, which readily undergoes decarboxylation to an enethiolate intermediate (13). Phosphopantetheine is coupled with ATP to form dephospho-coenzyme A 5 followed by phosphorylation of the 3′hydroxyl group to form CoA. The final two steps are catalyzed by a single bifunctional enzyme in mammals, although they are catalyzed by separate enzymes in bacteria (1). The genes coding for all of the enzymes of this pathway (coaA-E) have now been identified, as have the genes for the four enzymes of pantothenic acid biosynthesis in bacteria (1, 12, 14). Pantothenic acid biosynthesis has also been studied in plants (15). Pantetheine kinase is a key regulatory point in CoA biosynthesis and is subject to feedback inhibition by CoA in most species (1). Pantetheine kinase can also catalyze the phosphorylation of pantetheine and pantetheine analogs (16, 17), although phosphorylation occurs before coupling with L-cysteine in the biosynthetic pathway.

Reactions and processes involving CoA

The reactions of CoA involve only the thiol group and the acyl moieties attached to the thiol group as thioesters, with the remainder of the structure serving as a recognition element that facilitates binding to the appropriate enzymes. The notable exception is the phosphopantetheinytransferases that catalyze transfer of the phosphopantetheine moiety of CoA to a serine...
Inherent reactivity of thiols and thioesters

Thiols are structurally similar to alcohols and can generally undergo the same kinds of reactions. However, a typical thiol group (pk of 9-10) is substantially more acidic than the corresponding hydroxyl group (pk ca 16). The thiolate ion is also much more nucleophilic than the corresponding alkoxide. This combination of acidity to form the thiolate and nucleophilicity of the thiolate facilitates both alkylation and acylation reactions of thiols, with acylation being the primarily relevant reaction of CoA. A thioester has two inherent modes of reactivity that are both observed in enzyme-catalyzed reactions of CoA thioesters. One common type of reaction is initiated by reaction of a basic group to remove a proton from the carbon alpha to the carbonyl to form an enolate nucleophile, which then reacts with an electrophilic substrate. The other common type of enzymatic reaction of CoA thioesters involves acyl transfer by attack of an acyl acceptor nucleophile at the thioester carbonyl carbon. The inherent characteristics of the thioester make it more reactive than esters or amides in both of these types of reactions. The α-protons of a typical thioester (pk ca 23) are about 104-fold more acidic than those of an ester (pk ca 25), whereas the α-protons of an amide are even less acidic than those of an ester (19). The free energy of hydrolysis of a thioester is about 2 kcal/mol greater than that of an ester and, again, the difference relative to an amide is even greater (20), which provides a thermodynamic driving force in acyl transfer from CoA to an alcohol or amine nucleophile, whereas acyl transfer reactions to carbanion nucleophiles are also common. Thioesters are also kinetically much more reactive than esters toward most nucleophiles including thiols, amines, and carbamions, with amides being much less reactive even than esters (21, 22). However, the rates of reaction of esters and thiesters toward hydrolysis by aqueous base are essentially identical, with thiesters even appearing to be slightly less reactive than oxesters in some examples (21). This unique reactivity of thiocarboxylates provides for their reasonable stability in aqueous solution despite their substantial thermodynamic reactivity and inherently high kinetic reactivity toward most nucleophiles.

Acyl-CoA synthetases and related enzymes

Acetyl-CoA is formed from CoA and acetate by the enzyme acetyl-CoA synthetase, an ADP-forming ligase. Phosphotrans-acylase forms acetyl-CoA from CoA and acetyl-phosphate, which in turn is formed from acetate and ATP catalyzed by acetate kinase. Other enzymes that can form acetyl-CoA from CoA and other acetyl group donors include ATP citrate lyase and thiolase. Longer chain acyl-CoA thioesters are typically formed from CoA and a fatty acid catalyzed by ligases generally known as acyl-CoA synthetases.

Acyltransferases

Most acyltransferases catalyze acyl transfer to a hydroxyl or amine group of the acceptor substrate, whereas CoA ester hydrolysis by thioestersases is in effect acyl transfer to water. Choline acetyltransferase catalyzes transfer of the acetyl group from acetyl-CoA to the hydroxyl group of choline. The product acetylcholine is a major neurotransmitter (23). Serotonin acetyltransferase controls the sleep cycle by catalyzing acetyl transfer to the primary amine group of serotonin, which is the rate-limiting step in metabolism biosynthesis (24). Chloramphenicol acetyltransferase catalyzes acetyl transfer to a primary hydroxyl group, which destroys the antibiotic activity of chloramphenicol thereby conferring antibiotic resistance, whereas the antibiotics gentamicin and kanamycin are similarly inactivated by acetyltransferases (25). Other extensively studied acetyltransferases include the histone acetyltransferases (HATs), which mediate a major control element in gene expression (26). A cytochrome P450 also catalyzes acyl transfer from longer chain acyl-CoA. The highly studied protein N-myristoyltransferase catalyzes transfer of the myristoyl group to the amine group of N-terminal glycine of a protein substrate (27). The palmitoyltransferases catalyze transfer of palmitic acid to a cysteine thiol group of certain proteins (28). These post-translational acyltransfer reactions are important in membrane anchoring of proteins and may play a role in signaling events (29). Carnitine palmitoyltransferase catalyzes acyl transfer to the hydroxyl group of carnitine (30). The two general mechanisms for acyltransferases are a direct transfer to the nucleophilic acceptor or a two-step process involving initial acyl transfer to a nucleophilic group (usually a cysteine thiol) of the enzyme to form an acyl enzyme intermediate. Examples of both types of mechanisms are well documented, although the direct transfer mechanism appears to be more common.

Claisen enzymes

The enzymes that form a nucleophilic enolate intermediate by deprotonation of the methyl group of acetyl-CoA, which then reacts with a second electrophilic substrate, are generally referred to as Claisen enzymes in the literature, although not all catalyze true Claisen condensation reactions. Some common examples of Claisen enzymes and their electrophilic substrates are shown in Fig. 2. Thiocarbamylates a true Claisen condensation reaction in which the electrophilic substrate is a second equivalent of acetyl-CoA, whereas in enzymes recognizing longer chain substrates, the electrophilic substrate can be a medium or long chain acyl-CoA (31). From the perspective of the electrophilic substrate, thiolase is also an acyltransferase in which the acyl group is transferred to a nucleophilic carbon. Acetyl-CoA carboxylase catalyzes transfer of the carboxylate group from carboxybiotin to the enolate (32). Both citrate synthase and HM-CoA synthase catalyze reactions of the enolate with an electrophilic carbonyl group of a ketone substrate in what is formally an aldol reaction (33). In both of these examples, the enzyme also catalyzes a subsequent thioester hydrolysis reaction to release free CoA. Malate synthase catalyzes a very similar reaction (34). Although an enol intermediate for the Claisen enzymes and other mechanistically related enzymes has gained some consideration, an enolate intermediate stabilized by hydrogen bonding or in some cases possibly metal ion coordination to the carbonyl oxygen seems generally preferred (35). The low barrier or short-strong hydrogen bond proposal would favor the
proton of the hydrogen bond being equally shared between the substrate oxygen and the hydrogen bond donor group (36).

Enzymes catalyzing reactions of the acyl moiety
Certain CoA thioester using enzymes catalyze reactions at the $\beta$-carbon or other carbons of the acyl group more distant from the thioester functionality. The fatty acid $\beta$-oxidation cycle provides some examples (Fig. 3). Fatty acids enter the cycle by initial conversion to the CoA ester which is then oxidized to the $\alpha,\beta$-unsaturated thioester catalyzed by the enzyme crotonase, which is the centerpiece of the crotonase superfamily of enzymes that catalyze related reactions (37), which is followed by oxidation of the alcohol to form the $\beta$-keto thioester. Addition of water to the double bond to form the $\beta$-hydroxy thioester is catalyzed by the enzyme crotonase, which is the centerpiece of the crotonase superfamily of enzymes that catalyze related reactions (37), which is followed by oxidation of the alcohol to form the $\beta$-keto thioester. A retro-Claisen reaction catalyzed by thiolase forms acetyl-CoA along with a new acyl-CoA having a carbon chain two carbons shorter than in the initial or previous cycle.

Chemical Tools and Techniques

Structural studies of CoA utilizing enzymes
There is no single sequence motif common to all CoA utilizing enzymes, although homology has been observed across some fairly broad subsets. X-ray crystallography has been a powerful tool in the structural elucidation of CoA utilizing enzymes, and some structural studies have also been performed using NMR. Crystal structures for several of the enzymes of CoA biosynthesis have been solved while the solved structures of different CoA and CoA ester utilizing enzymes in the protein data bank numbers near 100, not including multiple structures of different complexes of the same protein and of the same protein from multiple sources. The conformation of CoA thioesters in solution has also been studied by NMR and comparisons made to the enzyme-bound conformation (38).

Applications of analogs of natural CoA thioesters
Synthetic analogs of natural CoA ester substrates have been widely used as mechanistic probes in studies of CoA ester utilizing enzymes (39). The most readily available are CoA thioesters of unnatural or uncommon acyl groups prepared by simple acylation of CoA. A number of haloacyl-CoA derivatives have been prepared, including fluoroacyl-CoA (40). Fluoroacyl-CoA is accepted as a substrate by citrate synthase to form fluorocitrate, which is an inhibitor of aconitase. This enzymatic processing to fluorocitrate and resulting inhibition of aconitase is responsible for the high toxicity of fluorocitrate. Bromoacyl-CoA was shown to inhibit carnitine acetyltransferase in the presence of carnitine as shown in Fig. 4. Enzyme-catalyzed transfer of the bromoacyl group to carnitine to form bromoacyl carnitine is followed by bromide displacement by the nucleophilic thiol group of CoA to form the bisubstrate adduct, which is the actual inhibitory species that binds tightly to the enzyme (41). Similar bisubstrate adducts have been employed as inhibitors of serotonin acetyltransferase (24). Other halo-acyl CoA thioesters as well as CoA thioesters having an epoxide in the acyl moiety have been employed as reactive electrophiles to trap nucleophilic residues in the active site of their target enzymes.

CoA thioesters of unsaturated acids have also been widely used. Hexadienoyl-CoA, cinnamoyl-CoA, and other $\alpha,\beta$-unsaturated acyl-CoAs having extended conjugation or having a heteroatom in place of the $\gamma$-carbon have been employed as spectroscopic probes of their enzyme complexes. The additional unsaturation or the heteroatom substituent serve to shift the UV absorbance to longer wavelength and, in some cases, also shifts the equilibrium of a reversible enzyme-catalyzed reaction toward the $\alpha,\beta$-unsaturated substrate (42, 43). 3-Alkynoyl-CoA were used as mechanism-based inhibitors of thiolase, with enzyme-catalyzed isomerization to the allenoyl-CoA being followed by conjugate addition of an active site nucleophile to form the inactive enzyme adduct (44).
Chemistry of CoA-Dependent Enzymes

Figure 3  The fatty acid β-oxidation cycle. E1: acylcoenzyme A synthetase; E2: acylcoenzyme A dehydrogenase; E3: enoylcoenzyme A hydratase (crotonase); E4: β-hydroxyacylcoenzyme A dehydrogenase; E5: thiolase.

Figure 4  Some unnatural CoA thioesters and related analogs. E1: carnitine acetyltransferase; E2: thiolase.

... was preceded by studies of inactivation of the dehydratase component of fatty acid synthase by 3-alkynoyl thioesters (45). In later work, simplified 3-alkynoyl pantetheine thioesters were employed similarly to identify the active site base in thiolase (46). Similar compounds have also been used to inactivate acyl-CoA dehydrogenases and enoyl reductase (47, 48). Several CoA thioesters of acids bearing cyclopropyl groups have been studied with special interest in those that undergo cyclopropyl ring opening of radical intermediates (49-51). Various other CoA thioesters bearing functionality, including photoaffinity labels (52, 53) and nitroxide spin labels (54) in the acyl moiety, have been prepared and employed as mechanistic tools. The dihydroxy analogs of acetyl-CoA and of fluoroacetyl-CoA and octanoyl-CoA were prepared and shown to have interesting spectral properties, with a λ max of 306 nm and enhanced acidity of the α-protons (pKa = 12.5) (55). CoA dithioesters have also been employed in studies of thiolase, HMG-CoA lyase and HMG-CoA reductase. Structure 22 is illustrative of a number of keto thioether analogs in which a methylene group is inserted between the sulfur and carbonyl carbon of the thioester. Such analogs are prepared by alkylation of CoA with a halomethyl ketone. These analogs do not have a cleavable bond between the CoA and acyl moiety and are often good inhibitors of acyltransferase enzymes (56). CoA has also been derivatized in the adenine base for attachment to a solid support for use in affinity chromatography (57, 58).

CoA thioester analogs have also been prepared that cannot be made by derivatization of natural CoA. Some early examples were made by nonenzymatic synthesis (59) generally following the original synthesis of CoA developed by Moffatt and Khorana (60). More recently analogs have been made enzymatically using the enzymes of CoA biosynthesis. One approach uses pantetheine kinase, phosphopantetheine adenylyltransferase, and dephospho-CoA kinase to convert synthetic pantetheine analogs (e.g., 23 to the corresponding CoA analogs, with phosphoenolpyruvate and pyruvate kinase included to regenerate the ATP consumed in the kinase steps as shown in Fig. 5 for the synthesis of an oxoester analog 26 of crotonyl-CoA (16). The three biosynthetic enzymes have been shown to accept substrate analogs having significant modification relative to the natural
substrates, including analogs having long chain acyl groups. This synthetic method takes advantage of the ability of pantetheine kinase to phosphorylate pantetheine and its analogs and derivatives in addition to the natural substrate pantothenic acid. The other primary synthetic approach has used the final enzymes of CoA biosynthesis to make a CoA analog in which the outermost amide bond is replaced with a thioester group (61). This analog serves as a general synthon for other CoA analogs by reaction with a primary amine bearing the functionality of interest in place of the thiol group of CoA (Fig. 6).

The oxoester analog of crotonyl-CoA shown in Fig. 5 was used as an alternative substrate for crotonase and exhibited about 300-fold decreased activity relative to the natural thioester (16). Some other representative analogs are shown in Fig. 6. Several dethia analogs in which the thioester sulfur atom is replaced with a methylene group have been prepared, as represented by the acetyl-CoA analog (59, 61). These analogs are generally inhibitors of enzymes that catalyze reactions involving cleavage of the thioester such as acyltransferases, while serving as substrates for enzymes catalyzing reactions that do not involve cleavage of the thioester such as some of the Claisen enzymes. The carboxylate analog is representative of compounds prepared as stable mimics of the enolate (or enol) intermediate of the Claisen enzymes (61, 62). It was shown to be a potent inhibitor of citrate synthase, binding about 1000-fold more strongly than acetyl-CoA, and the structure of the enzyme-inhibitor complex was solved (33). The analog 31...
having the orientation of the thioester reversed was shown to be an inhibitor of thiolase, apparently forming an acyl-enzyme with the CoA moiety in a step mimicking the formation of the acetyl-enzyme intermediate from acetyl-CoA (63).

**Other Technical Applications of CoA**

A recent interesting application of CoA and CoA analogs has been in protein labeling. Initially, a CoA molecule derivatized on the thiol group with a maleimide-linked reporter group was employed as a substrate for a phosphopantetheinyl transferase. The reporter-modified phosphopantetheine moiety was thereby enzymatically transferred from the CoA derivative 35 to the apo-form of the carrier protein of polyketide synthases and nonribosomal peptide synthetases to form the labeled holo-carrier protein analog 36 (Fig. 7) (64). The reporter group was either a fluorophore, for fluorescent visualization of the reporter-modified protein or biotin or other affinity tag to facilitate isolation of the labeled protein. In subsequent developments of this technology, it has been applied to the labeling of fusion proteins between the relatively small carrier protein and another protein, the fused carrier protein domain thereby directing labeling to the protein of interest (65). This methodology has also been applied to the labeling of fusion proteins on cell surfaces (66). More recently, it has been demonstrated that the CoA-reporter derivatives 35 can be prepared from synthetic pantetheine-reporter conjugates 32 using the pantothenate kinase, phosphopantetheine adenyltransferase, and dephospho-CoA kinase enzymes of CoA biosynthesis (67). The enzymes are quite promiscuous in converting analogs containing only the pantolic amide moiety of the natural pantetheine or pantetheine substrate, generally with a flexible tether between the amide and the labeling group of interest. Pantolic amide analogs containing reactive acide, alkyne, and ketone functionality as well as biotin and fluorescent labels have been employed to incorporate the functionality into the protein substrate of the phosphopantetheinyl transferase. The significance of the enzymatic generation of the CoA analog substrates for the phosphopantetheinyl transferase is that simple cell-permeable pantolic amide analogs can be processed all the way to the labeled carrier protein or fusion protein thereof in vivo. A set of bioorthogonal pantolic amide analogs have been developed for use in this in vivo protein tagging method (17). Other recent work has demonstrated that a simple 11-residue peptide incorporated into the targeted protein can be recognized by the phosphopantothenyl transferase Sfp from Bacillus subtilis (68), which promises to be a very general and versatile method for in vivo protein tagging and labeling.

**References**

Chemistry of CoA-Dependent Enzymes


See Also

Enzyme Catalysis, Chemical Strategies for Enzyme Co-factors, Chemistry of Vitamins, Chemistry of Small Molecules to Elucidate Biological Function


Collagen Triple Helix,
Stability of

Hans Peter Bächinger and Kazunori Mizuno, Shriners Hospital for Children, Portland, Oregon
doi: 10.1002/9780470048672.wecb096

Collagens are the most abundant extracellular matrix proteins in multicellular animals. They all contain a name-giving collagen triple helix, which connects their three chains and varying amounts of other noncollagenous protein domains. To form the triple helix a repeated sequence of -Gly-Xaa-Yaa- is required, where Xaa and Yaa can be any residue. Each chain forms a polyproline-II like left-handed helix. The three chains are staggered by one residue from each other, and form a right-handed helix. Twenty eight types of collagen molecules have been identified in mammals. The stability of the collagen triple helix is based on the length and the amino acid sequence of each polypeptide chain, and also by the presence of interchain cross-links and/or trimerization domains. The 4(R)-hydroxylation of proline residues in the Yaa position significantly increases the stability of the collagen triple helix.

Collagens are proteins that have a typical triple helical higher order structure, and the triple helix is a protein motif that can be found in other proteins. The chemistry of collagen covers more than seven orders of magnitude from subnanometer scale to centimeter scale (from amino acids to a tendon) including a wide range of both noncovalent and covalent interactions. Collagen research covers a wide range of fields such as biochemistry, organic chemistry, and biophysical chemistry.

Biologic Background

Collagen is the major component of the extracellular matrix of multicellular animals. In humans, the collagen molecules are classified into 28 types (Table 1). In addition, other proteins contain collagen-like triple helical domains (Table 2). Most mammals have a similar set of these proteins. Collagens in vertebrates are numbered in the order of their discovery using Roman numerals as type I, type II, type III, and so on. In invertebrates, collagens vary from one species to another including differences in posttranslational modifications. The Caenorhabditis elegans genome has upward of 150 distinct collagen genes (1). If a collagen molecule of an invertebrate is highly homologous to a specific type of vertebrate collagen, then it is referred to the specific number type of the vertebrate collagen. For example, one of the basement membrane collagens, type IV collagen, is found in both vertebrates and also invertebrates such as Drosophila melanogaster and C. elegans as highly homologous primary structures. In that case, the molecule is called a type IV collagen. In general, the gene of collagen molecules of invertebrates are different from those of vertebrates. Most proteins with a collagen triple helix were found as a component of the extracellular matrix (ECM). However, some proteins contain a transmembrane domain, for example, type XVII collagen molecule is a type II transmembrane protein in hemidesmosomes. In addition, some triple helical proteins are found in the serum as soluble proteins such as collectins and complement C1q. Usually, they are related to natural innate immunity. The most abundant collagen protein in vertebrates is type I collagen. It is distributed ubiquitously in the vertebrate body and forms fibrils with a variety of diameter (25 nm–200 nm). Type I collagen molecules self assemble at physiologic temperature to fibrils in vitro. When type I collagen molecules are denatured into single polypeptides at higher temperatures, they form a gel at lower temperature (gelatin), which consists of partially refolded molecules. Type I collagen interacts with a wide range of other extracellular matrix molecules to form the specific tissue. The local concentration of the collagen molecules, pH, temperature, the direction of mechanical force, and the order of interactions with other molecules affect the supramolecular organization of the collagen fibrils and the entire ECM. Tendon, skin, and cornea are composed mainly of type I collagen, but their physical properties, such as the length and the diameter of collagen fibrils and the direction of collagen fibrils are different. Cells and ECM interact with each other. The homeostasis of cell-ECM interaction is essential for all biologic phenomena of multicellular animals.

Collagens are the most abundant extracellular matrix proteins in multicellular animals. They all contain a name-giving collagen triple helix, which connects their three chains and varying amounts of other noncollagenous protein domains. To form the triple helix a repeated sequence of -Gly-Xaa-Yaa- is required, where Xaa and Yaa can be any residue. Each chain forms a polyproline-II like left-handed helix. The three chains are staggered by one residue from each other, and form a right-handed helix. Twenty eight types of collagen molecules have been identified in mammals. The stability of the collagen triple helix is based on the length and the amino acid sequence of each polypeptide chain, and also by the presence of interchain cross-links and/or trimerization domains. The 4(R)-hydroxylation of proline residues in the Yaa position significantly increases the stability of the collagen triple helix.

Collagens are proteins that have a typical triple helical higher order structure, and the triple helix is a protein motif that can be found in other proteins. The chemistry of collagen covers more than seven orders of magnitude from subnanometer scale to centimeter scale (from amino acids to a tendon) including a wide range of both noncovalent and covalent interactions. Collagen research covers a wide range of fields such as biochemistry, organic chemistry, and biophysical chemistry.

Biologic Background

Collagen is the major component of the extracellular matrix of multicellular animals. In humans, the collagen molecules are classified into 28 types (Table 1). In addition, other proteins contain collagen-like triple helical domains (Table 2). Most mammals have a similar set of these proteins. Collagens in vertebrates are numbered in the order of their discovery using Roman numerals as type I, type II, type III, and so on. In invertebrates, collagens vary from one species to another including differences in posttranslational modifications. The Caenorhabditis elegans genome has upward of 150 distinct collagen genes (1). If a collagen molecule of an invertebrate is highly homologous to a specific type of vertebrate collagen, then it is referred to the specific number type of the vertebrate collagen. For example, one of the basement membrane collagens, type IV collagen, is found in both vertebrates and also invertebrates such as Drosophila melanogaster and C. elegans as highly homologous primary structures. In that case, the molecule is called a type IV collagen. In general, the gene of collagen molecules of invertebrates are different from those of vertebrates. Most proteins with a collagen triple helix were found as a component of the extracellular matrix (ECM). However, some proteins contain a transmembrane domain, for example, type XVII collagen molecule is a type II transmembrane protein in hemidesmosomes. In addition, some triple helical proteins are found in the serum as soluble proteins such as collectins and complement C1q. Usually, they are related to natural innate immunity. The most abundant collagen protein in vertebrates is type I collagen. It is distributed ubiquitously in the vertebrate body and forms fibrils with a variety of diameter (25 nm–200 nm). Type I collagen molecules self assemble at physiologic temperature to fibrils in vitro. When type I collagen molecules are denatured into single polypeptides at higher temperatures, they form a gel at lower temperature (gelatin), which consists of partially refolded molecules. Type I collagen interacts with a wide range of other extracellular matrix molecules to form the specific tissue. The local concentration of the collagen molecules, pH, temperature, the direction of mechanical force, and the order of interactions with other molecules affect the supramolecular organization of the collagen fibrils and the entire ECM. Tendon, skin, and cornea are composed mainly of type I collagen, but their physical properties, such as the length and the diameter of collagen fibrils and the direction of collagen fibrils are different. Cells and ECM interact with each other. The homeostasis of cell-ECM interaction is essential for all biologic phenomena of multicellular animals.
### Table 1: Classification of collagen in mammals: Known and estimated information about collagen

<table>
<thead>
<tr>
<th>Type</th>
<th>Class</th>
<th>Alpha chain</th>
<th>Length of collagenous domains (human)</th>
<th>Chain composition</th>
<th>Main distribution</th>
<th>Supramolecular assembly</th>
<th>Another name</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>fibrillar</td>
<td>1(I)</td>
<td>1014</td>
<td>112</td>
<td>Ubiquitous</td>
<td>Major component of fibrils</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(I)</td>
<td>1014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>fibrillar</td>
<td>1(II)</td>
<td>1017</td>
<td>111</td>
<td>Hyaline cartilage</td>
<td>Major component of fibrils</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(II)</td>
<td>1029</td>
<td>111</td>
<td>Same as type I, little in bone and tendon</td>
<td>Minor component of fibrils</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>fibrillar</td>
<td>1(III)</td>
<td>1029</td>
<td>111</td>
<td>Same as type I, little in bone and tendon</td>
<td>Minor component of fibrils</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>basement</td>
<td>1(V)</td>
<td>1398</td>
<td>112, 345, 556</td>
<td>Basement membrane, sinusoid</td>
<td>Polygonal meshwork</td>
<td></td>
</tr>
<tr>
<td></td>
<td>membrane</td>
<td>2(V)</td>
<td>1428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(V)</td>
<td>1428</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4(V)</td>
<td>1405</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5(V)</td>
<td>1421</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6(V)</td>
<td>1417</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>fibrillar</td>
<td>1(V)</td>
<td>1014</td>
<td>112, 123, 1117</td>
<td>Ubiquitous, minor component of collagen fibril, abundant near basement membrane, thin collagen fibrils</td>
<td>Minor component of fibrils</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(V)</td>
<td>1014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(V)</td>
<td>1011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td></td>
<td>1(VI)</td>
<td>336</td>
<td>123</td>
<td>Ubiquitous</td>
<td>Beaded filaments</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(VI)</td>
<td>335</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(VI)</td>
<td>336</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>anchoring fibrils</td>
<td>1(VII)</td>
<td>1530</td>
<td>111</td>
<td>Dermal epidermal junction</td>
<td>Anti-parallel bundles (anchoring fibril)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(VII)</td>
<td>1530</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>short chain</td>
<td>1(VIII)</td>
<td>454</td>
<td>111, 222, 1127</td>
<td>Dermal connective tissue, blood vessels</td>
<td>Fine fibrils</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(VIII)</td>
<td>457</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IX</td>
<td>FACIT</td>
<td>1(X)</td>
<td>115 + 339 + 137</td>
<td>123</td>
<td>Cartilage</td>
<td>Bound on collagen fibril surface</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(X)</td>
<td>115 + 339 + 137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(X)</td>
<td>112 + 339 + 137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>short chain</td>
<td>1(X)</td>
<td>463</td>
<td>111</td>
<td>Hyaline cartilage</td>
<td>Fine fibrils</td>
<td></td>
</tr>
<tr>
<td>XI</td>
<td>fibrillar</td>
<td>1(IX)</td>
<td>1014</td>
<td>123</td>
<td>Minor component of fibril</td>
<td>Fine fibrils</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2(XI)</td>
<td>1014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3(XI) + 4(IX)</td>
<td>1014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XII</td>
<td>FACIT</td>
<td>1(XII)</td>
<td>103 + 152</td>
<td>111</td>
<td>Type I collagen rich tissues</td>
<td>Bound on collagen fibril surface</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Type</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XIII transmembrane 1(XIII)</td>
<td>95 + 172 + 209</td>
<td>Type I collagen rich tissues</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XIV FACIT 1(XIV)</td>
<td>106 + 149</td>
<td>Bound on collagen fibril surface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XV multiplexin 1(XV)</td>
<td>577 dc</td>
<td>Basement membrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XVI FACIT-like 1(XVI)</td>
<td>106 + 422 + 15 = 52 + 138 + 71 + 59 + 34 + 13 + 27</td>
<td>The C-terminal 185 aa part is retin.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XVII transmembrane 1(XVII)</td>
<td>988 dc</td>
<td>Skin, brain, hemidesmosome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XVIII multiplexin 1(XVIII)</td>
<td>688 dc</td>
<td>Basement membrane, BP180, Bullous pemphigoid antigen 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XIX FACIT-like 1(XIX)</td>
<td>70 + 168 + 108 = 224 + 144</td>
<td>Basement, eye, testis, basement membrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XX FACIT</td>
<td>103 + 155 (Chick)</td>
<td>Unknown, Bound on collagen fibril surface</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXI FACIT 1(XXI)</td>
<td>112 + 339</td>
<td>Blood vessel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXII FACIT 1(XXII)</td>
<td>109 + 339 + 234 + 374</td>
<td>Basement membrane</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXIII transmembrane 1(XXIII)</td>
<td>186 + 75 + 111</td>
<td>Developing corneal bone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXIV fibrillar 1(XXIV)</td>
<td>931</td>
<td>Brain, CLAC-P (collagen-like Alzheimer amyloid plaque component precursor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXV transmembrane 1(XXV)</td>
<td>44 + 238 + 189</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXVI fibrillar 1(XXVI)</td>
<td>69 + 33</td>
<td>Testis and ovary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXVII fibrillar 1(XXVII)</td>
<td>997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XXVIII fibrillar 1(XXVIII)</td>
<td>530</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*FACIT, fibril-associated collagens with interrupted triple helices; dc, discontinuous.*
<table>
<thead>
<tr>
<th>Name</th>
<th>COL domain (aa)</th>
<th>Other names</th>
<th>Class</th>
<th>Tissue of origin</th>
<th>Tissue of presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macrophage scavenger receptor I, II, III</td>
<td>69</td>
<td>SR-AI, SR-AII, SR-AIII (alternative splicing)</td>
<td>Transmembrane II, class A scavenger receptor</td>
<td>Macrophage</td>
<td>Macrophage</td>
</tr>
<tr>
<td>MARCO receptor</td>
<td>270</td>
<td></td>
<td>Transmembrane II, class A scavenger receptor</td>
<td>Macrophage</td>
<td>Macrophage of spleen, etc.</td>
</tr>
<tr>
<td>CL-P1/scavenger receptor with C-type lectin</td>
<td>147</td>
<td>SR-CL1, SR-CLII (alternative splicing)</td>
<td>Transmembrane II, class A scavenger receptor</td>
<td>Reptiles, vascular endothelial cells, nurse cells</td>
<td>Endothelial cells, thymus</td>
</tr>
<tr>
<td>Complement C1q (a chain)</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
<td>Serum (80 mg/L)</td>
</tr>
<tr>
<td>Complement C1q (b chain)</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement C1q (c chain)</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung surfactant protein A</td>
<td>73</td>
<td>SP-A, SP-35, SP28-35</td>
<td>Collectin</td>
<td>Lung epithelium, prostate, thymus, synovium, etc.</td>
<td>Bronchus, alveoli, mucosal surfaces, semen</td>
</tr>
<tr>
<td>Lung surfactant protein D</td>
<td>177</td>
<td>SP-D, CP4</td>
<td>Collectin</td>
<td>Lung epithelium, gastrointestinal epithelium, etc.</td>
<td>Bronchus</td>
</tr>
<tr>
<td>CL-L1</td>
<td>72</td>
<td>Collectin liver-1</td>
<td>Collectin</td>
<td>Tissues except skeletal muscle</td>
<td></td>
</tr>
<tr>
<td>Mannan-binding lectin</td>
<td>59</td>
<td>Mannan-binding lectin, MBL, MBP, ralf(p28)</td>
<td>Collectin</td>
<td>Liver (hepatocyte)</td>
<td></td>
</tr>
<tr>
<td>protein</td>
<td>location</td>
<td>organ</td>
<td>other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conglutinin 171</td>
<td>Collectin</td>
<td>Liver Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-43 114</td>
<td>Collectin</td>
<td>Liver Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-46 171</td>
<td>Collectin</td>
<td>Liver Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ectodysplasin A 58</td>
<td>Serum lectin P35, huculin, EBP-35, ficolin 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-ficolin 48</td>
<td>Ficolin 1, ficolin-alpha, ficolin-beta</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M-ficolin 57</td>
<td>Ficolin 33</td>
<td>H-ficolin 33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collectin Liver Serum</td>
<td></td>
<td>Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-43 114</td>
<td></td>
<td>Thymus, Liver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CL-46 171</td>
<td></td>
<td>Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H-ficolin 33</td>
<td></td>
<td>Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hibernation (HP-20, 25, and 27) proteins</td>
<td>30, 42, 39</td>
<td>Hibernation proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asymmetric form of acetylcholinesterase</td>
<td>188</td>
<td>AChE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adiponectin 60</td>
<td>ACRP30 (adipocyte complement-related protein of 30kDa), adipoQ, GB P28 (gelatin binding protein of 28kDa), apM1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adipocyte Serum</td>
<td>Serum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Number of human protein is indicated as regular fonts.
Conglutinin, CL-43, and CL-46 are found in bovine. Hibernation proteins are found in Tamias sibiricus.
Structure of the Collagen Triple Helix

The general sequence of the collagen triple helix requires a repeat of \(-\text{Gly-Xaa-Yaa}-\) where the Xaa and the Yaa residues can be any amino acid. Glycine in every third residue is required to form the triple helix. Figure 1 shows a model of the collagen triple helix. In the triple helix, all residues in the Xaa and the Yaa positions are exposed to the molecular surface. Each polypeptide forms a polyproline-II like left-handed helix. Three chains are staggered by one residue, and they form a right-handed superhelix (Fig. 1). The carboxyl group of the collagen helix is almost perpendicular to the molecular axis. This orientation is different from the alpha-helix, in which the carboxyl groups are almost parallel to the helical axis.

In the analysis of crystals of collagen model peptides, the carboxyl oxygen of the Yaa position residue has two hydrogen bonds with water molecules and that of Gly has one, because the other site for a hydrogen bond is hindered stericly by the neighboring peptide chain. In addition, the carboxyl oxygen in the Xaa position of Pro is directed to the center of the triple helix and participates in a direct hydrogen bond with the NH of Gly in the neighboring chain. This interchain Gly(NH\_\_O)Xaa hydrogen bond is almost perpendicular to the molecular axis. Based on the diffraction data of crystals, the left-handed 7/2-helical structure (Fig. 1) with a 20 Å axial repeat is obtained when both the Xaa and the Yaa positions are occupied by imino acids (2). Hence, left-handed 7/2-helical symmetry means the seven Gly-Xaa-Yaa tripeptide units from three chains make two left-handed turns in an axial repeat.

Each peptide strand forms a right-handed 7/1-helix in which seven tripeptide units and one helical unit turn in a 60 Å axial repeat. Therefore, the tripeptide unit twist and the tripeptide unit height of the strand are 51.4 (\(\approx 360/7\)) and 8.57 Å (\(\approx 60/7\)), respectively. The type I collagen molecule has a major triple helical domain with 1,014 amino acid residues of repeated \(-\text{Gly-Xaa-Yaa}-\) sequence. The pitch of each residue is 2.85 Å along the molecular axis. Therefore, the length of type I collagen molecule is about 300 nm. In collagen fibrils or the crystals of collagen model peptides, the distance between each triple helix is almost 1.5 nm.

The triple helical domain of collagen has a unique amino acid composition. Glycine accounts for one third of the total amino acid content. In homothermal animals, proline and 4(R)-hydroxyproline (Hyp) accounts for about 10% each. These three residues comprise more than half of the amino acids in collagen. Glycine and proline are usually regarded as two exceptional residues in proteins. Glycine is the only non-chiral amino acid, and it has only a hydrogen atom as side chain; usually, the composition is less than the other amino acid residues in globular proteins. Because of the smallest side chains, the dihedral angles of glycine are the most conformationally flexible among the 20 amino acids. In contrast, proline is the least conformationally flexible because of the five-membered pyrrolidine ring structure that restricts the phi angle. Proline is the only imino acid among the 20 coded amino acids. The high content of Gly and Pro (4(R)Hyp give the collagen triple helix unique properties not found in globular proteins. The content of hydrophobic residues in the triple helix is much lower than in globular proteins. In the collagen triple helix, more positively charged residues exist (Lys + Arg) than negatively charged residues (Glu + Asp). Because of the highly extended rod-like structure of the collagen triple helix, the accessible surface area per residue is much larger than that of globular proteins. All residues in the Xaa and the Yaa position are exposed to the molecular surface in a radial pattern.

Proline ring puckering

The pyrrolidine ring of (hydroxy)proline is not planar. The analyses of model collagen peptides at high resolution have revealed that there is distinct propensity of ring puckering in the Xaa and the Yaa position. In most cases, the imino acid residues in the Xaa position have the C\(_\text{y}-\text{endo}\) (down) puckering (Fig. 2). In contrast, most of imino acid residues in the Yaa position have the C\(_\text{y}-\text{exo}\) (up) puckering conformation.
Sometimes hydroxylysine is modified even more to O-β-Lys in the Yaa position is modified to 5-hydroxylysine. About a quarter of residues in the Yaa position; however, prolyl 3-hydroxylation avians, prolyl 4-hydroxylation occurs in almost all Pro 5-dioxygenases LH 1, 2a/2b & 3: EC 1.14.11.4) in some Gly-Pro-4Hyp-Gly- sequences, and 5-hydroxylation of lysine to 5-hydroxylysine (5-lysylhydroxylases, ly- syl hydroxylases (LHs), procollagen-lysine 2-oxoglutarate dioxygenases LH 1, 2a/2b & 3: EC 1.14.11.4) in some Gly-Xaa-Lys-Gly- sequences. In almost all mammals and avians, prolyl 4-hydroxylation occurs in almost all Pro residues in the Yaa position; however, prolyl 3-hydroxylation is much less frequent. Each type I collagen polypeptide chain has one 3-hydroxyproline. About a quarter of Lys in the Yaa position is modified to 5-hydroxylysine. Sometimes hydroxylysine is modified even more to O-β-galactosylhydroxylysine (enzyme: galactosyltransferase, EC 2.4.1.50) and 2-O-D-galactosyl-O-β-D-galactosylhydroxylysine (enzyme: galactosylhydroxylysyl glucosyltransferase, EC 2.4.1.66). This type of glycosylation has been found only in the collagen triple helical domain. LHS also possesses relatively low levels of collagen glucosyltransferase activity and very low levels of collagen galactosyltransferase activity in addition to the lysyl hydroxylase activity. The posttranslational modifications of invertebrate collagens are more complex than in vertebrates. The cuticle collagen of the sea hydrothermal tube worm contains glycosylated threo- proline and some other amino acid residues on the stability of the triple helix. Regarding the stabilizing effect of 4(R)-hydroxyproline on the triple helix, studies with model peptides showed that the or- der of the stability of the triple helix is -Gly-Pro-4(R)Flp- > -Gly-Pro-3(R)Hyp- > -Gly-Pro-4(R)Hyp-Gly-Pro-3(R)Hyp-Gly-Pro-NH2 and acetyl-(Gly-Pro-4(R)Hyp-Gly-Pro-3(R)Hyp-Gly-Pro-Yaa-(Gly-Pro-4(R)Hyp)3-Gly-Pro-4(R)Hyp-(Gly-Pro-4(R)Hyp)3-Gly-Pro-Yaa-(Gly-Pro-4(R)Hyp)3-(Gly-Pro-4(R)Hyp)3-Gly-Pro-NH2 (8).

Some residues, such as Gly and aromatic residues significantly destabilize the triple helix. Galactosylations of threonine (9) and that of the lysine (10) increases the stability of the triple helix. The thermal stability of collagen from different animal species is correlated with their highest environmental temperature. In general, the thermal stability of collagen is increased with the content of imino acid, proline, and 4(R)-hydroxyproline. Es- pecially, the contribution of 4(R)-hydroxyproline in the Yaa position is important. An additional stabilization of the collagen triple helix occurs on the formation of fibrils. The melting temperature of the collagen triple helix in fibrils is around 55°C. To clarify the mechanism of the stability of the triple helix, synthetic collagen model peptides have been used. The effect of 20 types of amino acid residues on the stability of the triple helix was analyzed in the context of host-guest system, that is, acetyl-(Gly-Pro-4(R)Hyp)3-Gly-Pro-4(R)Hyp-Gly-Pro-NH2 and acetyl-(Gly-Pro-4(R)Hyp)-Gly-Pro-Yaa-(Gly-Pro-4(R)Hyp)3-Gly-Pro-NH2 (8).

To determine the effect of 4(R)-hydroxyproline on the triple helix, studies with model peptides showed that the or- der of the stability of the triple helix is -Gly-Pro-4(R)Flp- > -Gly-Pro-4(R)Hyp-.

Stability of the Collagen Triple Helix

The temperature where one half of the polypeptide exists as a triple helix is termed the melting temperature (Tm), and is specific for different collagens from different species. The Tm of soluble collagen at neutral pH physiologic salt solution is around the body temperature of the organism from which it is isolated. The thermal unfolding of interstitial fibrillar collagens, type I, II, and III, occurs in a very narrow temperature interval of less than 3 degrees in passing from a 90% native to a 90% denatured conformation. Also, the guanidinium chloride induced transition of collagen is very sharp and occurs after increasing the concentration of the denaturant by only 0.2 M. The denaturation of these types of collagen is highly coopera- tive. The enthalpy change of the triple helix ↔ coil transition for type I collagen was determined to be ΔH = 15-18 kJ/mole triple peptide unit. The enthalpy change per residue of triple helix is significantly greater than that of globular proteins. The source of this greater enthalpy change is still controversial. Another unique thermodynamic property of collagen molecules is the absence of a change in heat capacity after the denaturation. The stability of the collagen triple helix depends on the se- quence and the length of the three polypeptide chains. The thermal stability of collagen from different animal species is correlated with their highest environmental temperature. In gen- eral, the thermal stability of collagen is increased with the content of imino acid, proline, and 4(R)-hydroxyproline. Es- pecially, the contribution of 4(R)-hydroxyproline in the Yaa position is important. A additional stabilization of the collagen triple helix occurs on the formation of fibrils. The melting temperature of the collagen triple helix in fibrils is around 55°C. To clarify the mechanism of the stability of the triple helix, synthetic collagen model peptides have been used. The effect of 20 types of amino acid residues on the stability of the triple helix was analyzed in the context of host-guest system, that is, acetyl-(Gly-Pro-4(R)Hyp)3-Gly-Pro-4(R)Hyp-Gly-Pro-NH2 and acetyl-(Gly-Pro-4(R)Hyp)-Gly-Pro-Yaa-(Gly-Pro-4(R)Hyp)3-Gly-Pro-NH2 (8).

Some residues, such as Gly and aromatic residues significantly destabilize the triple helix. Galactosylations of threonine (9) and that of the lysine (10) increases the stability of the triple helix. To determine the effect of 4(R)-hydroxyproline on the triple helix, studies with model peptides showed that the or- der of the stability of the triple helix is -Gly-Pro-4(R)Flp- > -Gly-Pro-4(R)Hyp-.
-Gly-Pro-4(R)Hyp- > Gly-Pro-Pro-, where Flp is fluoroproline. Fluorine is the most electronegative atom, and organic fluorine forms only weak hydrogen bonds. Neither -Gly-4(R)Hyp-Pro nor -Gly-Pro-4(S)Hyp- form a triple helix. These properties can be explained by a stereoelectronic effect of the 4-substitution (11). The inductive effect, the gauche effect of the pyrrolidine ring (4-substitution and the amide group), and an n → π* interaction. These effects adjust the optimum phi and psi angles of both the Xaa and the Yaa positions of collagen peptides and the puckering of the pyrrolidine ring.

Solvent affects the stability of collagen molecules and the triple helix of collagen model peptides. Polyols, such as glycerol, 1,2- and 1,3-propanediol, polyethylene glycol, sugars, and glycosaminoglycans increase the melting temperature of the triple helix. Trimethylamine N-oxide (TMAO) also stabilizes the triple helix of collagen model peptides. Urea and guanidinium hydrochloride decrease the Tm of collagen concentration linearly with increasing concentrations.

The stability of the triple helix is not affected as much by pH as globular proteins. There is a slight destabilization of the triple helix at acid pH. The effect of salt ions on the stability is similar to the Hofmeister series (12).

Chemical Tools and Techniques for Collagen Research

Collagen is one of the most popular proteins for biophysical and biochemical analysis because of its unique shape, ease of acquisition, and stability. The structure of the collagen triple helix seems more homogeneous and simpler than that of globular proteins. However, the difficulties of collagen preparations for biochemical and biophysical analysis are the heterogeneity of the sample by posttranslational modifications, intermolecular covalent cross-linking, the large molecular weight, the aggregation properties at physiologic temperatures, and the hysteresis of the folding-refolding reaction. Almost all biochemical and biophysical techniques have been applied to collagen such as circular dichroism (CD), analytical ultracentrifugation, differential scanning calorimetry (DSC), isothermal titration calorimetry, transmission electron microscopy, scanning electron microscopy, atomic force microscopy, scanning tunneling microscopy, second harmonic generation microscopy, laser scattering, electric birefringence, X-ray analysis, IR and NMR.

The collagen triple helix has a unique CD spectrum. The spectrum shows a positive peak at 220–225 nm, and a negative peak at 195–200 nm. In contrast, the polyproline II-like poly-4-hydroxyproline helix has a positive peak at 228 nm, and a negative peak at 206 nm. The polyproline II-like poly-4-hydroxyproline helix has a positive peak at 219 nm and a negative peak at 205 nm (13). The thermal stability of the triple helix can be monitored easily by the CD
The CD signal monitored as a function of temperature shows a sigmoidal denaturation curve. NMR is used for the characterization of the triple helical structure and also for the folding kinetics of the triple helix. NMR studies with synthetic model peptides of the triple helix are difficult because of overlapping resonances of the repeated sequence and by peak broadening from the shape. Isotopic labeling is used to observe specific residues using heteronuclear NMR techniques. Hydrogen exchange studies are used to show the Gly amide exchange. The exchange is faster in the imino acid-poor regions of a synthetic peptide compared with the Gly-Pro-4(R)Hyp region.

DSC was used to characterize the thermodynamic properties of the collagen triple helix to coil transition in individual molecules and in fibrils. Many microscopy techniques were used to visualize individual collagen molecules and their supramolecular assemblies.

Collagen model peptide synthesis

Because the natural collagen molecules are large, heterogeneous, and difficult to purify, collagen model peptides are used for many analyses. Recent developments in solid-phase peptide synthesis techniques have made the synthesis of longer peptides easier than before. The commercial availability of Fmoc-derivatives of 4(R)-hydroxyproline, 4(R)-fluoroproline and 3(S)-hydroxyproline, the synthesis of glycosylated-Fmoc-threonine and glycosylated-Fmoc-hydroxylysine, and other unusual amino acids have extended the range of collagen-like peptides that have been synthesized. Overlapping cysteine linked peptides of type III collagen have been made, which covers the whole triple helical region of type III collagen (1029 amino acids) (14). Several cross-links were introduced to stabilize the three chains of the triple helix (see below).
Collagen Triple Helix, Stability of

1. 2. 3-propane tricarboxylic acid (PTC)

Trimerized triple helical model design

In vivo, collagen molecules have domains that initiate trimerization of the synthesized alpha chains. Usually, globular domains at the carboxyl terminal end of the molecule have this role. Once this domain is removed, refolding of the three collagen polypeptides to the correctly staggered molecules is very difficult or impossible. The folding reaction from three polypeptides into a triple helix is affected by the concentration of the peptides. For the kinetic refolding analysis of triple helical model
peptides, several techniques have been developed to link the three polypeptides (Fig. 5). These are

a. The type III collagen homotrimeric disulfide knot (-Gly-Pro-Cys-Cys- sequence)
b. Regioselective artificial cystine knot for heterotrimers
   (15)
c. The foldon domain of T4 phage fibrin (16)
d. A di-lysine scaffold (17)
e. cis,cis-1,3,5-trimethylcyclohexane-1,3,5-tricarboxylic acid (K emt acid) (18)
f. TREN (Tris(2-aminoethyl)amine) succinic acid (19)
g. 1,2,3-propanecarboxylic acid (20)
h. A coordination of Fe(II) ions to built-in bipyrindane ligands (21)
i. Monosubstituted chains (22)
j. 18-membered cyclic hydropyran oligolide

References


Further Reading


See Also

Enzyme Kinetics

W. Wallace Cleland, University of Wisconsin, Madison, Wisconsin
doi: 10.1002/9780470048672.wecb159

Enzyme kinetics is an important tool for assaying enzyme activities and for determining enzyme mechanisms. Although other techniques can provide useful information on enzyme mechanisms, the kinetics has to be the ultimate arbitrer because it looks at the reaction while it is taking place. Initial velocity patterns, inhibition patterns, patterns of isotopic exchange, pH profiles, and isotope effects are all kinetic tools that allow one to determine kinetic mechanisms, chemical mechanisms, and transition state structures.

All chemical reactions in living cells are catalyzed by enzymes, because it is the only way to control reaction rates. Kinetics is one of the best ways to study enzyme-catalyzed reactions, because one follows the reaction while it takes place. Structural studies such as X-ray structures are extremely helpful in understanding enzymatic catalysis, but the kinetics must be the ultimate arbiter.

Reasons for Study of Enzyme Kinetics

One determines enzyme kinetics for several reasons. First, kinetic assays are needed to measure enzyme activity during purification and to establish that one has an active enzyme. Such assays also allow one to determine the effects of potential drugs as inhibitors. Second, kinetic studies are one of the best ways to determine the mechanism of the reaction, and we will focus in this article mainly on the use of kinetics for this purpose. For determination of mechanism we need:

1. The kinetic mechanism, which is the order of events during a catalytic cycle in terms of the order of substrates combining with and products dissociating from the enzyme. We need to know the complexes that form, how many sites exist, and their specificity for binding reactants. This information is qualitative.
2. The relative rates of steps in the mechanism, which is quantitative information on which steps are rate limiting at high or low concentrations of substrates. Are substrates sticky (that is, react once they combine faster than they dissociate)?
3. The chemical mechanism, which includes the nature of any intermediates and the identification of enzymatic groups involved in binding and catalysis.
4. Transition state structure. Isotope effects will often give a very clear picture of what the transition state looks like for an enzymatic reaction.

One Substrate Concentration Varied

An enzyme has a pocket or cleft in which substrates bind, and usually a flap exists that then closes around the substrates once they are bound. Thus, an enzyme has open and closed conformations, with reactants coming and going from the open form and catalysis taking place in the closed form. As the enzyme and substrate(s) form a complex, the basic rate equation when one substrate concentration is varied (any others being held constant) usually is:

\[ v = \frac{V A}{K + A} \]  

where \( v \) is velocity, \( V \) is the maximum velocity when the substrate concentration is saturating, \( K \) is the Michaelis constant, and \( A \) is the substrate concentration (bold type will indicate concentrations in this article). The Michaelis constant measures affinity in the steady state and is the level of substrate that gives half of \( V \). \( K \) can be greater than, smaller than, or the same as the dissociation constant of the substrate. We will assume that the enzyme is operating in the steady state, that is, that the substrate concentrations exceed the enzyme level by at least an order of magnitude, which is normally the situation in an enzyme assay, but for very slow mutants it may not be a valid assumption. Steady-state studies with enzyme levels equivalent to or exceeding substrate concentrations also require the use of more complicated rate equations.

The usual practice is to measure the initial velocity with which the reaction starts and to determine such initial velocities for separate reaction mixtures containing different levels of substrate. Although it is possible to integrate Equation 1 and determine residual substrate concentrations or product formation as the reaction proceeds, this procedure is subject to more uncertainties as the pH may change or the enzyme may die. The high correlation between the various data points also makes statistical analysis tricky. See References 1 and 2 for a discussion of this method and its problems.
In Equation 1, the rate when \( A \) is very high is \( V \), whereas at low \( A \), it is given by:

\[
v = \frac{V}{K}\frac{1}{A}
\]  

(2)

where \( V/K \) is an apparent first-order rate constant. \( V \) and \( V/K \) are the two kinetic constants that vary independently with the levels of other substrates, inhibitors and activators, \( pH \), ionic strength, and so forth. \( K \) is not an independent constant, but just the ratio of \( V \) and \( V/K \), it does indicate the likely concentration of the substrate \( A \) at which only in this concentration range does one get proportional control of the rate. At substrate levels well above \( K_a \), the rate does not vary with substrate concentration, whereas at low substrate concentrations, one is not using the potential activity of the enzyme.

As Equation 1 is not linear, but a rectangular hyperbola, it is usually inverted for graphing purposes to give a double reciprocal plot:

\[
\frac{1}{v} = \frac{1}{V} + \frac{1}{V/K}\frac{1}{A}
\]

(3)

Thus, \( 1/v \) plots linearly against \( 1/A \), with a slope of \( V/K \) and a vertical intercept of \( V \). Note that the slope and intercept are the reciprocals of the two fundamental kinetic constants, \( V/K \) and \( V \). In the discussion that follows, we will refer to slope and intercept effects, and one needs to keep in mind that these are effects on \( V/K \) and \( V \).

Although Equation 3 is the usual form used for graphing, one should not make unwarranted least squares fits of the experimental data to it. Such fits require \( v^0 \) weighting factors if errors in the velocities are thought to be roughly constant or \( v^2 \) weights if errors are proportional to velocities. The proper fits are made to Equation 1 directly by nonlinear least squares fitting (for constant errors), or to the equation with logarithms taken of \( v \) and \( v/K \) if errors are proportional to velocities, for graphing purposes, Equation 4 is inverted to give:

\[
\frac{1}{v} = \frac{1}{V}(1 + \frac{K_a}{V}) + \frac{1}{V/K}\frac{1}{A}
\]

(4)

Equation 6 describes an intersecting initial velocity pattern where \( 1/v \) is plotted versus \( 1/A \) at different values of \( B \), whereas Equation 6 describes the pattern where \( 1/v \) is plotted versus \( 1/B \) at different values of \( A \). Both the slopes and the intercepts of the reciprocal plots are functions of the other substrate concentration, and replots of slopes or intercepts versus the reciprocal of the other substrate concentration allow determination of all kinetic constants.

The reciprocal plots in these patterns cross to the left of the vertical axis at:

\[
\frac{1}{v_{cross}} = \frac{1}{V}(1 - \frac{K_a}{V})
\]

(7)

The crossover point is above the horizontal axis if \( K_a < K_{in} \) and below it if \( K_a > K_{in} \), and is the same regardless of which substrate concentration is varied.

Sequential mechanisms are typical of kinases and dehydrogenases. Some mechanisms are ordered and some are random, and one get proportional control of the rate. The approach outlined above is sufficient when there is only one substrate or in an assay where one varies only the concentration of one substrate. However, where two or more substrates exist, and one wants to know the order of their combination with the enzyme, one needs to determine an initial velocity pattern. One varies the concentrations of one substrate at several different levels of a second substrate and determines the initial velocities for the different reaction mixtures. It is, of course, necessary to have the same enzyme level in each reaction or correct the rates to constant enzyme concentration.

Three initial velocity patterns are commonly observed.

**Use of Kinetics for Study of Mechanism**

**Initial Velocity Studies**

The approach outlined above is sufficient when there is only one substrate or in an assay where one varies only the concentration of one substrate. However, where two or more substrates exist, and one wants to know the order of their combination with the enzyme, one needs to determine an initial velocity pattern. One varies the concentrations of one substrate at several different levels of a second substrate and determines the initial velocities for the different reaction mixtures. It is, of course, necessary to have the same enzyme level in each reaction or correct the rates to constant enzyme concentration.

Three initial velocity patterns are commonly observed. When both substrates have to add to the enzyme before any products are released, one has a sequential mechanism. The rate equation for most sequential mechanisms is:

\[
v = \frac{V}{V/K}(A + B - AB)
\]

(8)

where \( V \) is initial velocity; \( V \) is maximum velocity; \( A \) and \( B \) are substrate concentrations; \( K_a \) and \( K_b \) are Michaelis constants for \( A \) and \( B \), and are the levels that give half of \( V \) when the other substrate is saturating; and \( K_{in} \) is the dissociation constant of \( A \), the first substrate to add to the enzyme if the mechanism is ordered. If the mechanism is random, a dissociation constant for \( B \) of \( K_{ib} \) also exists, and \( K_aK_b = K_{ib}K_{in} \).

Although experimental data should be fitted directly to Equation 1, or to the log form of it if errors in velocities are proportional to velocities, for graphing purposes, Equation 4 is inverted to give:

\[
\frac{1}{v} = \frac{1}{V}(1 + \frac{K_a}{V}) + \frac{1}{V/K}\frac{1}{A}
\]

(9)

Equation 6 describes an intersecting initial velocity pattern where \( 1/v \) is plotted versus \( 1/A \) at different values of \( B \), whereas Equation 6 describes the pattern where \( 1/v \) is plotted versus \( 1/B \) at different values of \( A \). Both the slopes and the intercepts of the reciprocal plots are functions of the other substrate concentration, and replots of slopes or intercepts versus the reciprocal of the other substrate concentration allow determination of all kinetic constants.

The reciprocal plots in these patterns cross to the left of the vertical axis at:

\[
\frac{1}{v_{cross}} = \frac{1}{V}(1 - \frac{K_a}{V})
\]

(10)

The crossover point is above the horizontal axis if \( K_a < K_{in} \) and below it if \( K_a > K_{in} \), and is the same regardless of which substrate concentration is varied.

Sequential mechanisms are typical of kinases and dehydrogenases. Some mechanisms are ordered and some are random,
but the initial velocity pattern does not distinguish between them and inhibition studies are required (see below).

When one substrate adds to the enzyme and a product is released before the second substrate adds, one has a ping-pong mechanism, and the rate equation is

\[ v = V_{AB}/(K_{A}A + K_{B}B + AB) \]  

(8)

The constant term is missing, which leads to a parallel initial velocity pattern (Fig. 2) regardless of which substrate concentration is varied:

\[ 1/v = (K_{A}V)(1/A) + (1/V)(1 + K_{B}/B) \]  

(9)

\[ = (K_{A}V)(1/B) + (1/V)(1 + K_{A}/A) \]  

(10)

The slopes of the reciprocal plots do not vary with the level of the other substrate, but the intercepts do. Replots of intercepts, and comparison with slopes, allow determination of the kinetic constants.

Ping-pong mechanisms are shown by transaminases and nucleoside diphosphate kinase (4). One must be cautious about assuming that a parallel-looking initial velocity pattern represents a ping-pong mechanism, as some sequential mechanisms give patterns in which the crossover point is far down in the third quadrant and the pattern looks parallel. Isotopic exchange studies will distinguish the mechanisms (see below).

The third type of initial velocity pattern results from a mechanism in which 1) the substrates add in obligatory order and 2) the off-rate constant for the first substrate to bind exceeds the turnover number (V/Et or kcat) sufficiently that its binding is at equilibrium, which is called an equilibrium ordered initial velocity pattern (Fig. 3). The rate equation is

\[ v = V_{AB}/(K_{A}K_{B} + K_{A}A + AB) \]  

(11)

and in reciprocal form

\[ 1/v = (K_{A}K_{B}/V)(1/A) + (1/V)(1 + K_{B}/B) \]  

(12)

\[ = (K_{A}/V)(1 + K_{A}/A)(1/B) + (1/V) \]  

(13)

The patterns are not the same when plotted versus 1/A or 1/B. Equation 13 plots as a pattern intersecting to the left of the vertical axis at 1/V, with the slope at infinite B being zero. That is, a replot of slopes versus 1/B goes through the origin. Equation 13 plots as a pattern intersecting on the vertical axis at 1/V. Note that the differences in the patterns given by Equations 13 and 13 establish the order of combination of the two substrates as well as showing that they add in an obligatory order.

Equilibrium ordered mechanisms are rare, because in most ordered sequential mechanisms the off-rate constant of the first substrate to add is not much larger than the turnover number. Examples of equilibrium ordered mechanisms are creatine kinase at pH 7 or below (5), malic enzyme at pH 4 (6), phosphofructokinase with fructose-6-phosphate as an alternate substrate (7), and glycerokinase with amino analogs of glycerol as substrates (8). More common are cases in which reactant A is an activator that cannot dissociate once the substrate adds. In this case, the off-rate constant of the activator need not be large, as the activator is not used up during the reaction and its binding will come to equilibrium in the steady state. An example is the addition of Mn2+ followed by UDP-galactose to lactose synthetase (9).

When three substrates exist for an enzymatic reaction, the usual procedure is to vary the concentration of one of them at fixed levels of a second, with the concentration of the third held constant. This process gives three initial velocity patterns (A–B, A–C, and B–C). The pyruvate dehydrogenase complex shows three parallel patterns as the mechanism is a three-site ping-pong mechanism (10). A number of enzymes show intersecting patterns between two substrates and parallel patterns with either one of these and the third substrate. Examples are synthetases where MgATP and an acid combine to generate an acyl-adenylate intermediate with liberation of Mg-pyrophosphate. The intermediate then reacts with the third substrate to give the final product and A-MP. Biotin carboxylases show similar patterns with MgATP and bicarbonate generating carboxy-l-phosphate that carboxylates biotin with the release of Mg°D°P and phosphate. Carboxy-biotin then carboxylates the acceptor.
Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.

Inhibition Studies

The determination of kinetic mechanisms requires more than just initial velocity patterns, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.

Inhibition Studies

The determination of kinetic mechanisms requires more than just initial velocity patterns, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.

Inhibition Studies

The determination of kinetic mechanisms requires more than just initial velocity patterns, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.

Inhibition Studies

The determination of kinetic mechanisms requires more than just initial velocity patterns, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.

Inhibition Studies

The determination of kinetic mechanisms requires more than just initial velocity patterns, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Three common types of inhibition exist. The first causes changes only in V, and inhibition studies are usually required. Several types of inhibitors are useful. The products are substrates in the reverse reaction and thus have some affinity for the enzyme and will give inhibition unless their inhibition constants exceed their solubility. Dead-end inhibitors are molecules that play musical chairs with the substrates for open portions of the active site but do not react. Substrates may act as dead-end inhibitors by combining at points in the mechanism where they are not intended and thus cause substrate inhibition. The inhibition patterns caused by these inhibitors are useful in distinguishing between different kinetic mechanisms.

Sequential terreactant mechanisms can be ordered, random, or partly random. A fully random mechanism gives only intersecting patterns. An ordered mechanism gives a parallel pattern when the first and third substrate concentrations are varied and the level of the second one is saturating. An example is glutamate dehydrogenase, with NADPH, α-ketoglutarate, and ammonia adding in that order (11). The same pattern is seen if the first two substrates add in obligate order, but the third can add randomly. An example is β-hydroxy-β-methylglutaryl-CoA reductase (12). If one substrate must add first, but the other two can add randomly to the EA complex, the A–B pattern becomes parallel at infinite C and the A–C pattern becomes parallel at infinite B. For a detailed description of the kinetics of reactions with three substrates, see Reference 13.
In the reverse reaction, free ADP gives competitive substrate inhibition versus MgADP. MgATP and MgADP do not show substrate inhibition as they are too large to fit into the site for AMP or ADP.

Isotopic Exchange

Initial velocity and inhibition patterns involve following the chemical reaction catalyzed by an enzyme. In isotopic exchange studies, one follows the transfer of label between reactants at equilibrium, or from product back into substrate while the reaction is proceeding. In ping-pong mechanisms, exchange occurs without all reactants being present. Thus, one can observe an exchange between A and P in the absence of B and Q and vice versa. The reaction comes rapidly to equilibrium as indicated by E + A = F + P, where A and P are the reactants between which exchange takes place and E and F are the two stable enzyme forms. Observation of such exchanges establishes a ping-pong mechanism as long as Equation 21 holds.

\[ \frac{1}{V_{A+}} + \frac{1}{V_{B-}} = \frac{1}{V_{A}} + \frac{1}{V_{B}} \]  

(21)

\( V^{*}_{AP} \) and \( V^{*}_{BQ} \) are the maximum exchange rates given by Equations 24 and 24, whereas \( V_{1} \) and \( V_{2} \) are the maximum velocities of the chemical reaction in the two directions as given by equations such as Equation B.

\[ V_{1} = V_{B} + AP \left( K_{b}A + K_{b}P + AP \right) \]  

(22)

\[ V_{2} = V_{B} + BQ \left( K_{b}Q + K_{b}B + BQ \right) \]  

(23)

The forms of Equations 24 and 24 resemble that of Equation B, and reciprocal plots of exchange rate versus the level of one reactant at fixed levels of the other will give a parallel pattern. The kinetic constants determined are dissociation constants, rather than Michaelis constants, as the exchange takes place at equilibrium. One can vary the levels of reactants freely, because equilibrium is maintained by the ratio of E and F.

In more complicated ping-pong mechanisms where two substrates combine before release of one or two products, the exchange pattern between a substrate and product when A and B levels are varied will tell whether the addition of A and B is ordered or random. For example, with a synthetase where MgATP reacts with an acid to give Mg-pyrophosphate and an acyl-adenylate, one can observe exchange between pyrophosphate and ATP at different levels of MgATP and acid at a fixed level of Mg-pyrophosphate. If MgATP levels are varied at fixed levels of acid, one has three cases:

1. The mechanism is ordered with MgATP as A and the acid as B. One sees an intersecting pattern with competitive substrate inhibition by B.
2. The mechanism is ordered with the acid as A and MgATP as B. One sees an equilibrium ordered pattern (Eq. 13).
3. The mechanism is random. One sees an intersecting pattern with no substrate inhibition.

With sequential mechanisms, one can follow isotopic exchange at equilibrium only if one varies the levels of one substrate and one product at a constant ratio and has all reactants present. The usual protocol is to vary the concentrations of like reactants together at fixed levels of the others and determine exchange rates. For example, with an alcohol dehydrogenase, one varies the levels of NAD and NADH together and alcohol and aldehyde together. One determines exchange between NAD and NADH, and between alcohol and aldehyde. In an ordered mechanism, reciprocal plots of exchange rate versus reciprocal levels of reactants are linear when A and Q levels are varied together, and the plots for the A–Q and B–P exchanges have the same apparent dissociation constant (the plots cross on the horizontal axis). The B–P exchange is faster. When the levels of B and Q are varied together, the reciprocal plots for A–Q exchange and B–P exchange are parallel, but the A–Q exchange shows total substrate inhibition (Eq. 20), which is a very sensitive test for an ordered mechanism, as partly random ones will not show complete substrate inhibition, and the apparent dissociation constants when A and Q levels are varied together will not show the same dissociation constants for the two exchanges.

Random mechanisms will not show substrate inhibition of exchanges unless the levels of reactants that can form an abortive complex are varied together. The relative rates of the two exchanges will show whether catalysis is totally rate limiting (a rapid equilibrium random mechanism), or whether release of a reactant is slower. For kinases that phosphorylate sugars, the usual pattern is for sugar release to be partly rate limiting, but for nucleotides to dissociate rapidly (15, 16).

It is also possible in a sequential mechanism to follow the exchange of label from a product back into substrate while the reaction is occurring. Only the first product to be released will exchange back into substrate, which allows determination of the order of product release (17).

pH Studies

The pH variation of kinetic parameters allows one to identify enzyme groups required for binding or catalysis. The parameters...
that are plotted versus pH are log(V/K), log(V), and pK, [log (K/Ki)], where Ki is the dissociation constant of substrates, other than the last one to add to the enzyme, or of competitive inhibitors. Any pK values seen in a pK profile have their correct values for the enzyme form involved or the molecule that is binding, as K, is an equilibrium dissociation constant.

The log(V/K) profile shows the pKs of groups in substrate or enzyme required for binding, as well as ones involved in catalysis (Fig. 5). Thus, if a substrate binds only as a dianion, the V/K profile will decrease a factor of 10 per pH units below the pK of the substrate. If a group on the enzyme has a required protonation state for catalysis, this pK will also appear. The pK's seen in V/K profiles may not appear at their true values if the substrate is sticky (that is, reacts to give products as fast or faster than it dissociates). The pK will be displaced outward on the profile by log (1 - S), where S is the stickiness ratio (the ratio of the net rate constant to produce the first product and the off-rate constant for the substrate). Comparison of pKs in V/K and pK profiles allows one to determine stickiness of the substrate.

The log V profile shows only the pK's of groups responsible for the catalytic reaction after the substrate has bound and the pKs are typically displaced from values seen in free enzyme or substrate. The pKs of neutral acids (carboxyl, sulfhydryl, phosphate) are elevated by a pH unit or more by being removed from solvent. Specific hydrogen bonding interactions can displace a pK either upwards or downwards by 1-2 pH units. For fumarase, for example, the pK of the group probably a carboxyl) that removes the proton during reaction of malate is elevated by binding of either malate or fumarate, whereas the pK of the group that protonates the OH of malate to give water is elevated by 2 pH units by the binding of malate, and lowered by 2 units by the binding of fumarate (18).

Another reason for displaced pK values in log V profiles is the switch in rate-limiting steps. If the catalytic step that is pH-sensitive is 10-fold faster than pH-independent second product release at neutral pH, the pKs in the log V profile will be displaced outward by one pH unit, as catalysis has to be slowed by an order of magnitude before it becomes rate limiting.

pH profiles are log-log plots and consist of horizontal lines when all groups are properly protonated and lines with slopes of 1, 2, -1, -2, and so on. These segments are connected with a curved portion 2 pH units wide that is centered on the pK. If protonation causes loss of binding or catalysis, the profile decreases a factor of 10 per pH unit below the pK. If deprotonation does the same, the profile drops at high pH. When one group must be protonated and the other unprotonated, the profile will drop at both low and high pH. In this case, when the pKs are not far apart, one cannot tell from the pKs which group has to be protonated and which has to be unprotonated for activity. When the group with the lower pK has to be protonated and the one with the higher pK unprotonated, one has reverse protonation. The profile will have the same shape as if the other protonation states were required, but the amplitude will differ, which is evident in the V/K profiles for fumarase (18). As the equilibrium constant is pH-independent in the neutral pH range, the V/Kfumarate and V/Kmalate profiles are identical, with apparent pKs of 5.7 and 7.1. For reaction of maleate, the low pK group must be ionized and the other one protonated, whereas for reaction of fumarate, the states of protonation have to be reversed. This reversal is the typical situation when acid-base catalysis is involved in enzymatic reactions.

Isotope Effects

Isotope effects result from replacing the normal isotopes of hydrogen, carbon, nitrogen, or oxygen with deuterium or tritium, 13C, 14C, 15N, or 18O. The effects of such replacements are expressed as klight/kheavy for kinetic isotope effects, or Keq light/Keq heavy for equilibrium isotope effects. Equilibrium isotope effects compare the stiffness of bonding of the isotopic atom in product versus substrate, whereas kinetic ones compare the transition state to the substrate. The heavy isotope enriches in the more stiffly bonded position and the light isotope effect in the less stiffly bonded position.

Equilibrium isotope effects are normal (<1.0) if the substrate is more stiffly bonded and inverse (>1.0) if the product is more stiffly bonded. Primary kinetic isotope effects are almost always normal, as the atom being transferred is more weakly bonded in the transition state. Isotope effects in positions not subject to bond breaking or making are secondary and are good tools for determining transition state structure. They may be either normal or inverse.

Isotope effects on enzymatic mechanisms are determined for each V/K value and for V. A leading superscript indicates the nature of the isotope effect. Thus (V/K) is a deuterium isotope effect on V/K (the value for unlabeled substrate divided by that for deuterated substrate). The leading superscripts used by
Enzyme Kinetics

Isotope effects are useful in determining chemical mechanisms. In an oxidative decarboxylation where a hydride ion is removed from one carbon and CO₂ is lost from another, one can tell whether the reaction is concerted or stepwise by the effect of deuteration on the ¹³C isotope effect in the CO₂ (21). In a concerted mechanism, deuteration makes the chemical step more rate limiting and thus diminishes the same one if commitments were small to start with. In a stepwise mechanism, however, deuteration of the substrate makes the decarboxylation step less rate limiting and thus diminishes the size of the observed isotope effect.

If an appreciable equilibrium deuteron isotope effect exists on a stepwise mechanism (>1.1), one can tell which step comes first. In the direction where the deuteron-sensitive step comes first (21),

\[
\frac{1}{1.1}(V/K)_{D} - 1 = 0(V/K)/0K_{eq}
\]

and in the reverse direction with the ¹³C-sensitive step first,

\[
\frac{1}{1.1}(V/K)_{H} - 1 = 0V/K_{eq}(V/K)_{D} - 1.1K_{eq}
\]

These are really the same equation, but Equation 28 is expressed in terms of parameters of the forward reaction and Equation 29 in terms of ones for the reverse reaction.
The size of intrinsic isotope effects can be used to determine transition state structure. For example, in the concerted oxidative decarboxylation catalyzed by prephenate dehydrogenase, a large deuterium isotope effect (7.3) but a small 13C isotope one (1.5%) indicates an asynchronous transition state with considerable C-H bond cleavage, but only a small degree of C-C cleavage, as most decarboxylations show -- 5% 13C isotope effects (24). The shift in isotope effects as the redox potential of the nucleotide substrate was changed (thus changing $K_w$) for the formate dehydrogenase reaction showed that the transition state became earlier as $K_w$ increased (25). Similar results were seen with liver alcohol dehydrogenase (26).

References

Further Reading

See Also
Enzyme Inhibition, Mechanisms of Enzyme Inhibition, Tools to Study Enzyme Kinetics, Techniques to Study Kinetic Isotope Effects, Enzyme Catalysis, Chemistry of
Flavin-Mediated Hydroxylation Reactions

Barrie Entsch, School of Science and Technology, University of New England, Armidale, Australia

David P. Ballou, Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan
doi: 10.1002/9780470048672.wecb672

Flavins react with oxygen and can form stable flavin peroxides in an aprotic solvent or buried in a protein. It is this hydroperoxide or peroxide that is the oxygenating agent in flavoproteins. This property is used in nature to carry out aromatic hydroxylations, halogenations, Baeyer-Villiger oxygenations, hydroxylation of xenobiotics and some metabolites, as well as light emission from luciferase. Several groups of enzymes seem to have evolved hydroxylating properties independently of each other. One group consists of the two-component flavin-dependent hydroxylases that use many of the same principles as the single component hydroxylases, although they also have some special requirements. After a brief introduction to the reactivity of flavins with oxygen, we examine \( \text{p}-\text{hydroxybenzoate} \) hydroxylase as the paradigm for the chemistry and protein functions exhibited by these enzymes. We then discuss unique features of each group of enzymes and the exciting prospects for future research.

In the context of this section, hydroxylation refers to enzymatic catalysis in which one atom of molecular oxygen is incorporated into the structure of a substrate, which causes its oxidation. A large variety of enzymes can carry out this type of reaction, and they are called oxygenases or hydroxylases. They have one common property: They use cofactors in catalysis that are reactive with molecular oxygen. Flavoprotein hydroxylases are a major subgroup of oxygenases, and they are found in all types of aerobic organisms, although particularly in bacteria and fungi. A more detailed overview of flavoprotein hydroxylases can be found in References 1 and 2. These enzymes oxygenate various aromatic compounds and ketones, electron-rich atoms in many compounds, and halogen ions that then halogenate aromatic compounds. However, the enzymes do not hydroxylate less reactive compounds, such as hydrocarbons. We have chosen to emphasize in this section the enzyme \( \text{p}-\text{hydroxybenzoate} \) hydroxylase, because it has been the outstanding model for the chemistry of these enzymes (3, 4) and has been studied extensively. Most other enzymes can be understood as variations on this model.

Reaction of Flavins with Oxygen

The chemistry and biological context of flavins are described in the article, “Chemistry of Flavoenzymes.” Reduced forms of flavins are usually very reactive with \( \text{O}_2 \), which is somewhat surprising because most organic compounds are not very reactive with \( \text{O}_2 \). Although reactions between dioxygen and most organic molecules are extremely favorable in thermodynamic terms (consider what happens in wild fires), the reactions are usually slow. Reactivity is lacking because organic compounds have singlet spin states, whereas \( \text{O}_2 \) has a triplet spin state. However, in the case of reduced flavins and some other molecules, this impediment can be overcome because one-electron reduced states of flavins are readily accessible, and reactions with \( \text{O}_2 \) to form radicals is spin allowed (5). The accidental formation of various reactive oxygen species by electron transfers from reduced flavins to oxygen in cells is generally avoided because free flavins are usually present in very low concentrations. Hydroxylation can only occur if the formation and reaction of reactive oxygen species are tightly controlled. This control comes from the specific environments of enzyme active sites. One control strategy is to prevent the formation of the reactive reduced cofactor until conditions are suitable for hydroxylation—many flavoproteins use this strategy. Hydroxylases bind to and interact with flavins to decrease the thermodynamic stability of the free-radical form relative to the oxidized and reduced forms. As a result, when the reduced flavin reacts with oxygen inside the protein environment, the first electron is transferred to form a flavin radical and a superoxide radical (\( \text{O}_2^- \)) pair. This pair...
Flavin-Mediated Hydroxylation Reactions

PHBH catalyzes the reaction shown in Fig. 1 with a high degree of specificity and a turnover rate of 50–60 s⁻¹ at standard conditions of optimum pH (7.5–8) and 25 °C. A characteristic feature of PHBH and similar enzymes is the formation of some hydrogen peroxide during catalysis when conditions are not ideal for the enzyme (see Fig. 1, k8). This property, which is called uncoupling, comes about when the unstable flavin hydroperoxide (E F1 HpOH−S in Fig. 1) that is essential to catalysis decomposes. All of these flavin-dependent oxygenases function to limit the amount of toxic hydrogen peroxide formed during catalysis.

Catalytic cycle

PHBH catalyzes the reaction shown in Fig. 1 with a high degree of specificity and a turnover rate of 50–60 s⁻¹ at standard conditions of optimum pH (7.5–8) and 25 °C. A characteristic feature of PHBH and similar enzymes is the formation of some hydrogen peroxide during catalysis when conditions are not ideal for the enzyme (see Fig. 1, k8). This property, which is called uncoupling, comes about when the unstable flavin hydroperoxide (E F1 HpOH−S in Fig. 1) that is essential to catalysis decomposes. All of these flavin-dependent oxygenases function to limit the amount of toxic hydrogen peroxide formed during catalysis.

Catalytic cycle

PHBH catalyzes the reaction shown in Fig. 1 with a high degree of specificity and a turnover rate of 50–60 s⁻¹ at standard conditions of optimum pH (7.5–8) and 25 °C. A characteristic feature of PHBH and similar enzymes is the formation of some hydrogen peroxide during catalysis when conditions are not ideal for the enzyme (see Fig. 1, k8). This property, which is called uncoupling, comes about when the unstable flavin hydroperoxide (E F1 HpOH−S in Fig. 1) that is essential to catalysis decomposes. All of these flavin-dependent oxygenases function to limit the amount of toxic hydrogen peroxide formed during catalysis.
Flavin-Mediated Hydroxylation Reactions

**Figure 1** Catalytic cycle of p-hydroxybenzoate hydroxylase. In the first step, pOHB and NADPH bind (k1) and the FAD becomes reduced (k2). NADP is released (k3) and O2 reacts (k4) to form the C4a-hydroperoxy-FAD (E FlHOOH-S) in complex with substrate. Hydroxylation occurs via k5 to yield the dienone form of product and the C4a-hydroxy-FAD (Int II). Tautomerization yields 3,4-dihydroxybenzoate in complex with the enzyme (E FlH OH-P). Dissociation of 3,4-DOHB and H2O via k7 leads to free enzyme (E Flox). Uncoupling occurs via the loss of H2O2 from the C4a-hydroperoxy flavin (k8).

is incorporated into pOHB, and the other is reduced to water. Depending on the conditions, the primary rate-determining step in catalysis is either the dissociation of NADP from the enzyme after reduction, or the dissociation of 3,4 DOHB after oxidation (or both). The transient chemical intermediates in this oxidative half reaction were initially discovered and studied in the early 1970s by using stopped-flow spectrophotometry and fluorimetry to track the changes in FAD (8). Reduced, oxidized, and C4a-substituted flavins have unique absorption spectra (Fig. 2), and sometimes have unique fluorescence characteristics. For example, the C4a-hydroperoxyflavin and C4a-hydroxyflavin species often have very similar UV-visible spectra. However, the hydroxyflavin is often very fluorescent whereas the hydroperoxyflavin is not. These properties, which make it possible to study the kinetics directly of the chemical changes in catalysis with considerable specificity and sensitivity, are an enormous help to mechanistic investigations of flavoproteins.

**Role of proton exchanges and electrostatics in catalysis**

As is the case with most enzymes, catalysis by PHBH is sensitive to pH because proton associations and dissociations are important to the overall process. The structure of PHBH shows that the isoalloxazine in the active site is in a positive electrostatic field (10, 11). When pOHB is bound in the active site, its 4-OH group is linked into a chain of hydrogen bonds to the surface of the protein (Fig 3 Reference 12). These structural features are highly conserved in PHBH. This enzyme was an excellent model to study the effects of proton exchanges because pH-dependent changes occur to the spectra of both FAD and pOHB in the protein, and it is feasible to make specific stable mutants that involve the hydrogen bond chain and other residues in the enzyme.

When pOHB binds to the oxidized enzyme at neutral pH in the positive electrostatic field of the active site, the phenolic
The role of protein dynamics in catalysis has to be inferred from several pieces of mostly indirect evidence, which includes important kinetic observations. For example, why does NADPH reduce FAD in the enzyme 105-fold faster with pOHB bound than without, when the thermodynamics are essentially the same for each reaction? Under specific conditions, the 3-D structure of the enzyme is significantly different (10, 17), and these different structures can be logically ascribed to catalysis. For example, individual mutant forms of PHBH stabilize the structure in one or another particular conformation with specific dramatic effects on catalysis. Recently, single molecule fluorescence studies of PHBH have been used to demonstrate a specific conformational change in PHBH and to suggest a possible role of the protein dimer in promoting binding of pOHB (18). The accumulated evidence has provided the following minimum picture of functional protein dynamics.

The enzyme forms at least three highly interactive conformations during catalysis—named the in, out, and open conformations. Without ligands bound, the enzyme is in rapid equilibrium between the in and open conformations (which were measured in single-molecule studies) (18). It has been suggested that because of this dynamic equilibrium, the native protein fails to form high-resolution crystals without substrates bound. The open conformation has a more open active site and permits solvent access. This conformation has been illustrated by the structure of the Arg220Gln variant of PHBH, which gives a high-resolution crystal structure without any substrate bound (17). To form the open conformation, the highly conserved peptide loop that includes a strained peptide bond between Arg44 and Ala45, and that covers the substrate aromatic ring toward electrophilic attack. This effect was demonstrated by the ~100-fold decrease in the rate of hydroxylation in the mutant, Tyr201Phe, which cannot release the 4-OH proton (15). The positive electrostatic field in the active site is also important to the catalysis of hydroxylation. It was shown that making the field more positive increases the rate of hydroxylation, and making the field more negative decreases the rate of hydroxylation (16). These observations are consistent with the fact that the flavin hydroperoxide has a limited ability to carry out electrophilic attacks. For example, PHBH fails to hydroxylate either the less reactive benzene or p-fluorobenzene substrate analogs, even though they bind appropriately in the active site. After formation of the 3,4 DOHB product in the active site, the hydrogen bond network can again promote dissociation of the proton from the 4-OH position of 3,4 DOHB as part of a controlled conformational change in the protein that results in release of the 3,4 DOHB product and loss of water from the flavin (see the next section).

Protein dynamics in catalysis

In the positive electrostatic field of the active site, reduced FAD, which is formed by hydride transfer from NADPH, exists primarily as its anion (Fig. 3), which is ideal for its reaction with oxygen. The reaction of O2 with the anionic reduced flavin results in a peroxide that must be protonated to form the hydroperoxide for the subsequent electrophilic attack on pOHB. This protonation is extremely fast and has only been detected at high pH with a mutant form of PHBH (14). It is not known what is the source of the proton, but it is likely to be delivered through a water channel on the re side of the isalloxazine ring. The hydrogen bond network also has an important function during the oxidative half-reaction. With the formation of the neutral flavin hydroperoxide, the 4-OH proton of pOHB can be removed and transferred to the hydrogen bond network; this transfer promotes electrophilic attack of the distal hydroxyl of the C4a-hydroperoxide on the 3-position of pOHB (14). Thus, the enzyme activates the substrate aromatic ring toward electrophilic attack. This effect was demonstrated by the ~100-fold decrease in the rate of hydroxylation in the mutant, Tyr201Phe, which cannot release the 4-OH proton (15). The positive electrostatic field in the active site is also important to the catalysis of hydroxylation. It was shown that making the field more positive increases the rate of hydroxylation, and making the field more negative decreases the rate of hydroxylation (16). These observations are consistent with the fact that the flavin hydroperoxide has a limited ability to carry out electrophilic attacks. For example, PHBH fails to hydroxylate either the less reactive benzene or p-fluorobenzene substrate analogs, even though they bind appropriately in the active site. After formation of the 3,4 DOHB product in the active site, the hydrogen bond network can again promote dissociation of the proton from the 4-OH position of 3,4 DOHB as part of a controlled conformational change in the protein that results in release of the 3,4 DOHB product and loss of water from the flavin (see the next section).
right orientation and distance from the C4 of the nicotinamide of the bound NADPH to foster rapid hydride transfer and reduction of the flavin. In PHBH, this chemically productive orientation is associated with the formation of a charge-transfer complex between flavin and NADPH that can be monitored spectrally at long wavelengths (20). Similar conformational control of reduction is probably common to many one-component hydroxylases. However, the mechanism for triggering the change is likely to be different with each enzyme.

A free FAD is reduced, the enzyme reverts to the in conformation, probably because of the electrostatic attraction between the newly formed reduced flavin anion and the positive electrostatic field of the active site after protonation of pOHB by the hydrogen bond network (14). The conformational change is somehow linked to the dissociation of NADP, which is the rate-determining step in catalysis under optimal conditions for catalysis. The reaction with oxygen and the subsequent hydroxylation occurs with the protein in the in conformation. As with all of these hydroxylases, the enzyme reacts rapidly with oxygen in the enclosed hydrophobic active site. For PHBH, a hydroxylation with average reactivity with oxygen, the rate constant is $2 \times 10^9 \text{M}^{-1} \text{s}^{-1}$ at 4 °C, which is at least 10-fold faster than the reaction of oxygen with free reduced flavin. A structural explanation probably existed for this reaction. PHBH has a hydrophobic pocket on the face of the flavin with just the right space for oxygen to contact the isoalloxazine ring for reaction. Because the redox potential for the first electron transfer to oxygen (forming $\text{O}_2^-$) is not favorable, the electron transfer occurs over a short distance to oxygen, which is in the pocket, thus, the $\text{O}_2^-$ that is initially formed as a capped pair with the flavin radical rapidly collapses to form the peroxide. This property of short-range electron transfer may help to avoid the formation of free superoxide that might otherwise escape to solution. In the conformation, the flavin hydroperoxide has sufficient stability to execute hydroxylation of pOHB with 100% efficiency under the right conditions ($pH$ between 6 and 7). The ring of the product, 3,4-DHOB, is rotated in the active site by comparison to pOHB, which helps trigger a conformational shift to the open form, permitting product dissociation and loss of water from the flavin to complete the catalytic cycle. PHBH, like most of these hydroxylases, is subject to substrate (pOHB) inhibition. Kinetic analysis has shown that mM concentrations of pOHB rapidly bind to PHBH after dissociation of the 3,4 DHOB product and thus stabilize in conformation with the flavin trapped as the flavin-C4a-hydroxide (Fig. 3). This dead-end complex inhibits catalysis.

For PHBH to function as an efficient catalyst, the series of four conformational changes in a catalytic cycle have to be fast and coordinated compared with the chemical reactions of catalysis. For example, the observation that the reduction of flavin under optimal conditions for catalysis exhibits a full primary deuterium isotope effect (13) implies that the rate of reduction of flavin is limited by hydride transfer and not by conformational rearrangements. However, when the enzyme is stabilized in the in conformation (as with the mutant form, A1a45G/I), then a large fraction of flavin reduction becomes much slower under the same conditions and shows only a small deuterium isotope effect (21).

Two-Component Flavin-Dependent Hydroxylases

The one-component hydroxylases like PHBH have been recognized since the process of biological oxidation was first demonstrated in the 1960s. Many similar enzymes have been discovered over the years. For example, although phenol hydroxylation from yeast has very little sequence similarity to PHBH, it has a 3-D structure with the same folding pattern as PHBH (22), and almost certainly undergoes similar protein dynamics in catalysis. In the 1990s, another completely different group of flavin-dependent hydroxylases was found in bacteria—the two-component enzymes. These two-component hydroxylases consist of one protein that catalyzes the reduction of flavin and an oxygenase that binds the reduced flavin product and carries out the hydroxylation step. These enzyme systems are often called two-component-flavin diffusible monooxygenases (TC-FDM). In this section, we examine briefly p-hydroxyphenylacetate hydroxylase as a model for a large subset of this group of enzymes that, like PHBH, hydroxylate aromatic compounds. Another group of two-component enzymes has evolved with a different function—to halogenate substrates. We use tryptophan halogenase as an example of this expanding group. Finally, we refer to bacterial luciferase. This unique enzyme was the first two-component flavin-dependent hydroxylase to be recognized about 40 years ago. In the past 15 years, hydroxylations that involve catalysis by two-component enzymes of a huge diversity of substrates have been found. A summary of these can be found in Reference 2. The special case of epoxidation reactions should be noted. The best example is styrene monooxygenase (23, 24). As a two-component enzyme, it is similar to other enzymes of this type and uses a C4a-hydroperoxyflavin as a hydroxylating intermediate (24). However, the oxygenase component has no sequence similarity to other equivalent oxygenase components (23). We look forward to detailed mechanistic studies on hydroxylation by this enzyme because it may provide new insights into the process of oxygen insertion into double bonds by flavinoids as well as aromatic hydroxylations.

Para-hydroxyphenylacetate hydroxylase (HPAH)

This enzyme system, which has been isolated from several different organisms, subdivides into two distinct groups. The enzyme system from Acinetobacter baumannii is a model for one group, and the enzyme system from Pseudomonas aeruginosa is a model for the other. The P. aeruginosa hydroxylase is very similar to HPAH from E. coli (25), but it is more amenable for mechanistic studies. HPAH from A. baumannii is isolated as a reductase that contains flavin mononucleotide (FMN) and an oxygenase without flavin. The reductase is a homodimer with monomer mass of 35,000, whereas the oxygenase is a homotetramer with monomer mass of 47,000 Da. The two proteins together carry out a catalytic reaction analogous to that of PHBH; curiously, no detectable complex formation exists between the...
reductase and oxygenase proteins (26). The reductase reaction is almost identical to the reductive half-reaction of PHBH. NADH reduces the bound FMN slowly unless p-hydroxyphenylacetate (HPA) is present. However, when HPA is bound, the reductase binds FMN tightly and rapid reduction of the FMN ensues (27). The reduced FMN (FMNH\(^\cdot\)) is bound much less tightly to the reductase than is the oxidized FMN. In contrast, the oxygenase binds FMN\(^\cdot\) tightly and then catalyzes a reaction similar to the oxidative half-reaction of PHBH. Nevertheless, notable differences from PHBH are observed. Without HPA, the oxygenase forms a stable flavin hydroperoxide, then binds HPA to form the product, 3,4-dihydroxyphenylacetate and releases it into solution (28). To complete catalysis, the separate proteins must coordinate their functions. It has been shown that when FMN is reduced by the reductase in the presence of HPA, FMNH\(^\cdot\) dissociates into solution with a rate constant of 80 s\(^{-1}\) at 4 °C (26). FMNH\(^\cdot\) binds FMN\(^\cdot\) very rapidly to the oxygenase and reacts with oxygen with a rate constant of 1.1 × 10^8 M\(^{-1}\) s\(^{-1}\) at 4°C. Because these events are so fast, the reaction of oxygen with free FMNH\(^\cdot\) in transit between the two proteins is negligible, providing that the oxygenase is in excess over the available FMN, which is generally the case (25).

This two-component system is most effective with no external flavin added and with a slight excess of oxygenase active sites over reductase sites. Under these conditions, coupling of reducing equivalents from NADH to hydroxylation is maximal. How does the performance of this system compare with PHBH? Under the same conditions, turnover of PHA is ∼2 s\(^{-1}\) compared with ∼50 s\(^{-1}\) for PHBH. Moreover, HPAH always generates a small amount of H\(_2\)O\(_2\), but PHBH generates almost none. The HPAH from P. aeruginosa consists of a reductase that is a flavin hydratase with monomer molecular weight of 13,000 and is isolated with only some protein that contains FAD. The oxygenase is isolated as a colorless homotramer with monomer mass of 46,000. In this system, in contrast to the A. baumannii HPAH, the reductase does not bind HPA, and it catalyzes reduction of FAD by NADH and release of FAD\(^\cdot\) into solution with no regulation of catalysis by HPA. Thus, for this HPAH to be effective, regulation of the reaction with oxygen must reside with the oxygenase. The reductase that is present reduces any free FAD for immediate uptake by the oxygenase, which, as established in E. coli (25), is present in cells at higher concentrations than the reductase. It has been found that the oxygenase forms a very stable C4a-hydroperoxyflavin, and this species is probably the predominant form of the enzyme in cells (25, 29). When the C4a-hydroperoxyflavin form of oxygenase binds HPA, it rapidly converts the HPA to product and releases it into solution. The release of FAD from the oxygenase follows and is the rate-determining step in catalysis (26). This HPAH has a turnover rate similar to the enzyme from A. baumannii. The structures of the oxygenases from the two types of HPAHs are different in detail (particularly the active sites) but have similar core scaffolds (30). A great deal of research must be performed to understand the unique dynamics of the P. aeruginosa-type oxygenase that probably do not occur in the A. baumannii-type oxygenase.

**Tryptophan-7-halogenase (TH)**

Many thousands of halogenated metabolites in nature become halogenated by reactions catalyzed by flavoproteins. To date, only a couple of enzymes have been studied in any detail. TH has now become the model for these enzymes. TH catalyzes the chlorination of the 7-position of tryptophan for the synthesis of pyrroloindin, which is an antifungal agent from P. fluorescens, and for the synthesis of rebeccamycin, which is an anticancer agent from Saccharopolyspora rebaudiana. TH consists of a flavin reductase and a halogenase. The halogenase is similar to the simple reductase of HPAH from P. aeruginosa, whereas the halogenase has structural similarity to PHBH (31). Thus, in general terms, TH is a two-component system with FAD that difuses between two proteins, but TH has evolutionary links to the one-component hydroxylases.

Recent transient-state kinetic analysis of the reaction of the halogenase component of TH has been very revealing (32). The halogenase binds FAD\(^\cdot\) and reacts with oxygen to form a stabilized C4a-hydroperoxyflavin. With Cl\(^{-}\) present, this intermediate converts to a fluorescent C4a-hydroxyflavin, which finally decomposes to FAD and water. Tryptophan has no influence on the kinetics of the flavin reactions. The implication of these observations is that Cl\(^{-}\) is hydroxylated by the hydroperoxide to form HOCl, which is subsequently used in the chlorination of tryptophan. The details of the regiospecific chlorination remain an outstanding question, although it is known from the crystal structures of TH from both P. fluorescens (PmA) and from S. aerocolonigenes (RebH) that tryptophan is separated from FAD by ∼10 Å, and they are connected by a channel in the active site. Chlorination is a slow reaction (first-order with a rate constant of 0.05 s\(^{-1}\) at 25°C) and inefficient—only ∼3% of starting reduced FAD equivalents result in chlorinated product (32). A static crystal structure is probably inadequate for a complete understanding of the details of the catalytic process. Unlike other hydroxylases, the halogenase undergoes some dynamic changes in the oxygen reactions that are not understood. For example, if the halogenase binds FAD\(^\cdot\), then the subsequent reaction with oxygen fails to form the flavin hydroperoxide, and only FAD and H\(_2\)O\(_2\) are produced. In contrast, if reduced flavin is added to RebH and O\(_2\) is present, this intermediate converts to a fluorescent C4a-hydroxyflavin, which is subsequently used in the chlorination of tryptophan. Recently, it was shown that when RebH, FADH\(^\cdot\), and O\(_2\) react in the absence of tryptophan, a long-lived intermediate (t\(_{1/2}\) = 63 h at 4°C) formed that could halogenate tryptophan (33). This result suggested that the initially formed HOCl reacted with a group on the enzyme to produce an intermediate, such as a chloramine, that was responsible for halogenating tryptophan. Lysine 97, which is required for catalysis, is a likely candidate, and it is in a position that could halogenate tryptophan in the 7-position.

**Bacterial luciferase**

A luciferase is an enzyme that catalyzes chemiluminescence. This property is widespread in marine environments (particularly deep-sea), in which many higher organisms cultivate specific bacteria for the production of light. All known bacterial luciferases are homologous. Although this enzyme has been
Flavin-Mediated Hydroxylation Reactions

studied for decades, its mechanism is still partly a mystery. Like the enzymes just discussed, luciferase is composed of a reductase and an oxygenase. These proteins are usually expressed from a single operon similar to other two-component enzymes (34). The reductase catalyzes an unregulated reduction of FMN by NADH (34). The oxygenase, which is usually referred to as luciferase, is structurally and functionally unique. It catalyzes the hydroxylation of a fatty aldehyde to form a fatty acid in a manner similar to the Bayer-Villiger enzymes discussed in the next section. However, mysteriously, the free energy liberated in the reaction is captured in an excited-state flavin. The usual intermediates in an oxygenase reaction are observed (35, 36)—formation of a flavin-C4a-peroxide, followed by a flavin hydroxide, return to oxidized flavin, and release of flavin from the enzyme. It is the excited-state C4a-hydroxyflavin that is the light emitter. Hypotheses put forward for the capture of an excited-state C4a-hydroxyflavin can be found in Reference 37. Currently, we do not even know how FMN is bound in luciferase. When we have a clear picture of the active site of this enzyme (showing the interactions between protein and FMN in different oxidation states) and the complex dynamics that seem to be part of the process, we may finally understand this chemistry.

Other One-Component Hydroxylases

Although the one-component aromatic hydroxylases (represented by PHBH above) are a major group of flavoproteins, they are not the only one-component flavin-dependent oxygenases. Many flavoproteins oxidize aliphatic compounds that are susceptible to nucleophilic attack by flavin peroxides. Oxygenation of aldehydes and ketones is a common reaction catalyzed. The mechanism is considered to involve a Bayer-Villiger rearrangement. We briefly examine the model enzyme cyclohexanone monooxygenase, which oxygenates the carbonyl of cyclohexanone and forms a lactone (cyclic ester) via a Bayer-Villiger rearrangement. A mother group of flavoproteins is important in the defensive armory of animals against electron-rich foreign compounds, and it is also involved in some metabolic pathways in animals and plants. An appropriate model for this group of enzymes is mammalian flavin monooxygenase. We finish with an example of a recently discovered enzyme BluB that although structurally unrelated to other flavoprotein hydroxylases, it catalyzes the oxygenation and conversion of FMN to dimethylbenzimidazole using flavin C4a-intermediates. Other novel flavoprotein hydroxylases/oxygenases are yet to be discovered.

Cyclohexanone monooxygenase (CHMO)

The CHMO that has been studied most extensively comes from a strain of Acinetobacter calcoaceticus. The monomeric form of the enzyme contains one FAD and has a molecular weight of 62,000. The structure of a homolog of this enzyme (phenylacetone monooxygenase) has been published (38). The structure and kinetic analysis of CHMO show that the overall catalytic cycle is similar to that of PHBH, although there are important differences (39). In contrast to PHBH, the reductive half-reaction is not regulated by substrate. Thus, the formation of cellular reactive oxygen species is prevented by a different mechanism.

After NADPH binds and reduces the FAD, independently of cyclohexanone, the NADP product remains bound to the enzyme, and the resultant reduced enzyme reacts rapidly (10^7 M^-1 s^-1) to form a stable NADP-flavin C4a-peroxide complex. In fact, it is this species that is the predominant form of the enzyme in the cell. The bound NADP has a critical role in stabilizing the flavin peroxide. When cyclohexanone binds to the enzyme, it is converted rapidly to the 7-membered ring, 1-caprolactone, with the ring-oxygen coming from the peroxide, which leaves the flavin in the form of a C4a-hydroxy-FAD. A faster it loses water to form oxidized FAD, the rate-determining step in catalysis is the dissociation of NADP. Kinetic and structural evidence suggests that the mechanism involves a nucleophilic attack of the flavin peroxide on the carbonyl carbon to form a Criegee intermediate that collapses to form product and flavin hydroxide (Fig. 4). These enzymes avoid producing reactive oxygen species because the C4a-peroxy-FAD is very stable until substrate binds and reacts rapidly with it. There is great potential for research into the relationship between the protein structure and function of these enzymes that use C4a-flavin peroxides.

Flavin monooxygenase (FMO)

The first known enzyme of this class was isolated from pig liver and was described as liver microsomal FAD-containing monooxygenase (40). It has several properties similar to CHMO, but it mainly oxidizes a wide range of heteroatom-containing soft nucleophiles, which are generally electron-rich compounds. Mammalian FMOs participate with cytochrome P450s in the oxygenation of hydrophobic xenobiotic compounds, which includes many drugs, making them more water soluble and ready for coupling with glutathione and other compounds so they can be excreted. Extensive rapid and steady-state kinetics studies have been carried out on pig liver FMO (40-42). FMO is first reduced by NADPH, and, like CHMO, the enzyme retains the NADP product. The reduced flavin reacts with O2 to form a stable NADP-C4a-hydroperoxyflavin. Its return to oxidized FAD with release of H2O2 occurs over a period of many minutes, which depends on the conditions. However, if a suitable heteroatom-containing substrate is present, FMO reacts quickly to oxygenate that substrate, which forms hydroxyl amines, sulfenic acids, and so on. The resulting C4a-hydroxyflavin releases water to reform the oxidized FAD in what is often the rate-determining step of catalysis. Therefore, the turnover numbers for most substrates are very similar. In the absence of NADP, the C4a-hydroperoxide is so unstable that it is almost undetectable. Five documented functional genes and some pseudogenes for FMO are observed in the human genome (43). FMOs also exist in plants (44) where they participate in numerous activities, which include the biosynthesis of the plant hormone, auxin.
BluB—A protein that converts FMN into dimethylbenzimidazole

The biosynthesis of vitamin B_{12} has intrigued some of the most famous chemists, and research on this topic has resulted in four Nobel prizes. Yet until recently, the synthesis of dimethylbenzimidazole (DMB), which is the alpha-axial ligand to B_{12}, has remained a mystery, except that it was known to derive from FMN. It has now been demonstrated that the BluB proteins from Sinorhizobium meliloti (45) and Rhodopseudomonas rubrum (46) catalyze the reaction of FMNH_{2} with O_{2} to form DMB and erythrose-4-phosphate. The C1′ from the ribityl moiety of FMN becomes the C2 of DMB, and the dimethyl-diamino benzene portion comes from the isoalloxazine ring of the flavin. This amazing reaction is unprecedented, and clearly, it must involve some unusual chemistry. Recent unpublished studies (D. Ballou and M. Taga) have shown that BluB binds FMNH, and this complex reacts with O_{2} with a rate constant $k_{1}$ of $10^6 M^{-1}s^{-1}$ at $4^\circ C$ to form a C4a-hydroperoxy-FMN. This species oxygenates itself and decays in several steps to form about 50% DMB with the remaining product being FMN. Thus, BluB is a special case of a flavin oxygenase. Future identification of intermediates along the pathway to DMB will surely provide some novel chemistry.

References


Further Reading

The further reading list is incorporated into the References list above (especially Refs. 3–5) in the form of review articles and book chapters. Breaking news about the field of flavoproteins can be found in journals such as Biochemistry, Chemical Reviews, and Current Opinion in Chemical Biology.
Flavin-Mediated Hydroxylation Reactions

found in Flavins and Flavoproteins 2008, which will be published in 2008. This book will contain proceedings from the 16th International Symposium on Flavins and Flavoproteins to be held in Spain in June 2008.

See Also

Cytochrome P450 Monooxygenases, Chemistry of
Enzyme Catalysis, Roles of Structural Dynamics in
Enzyme Catalysis, Chemical Strategies for
NAD+ -Dependent Enzymes, Chemistry of
Oxygen-Activating Enzymes, Chemistry of
Transient State Enzyme Kinetics
Flavoenzymes, Chemistry of
Flavoenzymes are omnipresent in nature and are involved in many cellular processes. Flavoenzymes typically contain the vitamin B2 derivatives FAD and FMN as a redox-active prosthetic group. By varying the protein environment around the isoalloxazine ring of the flavin, evolution has created a great diversity of flavoprotein active sites and catalytic machineries. Most flavoenzymes perform one- or two-electron redox reactions and belong to the following groups: Flavoprotein reductases primarily use NAD(P)H as electron donor and pass these electrons to a protein substrate or another electron acceptor. Flavoprotein dehydrogenases oxidize organic substrates and mainly use quinones and electron transfer proteins as electron acceptors. Flavoprotein disulfide oxidoreductases contain active-site thiols. They use a dithiol substrate and NAD\(^+\) to form a disulfide product and NADH or act in the other direction yielding NAD(P)\(^+\) and a reduced (dithiol) product. Some flavoprotein (d)thiol oxidoreductases stabilize a reactive and reversibly oxidized cysteine in their active site. Flavoprotein oxidases catalyze the conversion of a substrate single bond to a double bond. The reduced flavin generated during this reaction is reoxidized by molecular oxygen to form hydrogen peroxide. Flavoprotein monoxygenases mainly use NAD(P)H as an electron donor and insert one atom of molecular oxygen into their substrates. By doing so, they act in different biological processes, ranging from lignin degradation and detoxification to the biosynthesis of polyketides and plant hormones.

Flavoenzymes are widespread in nature and are involved in many different chemical reactions. Flavoenzymes contain a flavin mononucleotide (FMN) or more often a flavin adenine dinucleotide (FAD) as redox-active prosthetic group. Both cofactors are synthesized from riboflavin (vitamin B2) by microorganisms and plants. Most flavoenzymes bind the flavin cofactor in a noncovalent mode (1). In about 10% of all flavoenzymes, the isoalloxazine ring of the flavin is covalently linked to the polypeptide chain (2, 3). Covalent binding increases the redox potential of the flavin and its oxidation power, but it may also be beneficial for protein stability, especially in flavin-deficient environments.

Flavoenzymes constitute about 2% of all biological catalysts and are classified in several ways. One classification is based on EC number (enzyme nomenclature) and refers to the type of reaction catalyzed. More sophisticated classifications concern the inclusion of sequence, fold, and function. Historically, a distinction is made between “simple” and “complex” flavoenzymes (4). The latter proteins contain besides flavin other cofactors like heme, tetrahydrobiopterin, and metal ions.

The catalytic cycle of each flavoenzyme consists of two distinct processes, the acceptance of redox equivalents from a substrate and the transfer of these equivalents to an acceptor. Accordingly, the catalyzed reactions consist of two half-reactions: a reductive half-reaction in which the flavin is reduced and an oxidative half-reaction, in which the reduced flavin is reoxidized. This review summarizes the chemistry of “simple” flavoprotein reductases, dehydrogenases, (d)thiol oxidoreductases, oxidases, and monoxygenases (Table 1) (5–40). This grouping provides a good appreciation about what type of common mechanisms can be distinguished and what type of substrates can be converted. Information on the chemistry of “complex” flavoenzymes can be found in the Further Reading section.

Biological Background

The intrinsic chemical properties of the isoalloxazine nucleus, as modulated by the protein environment, are at the heart of the success of flavoenzymes in nature. The chemical versatility,
Table 1 Examples of "simple" flavoenzymes

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>EC number</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reductases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAD(P)H:quinone reductase</td>
<td>1.6.5.2</td>
<td>(5)</td>
</tr>
<tr>
<td>Flavin reductase</td>
<td>1.5.1.30</td>
<td>(6, 7)</td>
</tr>
<tr>
<td>Ferredoxin NADP⁺ reductase</td>
<td>1.18.1.2</td>
<td>(8, 9)</td>
</tr>
<tr>
<td>Dehydrogenases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acyl-CoA dehydrogenase</td>
<td>1.3.99.3</td>
<td>(10-12)</td>
</tr>
<tr>
<td>L-galactono-1,4-lactone dehydrogenase</td>
<td>1.3.2.3</td>
<td>(13)</td>
</tr>
<tr>
<td>Disulfide oxidoreductases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dihydrolipoamide dehydrogenase</td>
<td>1.8.1.4</td>
<td>(14)</td>
</tr>
<tr>
<td>Glutathione reductase</td>
<td>1.6.4.2</td>
<td>(15, 16)</td>
</tr>
<tr>
<td>Thioredoxin reductase</td>
<td>1.6.4.5</td>
<td>(15, 17, 18)</td>
</tr>
<tr>
<td>Sulfinic acid oxidoreductases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NADH peroxidase</td>
<td>1.11.1.1</td>
<td>(16, 19)</td>
</tr>
<tr>
<td>NADH oxidase</td>
<td>1.6.99.x</td>
<td>(16, 19)</td>
</tr>
<tr>
<td>Oxidases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-amino acid oxidase</td>
<td>1.4.3.3</td>
<td>(20, 21)</td>
</tr>
<tr>
<td>M onoamine oxidase</td>
<td>1.4.3.4</td>
<td>(22, 23)</td>
</tr>
<tr>
<td>N/nitroblue oxidase</td>
<td>1.7.3.2</td>
<td>(24)</td>
</tr>
<tr>
<td>Vanillyl-alcohol oxidase</td>
<td>1.1.3.38</td>
<td>(3, 25, 26)</td>
</tr>
<tr>
<td>Glucose oxidase</td>
<td>1.1.3.4</td>
<td>(27)</td>
</tr>
<tr>
<td>Cholesterol oxidase</td>
<td>1.1.3.6</td>
<td>(12, 28, 29)</td>
</tr>
<tr>
<td>Acyl-CoA oxidase</td>
<td>1.3.3.6</td>
<td>(30)</td>
</tr>
<tr>
<td>Monooxygenases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-hydroxybenzoate 3'-hydroxylase</td>
<td>1.14.12.2</td>
<td>(31-34)</td>
</tr>
<tr>
<td>Cyclohexanone monoxygenase</td>
<td>1.14.13.12</td>
<td>(35, 36)</td>
</tr>
<tr>
<td>Phenylacetone monoxygenase</td>
<td>1.14.13.92</td>
<td>(37)</td>
</tr>
<tr>
<td>Flavin-containing monoxygenase</td>
<td>1.14.13.8</td>
<td>(38)</td>
</tr>
<tr>
<td>Tryptophan 7-halogenase</td>
<td>1.14.13.x</td>
<td>(39, 40)</td>
</tr>
</tbody>
</table>

of the flavin cofactor is used in a wide variety of biological processes that range from energy production, light emission, protein folding, and neural development to detoxification, apoptosis, chromatin remodeling, and DNA repair (41). With the developments in molecular life sciences, it is expected that many more functions of flavoenzymes will develop. The biological background of certain flavoprotein reductases, dehydrogenases, (di)thiol oxidoreductases, oxidases, and monooxygenases is summarized below.

Reducases

Flavoprotein reductases have many important cellular functions. NADH-cytochrome b5 reductase (EC 1.6.2.2) is a crucial housekeeping enzyme that controls the level of iron in the blood. Mitochondrial NADPH-cytochrome P450 reductase (EC 1.6.2.4) is involved in hepatitis drug metabolism by transferring electrons to many cytochrome P450 isoenzymes. Cytosolic NAD(P)H quinone reductase (NQO1; EC 1.6.5.2) protects cells from oxidative stress by catalyzing the reduction of exogenous and endogenous quinones to the corresponding hydroquinones. Recent studies suggest that mammalian quinone reductases also control the lifespan of transcription factors, such as p53, and hence participate in the development of apoptosis and cell transformation (5). Flavin reductases (EC 1.5.1.30) are widespread in microorganisms and use riboflavin, FMN, or FAD as substrate (6). The reduced flavin product is used by other enzymes for various purposes like for instance the emission of light (7). Plant-type ferredoxin NADP⁺ reductase (FNR; EC 1.18.1.2) acts in the reverse sense. Being involved in carbon fixation this enzyme receives electrons one at a time from ferredoxin and then transfers them to a two-electron step to NADP⁺ (8).

Dehydrogenases

Flavoprotein dehydrogenases are also widespread. Many of them occur in mitochondria where they are involved in energy production and the biosynthesis of essential nutrients. Dysfunction of these enzymes may result in oxidative stress and neurobehavioral deficits. Acyl-CoA dehydrogenases (EC 1.3.3.9) play a crucial role in the mitochondrial β-oxidation of fatty acids (10). The reduced forms of these enzymes are reoxidized in two one-electron steps by ETF (electron transferring flavoprotein) (42). Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency is a widely occurring inherited genetic disorder. During fasting or illness, it results in continued glucose consumption and strongly reduced ketone body production. L-galactono-1,4-lactone dehydrogenase (GALDH; EC 1.1.3.38) catalyzes the terminal step of vitamin C biosynthesis in plants (13). The mitochondrial enzyme uses cytochrome c as electron acceptor. GALDH homologues in animals (L-gulono-1,4-lactone...
oxidase), yeast (D-arabinono-1,4-lactone oxidase), and fungi (D-glucosonolactone oxidase) use molecular oxygen as electron acceptor and are involved in the synthesis of L-ascorbate or its analogs D-erythorbate and D-erythroascorbate (3).

Disulfide oxidoreductases
Flavoprotein disulfide oxidoreductases are crucial metabolic and detoxification enzymes. Dihydrolipamide dehydrogenase (EC 1.3.1.4) catalyzes the NAD^+ dependent oxidation of dihydrolipoyl groups, which are covalently attached to the lipoyl domain of the acyltransferase components of the mitochondrial a-ketooacid dehydrogenase and glycine decarboxylase multienzyme complexes. Mutations to this homodimeric flavoprotein cause the often-fatal human disease known as E3 deficiency (14). Glutathione reductase (EC 1.6.4.2) and thioredoxin reductase (TrxR; EC 1.6.4.5) supply cells with high levels of dithiols, which are essential in keeping other thiols reduced (15). In red blood cells, up to 10% of the glucose consumption may be used for the production of reduced glutathione. The TrxR/thioredoxin system is involved in important physiological functions, such as cell growth, inflammation reactions, and apoptosis and regulation.

NADH peroxidase (Npx; EC 1.11.1.7) and NADH oxidase (Nox; EC 1.6.99.x) are flavoprotein oxidoreductases-related enzymes that contain a single redox-active cysine (16). They supply strictly fermentative bacteria with NAD^+ for glycolysis and play an important role in redox signaling in response to oxidative and nitrosative stress (18).

Oxidases
Flavoprotein oxidases are ubiquitous enzymes involved in the biosynthesis and biodegradation of a huge variety of compounds. D-amino acid oxidase (DAAO; EC 1.4.3.3) is the prototype amino acid oxidase. This peroxisomal enzyme is spread from yeast to humans. Mammalian DAAO has been connected to the brain D-serine metabolism and to the regulation of glutamatergic neurotransmission (20). The outer mitochondrial membrane monooxygenase (MAD; EC 1.4.3.4) is perhaps the most well-known amine oxidase. In addition to the oxidation of neurotransmitters, such as dopamine and serotonin, this enzyme also oxidizes ingested amines such as phenethylamine and tyramine to prevent their functioning as false neurotransmitters. The human isoforms MAO-A and MAO-B are involved in many diseases and are important targets for antidepressant and neuroprotective drugs (22). Acyl-CoA oxidases (EC 1.3.3.6) are acyl-CoA dehydrogenase homologs involved in peroxisomal fatty acid breakdown in plants (30).

Monoxygenases
Flavoprotein monoxygenases are widely found in microorganisms and plants, and some are present in mammalian species (43). Aromatic hydroxylases like p-hydroxybenzoate 3-hydroxylation (PHBH; EC 1.14.13.2) are involved in lignin degradation. Bayer-Villiger monoxygenases like cyclohexanone monoxygenase (EC 1.14.13.22) participate in microbial catabolic pathways by converting ketones (or aldehydes) into esters or lactones (35). Flavin-containing monoxygenases (EC 1.14.13.8) oxidize nitrogen-containing compounds and primarily are found in mammals and plants. The mammalian isoforms assist in the detoxification of drugs and other xenobiotics, whereas the plant enzymes are involved in the biosynthesis of auxin, the metabolism of glucosinolates, and in pathogen defense (38). Flavoprotein halogenases play an important role in the biosynthetic pathways of antibiotics, antitumor agents, and other natural products (39).

Chemistry
The isoalloxazine moiety of the flavin cofactor forms the catalytic heart of a flavoenzyme. It can undergo one- and two-electron redox transitions and form covalent adducts with substrates and protein residues. The redox properties of the flavin cofactor are modulated by the protein environment. In free flavin, the one-electron reduced state is thermodynamically unstable. Flavoenzymes, however, can stabilize the neutral or anionic semiquinone state (see also Flavoprotein Protocols article). Oxidized (yellow) flavin has characteristic absorption maxima around 375 and 450 nm (Fig. 1b and 1c). The anionic (red) and neutral (blue) semiquinone show typical absorption maxima around 370 nm and 580 nm, respectively (Fig. 1b and 1c). During two-electron reduction to the (anionic) hydroquinone state, the flavin turns pale, and the absorption at 450 nm almost completely disappears (Fig. 1b and 1c). The optical properties of the flavin can be influenced through the binding of ligands (substrates, coenzymes, inhibitors) or the interaction with certain amino acid residues. In many cases, these interactions result in so-called charge-transfer complexes that give the protein a peculiar color.

The catalytic cycle of each flavoenzyme consists of a reduced half-reaction, in which the flavin is reduced, and an oxidative half-reaction, in which the reduced flavin is oxidized. The reduction and oxidation steps are in many cases irreversible, which enables the direct characterization of reaction intermediates (see “See Also” section and the Further Reading List).

Reductases
Flavoprotein reductases primarily use NAD(P)H (AH2) as electron donor:

\[
\text{EFAD} + \text{AH}_2 \rightarrow \text{EFADH}_2 + \text{A}
\]

and pass these electrons to a protein substrate or another electron acceptor (B) in two single-electron steps:

\[
\text{EFADH}_2 + \text{B} \rightarrow \text{EFADH}^* + \text{BH}^* \\
\text{EFADH}^* + \text{B} \rightarrow \text{EFAD} + \text{BH}^*
\]
Chemistry of Flavoenzymes

Figure 1  (a) Redox states of the flavin cofactor. Flavoenzymes generally stabilize the anionic hydroquinone state ($pK_a$ free reduced flavin = 6.7).
(b) Oxidized (---), anionic semiquinone (···), and hydroquinone (—) forms of the FAD cofactor of Arabidopsis thaliana GALDH (adapted from Reference 13).
(c) Oxidized (---), neutral semiquinone (···), and hydroquinone (—) forms of the FMN cofactor of Bacillus subtilis flavodoxin YkuP (adapted from Reference 44).

or, alternatively, in one two-electron step:

$$\text{EFADH}_2 + B \rightarrow \text{EFAD} + \text{BH}_2$$

Several of the reductases mentioned here belong to the same structural family (the FNR family), and they are mechanistically related to each other (9). A two-electron reduction of the flavin by NAD(P)H in these enzymes typically involves the transient formation of an oxidized flavin-reduced pyridine nucleotide charge-transfer complex, which is followed by hydride transfer. After reduction, the flavin can transfer its electrons to different redox partners. With NADH:cytochrome b5 reductase, this transfer occurs in separate single-electron transfer steps. With NADPH:cytochrome P450 reductase, an enzyme containing which a “complex” flavoenzyme that contains two flavins, one electron is first intramolecularly transferred from FAD to FMN, before the reaction with cytochrome P450 takes place. With FNR, NADP$^+$ first has to bind to the oxidized form, before the very fast one-electron transfer from the specifically interacting reduced ferredoxin (Fdred) occurs (8). Subsequent dissociation of the oxidized ferredoxin (Fdox) is rate-limiting in catalysis. The enzyme semiquinone-NADP$^+$ complex then reacts with another reduced ferredoxin molecule to yield the flavin hydroquinone state. In the final steps of the catalytic cycle, the NADP$^+$ is reduced and the NADPH dissociates.
Disulfide oxidoreductases contain active-site thiols. They use a diol or NAD+ to form a disulfide and NADH or act in the reverse direction yielding NAD(P)+ and a reduced (dihiol) product.

\[
\begin{align*}
\text{EFAD} + \text{NAD}^+ & \rightarrow \text{EFAD-NAD}^+ \\
\text{EFAD-NAD}^+ + \text{F}dred & \rightarrow \text{Fdox-} \text{EFAD}^+ \text{NAD}^+ \\
\text{Fdox-} \text{EFAD}^+ \text{NAD}^+ & \rightarrow \text{EFAD}^+ \text{NAD}^+ + \text{Fdred} \\
\text{EFAD}^+ \text{NAD}^+ + \text{Fdred} & \rightarrow \text{Fdox-} \text{EFAD}^+ \text{NAD}^+ \\
\text{Fdox-} \text{EFAD}^+ \text{NAD}^+ & \rightarrow \text{FDox-} \text{EFAD-NADPH} \\
\text{FDox-} \text{EFAD-NADPH} & \rightarrow \text{FDox} + \text{EFAD} + \text{NADPH}
\end{align*}
\]

Disulfide oxidoreductases are structurally related homodimers with residues from both subunits participating in the catalysis of the two active sites (see the Further Reading for more information). The reductive half-reaction of lipoamide dehydrogenase involves the reduction of the active-site disulfide by the dihydrolipoamide substrate. During this reaction, the enzyme turns red, and a typical absorbance develops around 530 nm because of the formation of a charge-transfer species between the oxidized flavin and the nearby thiolate. The intensity of the charge-transfer absorption band is strongly pH dependent as influenced by the pKₐ modulating properties of a C-terminal histidine, which acts as an acid-base catalyst. Binding of NAD+ then shifts the reducing equivalents from the disulfide to the flavin because of an increase of the redox potential of the FAD.

The catalytic cycle is terminated by flavin reoxidation and dissociation of NADH.

Thioredoxin reductase (TrxR) acts in the reverse direction and shows a somewhat different mechanism, which is dependent on the protein source. Prokaryotes, plants, and lower eukaryotes contain a 35-kDa TrxR with one redox-active disulfide. Higher eukaryotes produce a 55-kDa TrxR that has either an additional redox-active disulfide or a selenenylsulfide in the flexible C-terminal part of the neighboring subunit. In low Mr TrxR, a large conformational change is required to move reducing equivalents from the apolar flavin site to the surface of the protein where the thioredoxin redox partner binds. In high Mr TrxR, this transfer is mediated by the second disulfide or selenenylsulfide, and the conformational changes required are comparatively small (17).

\[
\begin{align*}
\text{EFAD(S-S)} & \rightarrow \text{NADH} + \text{H}^+ \rightarrow \text{EFAD(SH)} + \text{NAD}^+ + \text{H}_2\text{O} \\
\text{EFAD(SH)} + \text{NAD}^+ & \rightarrow \text{EFAD(SH)} + \text{NAD}^+ + \text{H}_2\text{O}
\end{align*}
\]

The thiolate is stabilized by a histidine, which facilitates the nucleophilic attack of the reduced enzyme-NADH complex by hydrogen peroxide:

\[
\begin{align*}
\text{EFAD(SH)} & \rightarrow \text{EFAD(SH)} + \text{NAD}^+ \\
\text{EFAD(SH)} + \text{NAD}^+ + \text{H}_2\text{O} & \rightarrow \text{EFAD(SH)} + \text{NAD}^+ + \text{H}_2\text{O}
\end{align*}
\]

Again, NADH reduces the sulfenate via the flavin, which prepares the enzyme for the next cycle:
Chemistry of Flavoenzymes

(a) Proposed hydride transfer mechanism for substrate oxidation in MCAD. (b) Proposed polar nucleophilic mechanism for the reductive half-reaction in MAO.

In Nox, NADH reduction of the thiolate form involves the stabilization of the four-electron reduced enzyme:

$$\text{EFAD(S}^-\text{)} + \text{NADH} + \text{H}^+ \rightarrow \text{EFADH}_2(\text{S}^-\text{)}-\text{NAD}^+$$

Subsequent reaction with molecular oxygen then yields the active-site sulfenate:

$$\text{EFADH}_2(\text{S}^-\text{)}-\text{NAD}^+ + \text{O}_2 \rightarrow \text{EFAD(SO}^-\text{)}^-\text{NADH} + \text{H}_2\text{O}$$

The Nox oxidizing substrate must be activated prior to its reaction with the active site thiolate. Kinetic studies with the Cys42Ser variant have indicated the formation of a flavin C4a-hydroperoxide as a primary oxygenated intermediate in reoxidation of the reduced enzyme-NAD$^+$ complex (16). In the Cys42Ser variant, hydrogen peroxide is eliminated directly to give the oxidized enzyme. In wild-type Nox, Cys42 is in a favorable position for nucleophilic attack on the distal peroxylflavin oxygen, which yields the sulfenate and the flavin C4a-hydroxide in a monooxygenase type of reaction (vide infra). The flavin C4a-hydroxide eliminates water directly to give the oxidized enzyme.

Oxidases

Flavoprotein oxidases catalyze the conversion of a substrate single bond to a double bond. They differ from flavoprotein dehydrogenases in that the reduced flavin is oxidized by molecular oxygen to form hydrogen peroxide.

$$\text{EFAD} + \text{AH}_2 \rightarrow \text{EFADH}^- + \text{A} + \text{H}^+$$

$$\text{EFADH}^- + \text{O}_2 + \text{H}^+ \rightarrow \text{EFAD} + \text{H}_2\text{O}_2$$

The reaction of singlet-reduced flavin with triplet oxygen is spin forbidden and therefore involves the initial formation of a flavin semiquinone-superoxide anion radical pair (11). This caged radical pair can dissociate into oxygen radicals, or it can undergo a second electron transfer that produces hydrogen peroxide and oxidized flavin. Or, it can collapse to form a flavin-C4a-(hydro)peroxide covalent adduct. This adduct is the essential oxygenation species in flavoprotein monooxygenases (vide infra) but has never been detected in flavoprotein oxidases. The oxygen reactivity of flavoprotein oxidases can vary dramatically, and it is not entirely clear what determines this reactivity (45).

Flavoprotein oxidases come in many flavors. They can have different folds and topologies (46) and are active with many different substrates, which include (amino) acids, mono- and polyamines, nitroalkanes, aliphatic and aromatic alcohols, monosaccharides and oligosaccharides, thios, thioesters, and so on. Flavoprotein oxidases obey either a ping-pong or ternary complex kinetic mechanism. In the latter case, the product (A) remains bound during the oxidative half-reaction. The rate-limiting step in catalytic turnover is often represented by the rate of flavin reduction or the rate of product release. The kinetic mechanism and also the rate-limiting step of catalysis may vary depending on the type of (model) substrate used.
DAAO is one of the most extensively studied flavoprotein oxidases. The homodimeric enzyme catalyzes the strictly stereospecific oxidative deamination of neutral and hydrophobic D-amino acids to give α-keto acids and ammonia (Fig. 3a). In the reductive half-reaction the D-amino acid substrate is converted to the imino acid product via hydride transfer (21). During the oxidative half-reaction, the imino acid is released and hydrolyzed. Mammalian and yeast DAAO share the same catalytic mechanism, but they differ in kinetic mechanism, catalytic efficiency, substrate specificity, and protein stability. The dimeric structures of the mammalian enzymes show a head-to-head mode of monomer–monomer interaction, which is different from the head-to-tail mode of dimerization observed in Rhodotorula gracilis DAAO (20). Benzamide is a potent competitive inhibitor of mammalian DAAO. Binding of this ligand strengthens the apoenzyme-flavin interaction and increases the conformational stability of the porcine enzyme.

MAO-A and MAO-B catalyze the oxidative deamination of aromatic amines to the corresponding aldehydes. Two active-site tyrosyl residues function in both isoforms as an “atomic cage” (23). The kinetic mechanism of MAO is similar to that of DAAO (Fig. 4). The active sites of either MAO-A or MAO-B do not contain basic residues that could possibly function as a proton acceptor in the reductive half-reaction. However, the bent conformation of the isoalloxazine ring of the flavin might provide a clue as to how proton abstraction might occur (23). The strained conformation of the isoalloxazine moiety results in a higher electron density at N5 and lowered electron density at the C4a position of the flavin ring. This conformation facilitates the nucleophilic attack of the basic substrate amine lone pair on the C4a position of the cofactor that results in a flavin-substrate adduct which would be isoelectronic with the reduced flavin ring (Fig. 2b). The N5 of the reduced flavin could constitute the strong base required to abstract the proton from the benzyl carbon of benzylamine substrates, and the “atomic cage” might polarize the amine moiety of the substrate to make it more nucleophilic in accord with the proposed mechanism (23).

Nitroalkane oxidase (NAO; EC 1.7.3.11) from Fusarium oxysporum catalyzes the oxidation of neutral nitroalkanes to the corresponding aldehydes or ketones with the production of nitrile and hydrogen peroxide (24). The enzyme is evolutionarily related to the acyl-CoA dehydrogenases and acyl-CoA oxidases, but it contains an asparagine (Asp402) instead of a glutamate as the catalytic base involved in water attack to the other site of the substrate-binding pocket (26).

Monooxygenases

Flavoprotein monooxygenases mainly use NADPH as electron donor and insert one atom of molecular oxygen into their substrates. Oxygen activation of flavoprotein monooxygenases involves the (transient) stabilization of a flavin C4a-(hydro)peroxide. This species performs either a nucleophilic or electrophilic attack on the substrate (Fig. 5). Oxygenation reactions catalyzed by flavoprotein monooxygenases include hydroxylation, epoxidations, Baeyer-Villiger oxidations, and sulfoxidations (43).

Porcine PHBH is the prototype of the flavoprotein aromatic hydroxylases. Each subunit of this dimeric enzyme contains two active sites which, during catalysis, are alternately visited by the isoalloxazine ring of the FAD cofactor (31). Catalysis is initiated by reduction of the flavin in the exterior active site. The reduced flavin then moves to the interior active site where the reactions with oxygen occur. A similar conformational flexibility of the FAD cofactor has been observed in the crystal structures of phenol hydroxylase (EC 1.14.13.7) and 3-hydroxybenzoate 4-hydroxylase (EC 1.14.13.23). PHBH obeys the following kinetic mechanism:

\[
\begin{align*}
\text{ESFAD} + S &\rightarrow \text{ESFAD}S \\
\text{ESFAD}S + \text{NADPH} &\rightarrow \text{ESFAD}SH + \text{NADP}^+ \\
\text{ESFAD}SH + \text{O}_2 + \text{H}^+ &\rightarrow \text{ESFADHOH} \\
\text{ESFADHOH} + \text{EPFAD} &\rightarrow \text{EPFADHOOH} \\
\text{EPFADHOOH} + \text{EFAD} &\rightarrow \text{EFAD} + \text{P} + \text{H}_2\text{O}
\end{align*}
\]

Binding of 4-hydroxybenzoate (S) in the phenolase form facilitates flavin reduction by NADPH. After NADP+ release, the flavin hydroquinone reacts with molecular oxygen to yield the flavin C4a-hydroperoxide oxygenation species. Protonation of the distal oxygen of the peroxiflavin facilitates the electrophilic attack on the nucleophilic carbon center of the substrate phenolate. After monooxygenase activity, the resulting hydroxyflavin is
decomposed, and the 3,4-dihydroxybenzoate product (P) is released (Fig. 6). Studies from site-directed mutants have provided many insights into the process of substrate hydroxylation (32). In general, flavoprotein aromatic hydroxylases display a narrow substrate specificity and are very regioselective.

Baeyer-Villiger monooxygenases are another class of flavoprotein monooxygenases. These enzymes typically depend on NADPH as an electron donor and catalyze a relatively broad range of asymmetric oxygenation reactions with high enantioslectivity or enantiotoposelectivity (35). The kinetic mechanism of Baeyer-Villiger monooxygenases differs from that of the aromatic hydroxylases:

Here, NADP+ stays bound throughout the entire reaction cycle. Furthermore, Baeyer-Villiger monooxygenases usually promote the deprotonation of the flavin C4a-peroxide (Fig. 6), which thereby facilitates nucleophilic substitution reactions.

**Chemical Tools and Techniques**

Understanding the action mechanism of flavoenzymes heavily relies on the combination of different chemical tools and techniques. First, it is of utmost importance to have a pure and stable (recombinant) protein. Size exclusion chromatography will provide information about the enzyme quaternary structure, and mass spectrometry can establish posttranslational modifications. For a detailed insight into the protein structure, well-diffacting crystals are needed to determine the X-ray structure. The properties of the flavin are exploited for flavoenzyme characterization. Absorbance spectroscopy (under anaerobic, as needed) can be used for binding studies, redox titrations, and rapid-reaction kinetics (Note the relevant articles in the “See Also” section). Other useful techniques include fluorescence spectroscopy (at equilibrium or time-resolved, although in most cases the flavin fluorescence is quenched in the holoenzyme), EPR, and NMR. Several of these techniques have been developed early on (also) thanks to research on flavoproteins. Preparation of the apoflavoprotein (1) and its reconstitution with 13C- and 15N-enriched flavins and subsequent NMR analysis yields information about the n-electron density of the atoms of the isoalloxazine ring in the different redox states (47). Reconstitution with chemically modified ("artificial") flavins can be of help in substrate structure-activity relationship studies (23) and provide information about the solvent accessibility of the active site (48).
Site-directed mutagenesis can be used to establish the function of individual amino acid residues. This method also allows to introduce chemical probes (e.g., fluorophores) that may give insight into the dynamic features and folding properties of a flavoenzyme and its interaction with other protein partners (33, 34). In this context, it is important to stress that catalytically relevant conformational changes of flavoenzymes have been demonstrated that actually exploit the flavin itself as an excellent intrinsic spectroscopic probe (18, 31).

Bioinformatic tools are of increasing importance for the characterization of flavoenzymes. This finding holds for protein sequence and protein structural analysis as well as for gaining insight into the reactivity of the flavin cofactor by combined quantum mechanical and molecular mechanical (QM/MM) simulations (13).

Practical Applications and Future Research Directions

The available knowledge about the structural and mechanistic properties of flavoenzymes is extremely valuable for the discovery and characterization of new flavoenzymes and for the development of practical applications. Several flavoenzyme-inspired applications already exist but many more can be foreseen. Glucose oxidase (EC 1.1.3.4) (27) and cholesterol oxidase (EC 1.1.3.6) (28) are widely applied in diagnostics. Many flavoprotein oxidases and monoxygenases serve as biocatalysts for the production of fine chemicals and pharmaceuticals (41). DAAO is used for the enzymatic synthesis of the cephalosporin precursor 7-aminocephalosporanic acid and in gene therapy for tumor treatment (20). Other flavoenzymes, like e.g., MAO (22) and TrxR (15) serve as drug targets.

Flavoenzymes may be ideal model systems for the development of new tools to be made available for the study of enzymes in general. One example may be the use of cholesterol oxidase for single molecule enzymology, which exploits the flavin fluorescence changes during the catalytic cycle (29). Sophisticated computational tools will facilitate the prediction of novel flavoenzyme functions from sequence, provide a better insight into enzyme and ligand dynamics, and improve methods for flavin-dependent biocatalyst design (41). Another challenge is to use the tools of chemical biology to study the functional properties of flavoenzymes in their natural environment, the living cell. In this article, only a selected group of "simple" flavoenzymes has been discussed. However, one should remember the "complex" flavoenzymes, and the flavin-dependent photoreceptors, and the emerging group of flavoenzymes that do not catalyze redox reactions (see Further Reading List).

References

2. Reeves M, Mcclintin WS, Scrutton NS. Covalent attachment of flavin adenosine diphosphonucleotide (FAD) and flavin mononucleotide.
During the entire reaction cycle, flavoprotein monooxygenases play a crucial role. With Baeyer-Villiger monooxygenases (nucleophilic oxygenation), NADP+ serves as the electron donor, enabling the oxygenation of various substrates.

Figure 6. General mechanism for flavoprotein monooxygenases. With Baeyer-Villiger monooxygenases (nucleophilic oxygenation), NADP+ stays bound during the entire reaction cycle.


45. Matterni A. To be or not to be an oxidase: challenging the oxygen reactivity of flavoenzymes. Trends Biochem. Sci. 2006;31:276–283.


Further Reading


See Also
NAO(P) Dependent Dehydrogenases
Oxygen-Activating Enzymes, Chemistry of Transient State Enzyme Kinetics
Membrane proteins constitute a significant fraction of the proteins that are encoded in the typical genome, and they are critical to many cellular processes, which include transport, cell signaling, and energy transduction. This importance is underscored by the fact that membrane proteins represent the major class of protein targets for pharmaceuticals that are currently in use. However, despite their importance, there are relatively few high-resolution models for membrane proteins, and little is known about their molecular function. The lack of information on membrane proteins in part reflects the difficulties in efficient expression of this class of proteins and difficulties with structural approaches, such as high-resolution nuclear magnetic resonance (NMR), that are not well suited to membrane proteins.

In the field of structural biology, membrane proteins represent a significant challenge, and in recent years new structural and biochemical tools have been developed to characterize these systems. In this review, we discuss the functions and properties of membrane proteins, and we discuss the approaches and tools that yield new information on their structure and molecular function.

Membranes define the boundaries and compartments that make up the cellular matrix, and they are responsible for defining the chemical environment within the cell. Membranes also provide an interface that facilitates protein–protein interactions and other biochemical reactions necessary for cell signaling and trafficking. It is estimated that protein domains that reversibly associate with membrane interfaces represent the most abundant domains found in water-soluble proteins (1). Although the number of proteins that interact with or function at the membrane interface is large, this review will focus on membrane proteins that are sometimes termed integral membrane proteins. Integral membrane proteins are generally defined as proteins that cannot be isolated and purified without first dissolving the bilayer structure, typically with detergents. When compared with water-soluble proteins, membrane proteins have unique properties and present unique challenges in terms of their expression, isolation, and structural characterization. Inducing membrane protein expression in high yield can be more challenging than that of water-soluble proteins. Because the forces that dominate the fold of a membrane protein differ from those that stabilize water-soluble protein folds, purified membrane proteins must be studied in heterogeneous environments, such as membrane mimetic detergent micelles or reconstituted lipid bilayers. It is estimated that approximately 30% of the genome codes for membrane proteins (2), although the exact percentage is unknown. However, when compared with the total number of entries in the protein databank, membrane proteins represent less than 1% of the total number. Although X-ray crystallography remains the single most important method for generating structures of membrane proteins, several new spectroscopic methods are being employed to study membrane protein structure and dynamics.
Membrane Proteins, Properties of Biological Functions and Distribution of Membrane Proteins

Classification of membrane proteins: nomenclature and architecture

Membrane proteins carry out a wide range of critical functions in cells, and they include passive and active transporters, ion channels, many classes of receptors, cellular toxins, proteins involved in membrane trafficking, and the enzymes that facilitate electron transport and oxidative phosphorylation. For example, the voltage-gated ion channels that facilitate the passive diffusion of sodium and potassium across the axonal membrane are responsible for the formation of an action potential. Active transport proteins establish ion gradients and are necessary for the uptake of nutrients into cells. Soluble hormones bind to membrane receptors, which then regulate the internal biochemistry of the cell.

At the present time, representative structures exist for approximately 21 unique β-barrel membrane protein families and 35 polytopic α-helical protein families. This sample is a small fraction of the predicted 300–500 α-helical folds and 700–1700 families (3). Although the structural biology of membrane proteins is in its infancy, it is clear that membrane proteins display a rich variety of structures that vary greatly in size and topology (Fig. 1). Of the structures observed thus far, all are based on two fundamental architectures: the α-helical bundle (4) and the β-barrel (5, 6).

The membrane β-barrel fold is unique to the outer membrane of mitochondria, chloroplasts, and Gram-negative bacteria, and they are found to be composed of an even number of β-strands that vary in number between 8 and 22. The inter-strand hydrogen-bonding pattern within the barrel allows structures to form, which do not have unsatisfied hydrogen bonds within the membrane interior. Frequently, these strands are configured so that amino acid side chains with an aliphatic composition reside on the barrel exterior facing the membrane hydrocarbon, and as observed in membrane proteins based on helical bundles, β-barrel proteins display more aromatic side chains at regions near the membrane bilayer interface. Membrane proteins based on transmembrane α-helices (Fig. 1) are typically localized to the plasma membrane, organelle membranes, and inner membrane of mitochondria and bacteria. Thus, these proteins not only differ in secondary structure, but also they differ in localization. Helical membrane proteins can be formed from 1 to 19 transmembrane segments. When they possess a single transmembrane pass, they are sometimes referred to as either monotopic or bitopic. When these proteins have two or more transmembrane helices, they are referred to as polytopic. The transmembrane helices of a polytopic membrane protein associate into a bundle, and to maintain unsatisfied hydrogen bonds to a minimum, these helices are usually regular. Although helical membrane proteins may be quite flexible and dynamic, elements of helical structure within the bilayer are thought to be rigid.

Membrane protein folding and membrane insertion

Integral membrane proteins must be stable and function in a unique and highly anisotropic environment. The aqueous facing...
domains of a membrane protein experience a very different environment than do the membrane facing regions of the protein, which must be stable in a low-dielectric region devoid of water (7). Because of the absence of a bulk aqueous phase, significant numbers of unsatisfied hydrogen bonds are highly unfavorable, and the folds of β-barrel or helical membrane proteins, as indicated above, must always be arranged to satisfy backbone hydrogen bonding.

In water-soluble proteins, the hydrophobic effect is believed to be a major force that drives the folding of proteins, dominating over van der Waals forces (8, 9). However, in membrane proteins, the hydrophobic effect is relatively unimportant in comparison to side-chain hydrogen bonding and van der Waals forces. Van der Waals forces in a helical bundle would be maximized by complementarity of the interacting surfaces on transmembrane helices; as expected, sequence motifs, such as GxGxG, have been identified that maximize the packing of membrane helices (10–12). Interhelix hydrogen bonding is also important in driving the association of helices, and it may be an important determinant of helix association during membrane protein synthesis (13). Although both interactions play important roles, a mutational study on one helical bundle suggests that van der Waals forces make the largest contribution to the stability of the bundle (14).

The arrangement of the transmembrane segments of a membrane protein and the orientation of the protein C and N termini are determined by two main features of a membrane protein sequence: the stretch of hydrophobic residues that ultimately span the bilayer, and the position of positively charged segments (15, 16). The hydrophobic residues (Ala, Ile, Leu, Val) have the highest frequency in helical regions that lie in the center of the bilayer, whereas the two aromatic residues Tyr and Trp are most abundant near the bilayer interface (17). The balance of charge on either side of a transmembrane segment determines the orientation of the segment, so that the more positively charged portions of a membrane protein sequence are found on the cytoplasmic side of the bilayer. This structure is sometimes referred to as the “positive-inside rule” (18). By altering hydrophobicity and charge, differing topologies can be generated, and it is even possible to find homologous proteins that have naturally evolved different topologies by modifying these features (19). Biosynthetically, the insertion of the growing polypeptide chain into the bilayer is mediated by a translocon, which is a helical membrane protein that is believed to have a lateral gate that opens to the bilayer interior. In the rough endoplasmic reticulum, this protein is Sec61. It is generally thought that the growing polypeptide chain may sample the surrounding bilayer throughout this stage, so that a thermodynamic equilibrium is established with the surrounding lipid (16, 19).

The lipid bilayer is not passive in determining membrane protein activity and function, and an accumulating body of evidence indicates that there is a coupling of membrane proteins to lipid bilayer properties. These properties include the effect of bilayer curvature strain (20), the role of specific lipids such as phosphoinositides, (21) and the effect of thickness on membrane protein function (22). The lipid composition, as well as the bilayer properties that result from this composition, act as allosteric regulators of membrane protein function.

**Molecular function of membrane proteins**

Structural methods and other genetic and biochemical studies have provided clues to the molecular function of membrane proteins. For example, in the case of the bacterial K^+ channel KcsA, the protein is designed to lower the energy for a potassium ion as it passes through the center of the channel. This function is facilitated both by a central cavity that is hydrophobic and by a helix that is positioned toward the channel pore so that its negatively charged end points toward this cavity (23). In the case of visual rhodopsin, a salt bridge that exists between Lys296 and Glu133 functions to maintain the protein in an inactive state. Disruption of this ionic interaction is responsible for the movement of helix 6, which activates the G-protein transducin (24), and mutations that disrupt this ionic interaction are responsible for the retinal diseases retinitis pigmentosa and congenital night blindness (25). Among outer membrane bacterial transporters, TonB-dependent transporters function to move rare nutrients, such as iron chelates, into the cell. They contain a large N-terminal domain of approximately 150 residues that is sometimes referred to as a “hatch.” The available high-resolution structures suggest that substantial rearrangements or an unfolding of this hatch domain takes place to allow the passage of substrates (26).

**Expression, Isolation and Purification of Membrane Proteins**

Difficulties in obtaining protein samples

In general, membrane proteins present challenges at every step on the way to structural determination. The synthesis and processing of these proteins is complex and often involves specific folding factors or chaperones (see References 27 and 28 for reviews). Expression in bacteria is favored because of the low cost to grow large culture volumes by fermentation, the potential for high yields (up to 100 mg protein per liter of culture), the very fast growth rate, and the simplicity and flexibility of expression systems. However, intrinsic differences in how proteins are processed often prevent the expression of adequate amounts of protein with the proper fold, and it may not be possible to isolate and purify sufficient quantities of a membrane protein to a homogenous state. As a result, understanding the processing pathways for the specific protein of interest can be the important factor to achieve successful expression.

The biosynthetic membrane insertion of eukaryotic membrane proteins almost always takes place cotranslationally, whereas many prokaryotic proteins can be posttranslationally inserted, typically with the aid of chaperones. In eukaryotic cells, the proofreading mechanisms in the endoplasmic reticulum prevent misfolded proteins from leaving to the Golgi. Some of these mechanisms include glycosylation of the protein on
its luminal domain (29). In addition, other protective mechanisms such as the yeast unfolded protein response may lead to increased degradation rates of improperly folded proteins (30).

Some proteins require specific types of processing. For example, β-barrel proteins in the outer membrane of Gram-negative bacteria must pass through the cytoplasmic membrane in a linear fashion before being assembled in the outer membrane (for a review, see Reference 31). Processing of β-barrel proteins in the outer membranes of mitochondria and chloroplasts may also involve passage of the protein through the organelle’s outer membrane before insertion, but this process is not well understood (32). Several chaperones have been identified that help the assembly of mitochondrial and chloroplast membrane proteins. Some of these proteins are encoded in the nucleus and are imported to the proper compartment posttranslationally, whereas a few are encoded on the organelle’s own chromosome and processed from within. For these reasons, structural approaches for such proteins always rely on isolation from the native organelle.

Bacterial expression systems are often not an option for eukaryotic membrane proteins. Even if the protein is found embedded in the bacterial membrane, obtaining correctly folded protein is always a concern. Frequently, most or all expressed protein is found in an intracellular aggregate or inclusion body, and only a few examples have been refolded to their native state (33). In many cases, the apparent toxicity of the expressed membrane protein blocks expression as well as growth. Several reasons can explain protein toxicity (see Reference 34 for a review). For example, Miroux and Walker (35) suggest that the overproduction of a single mRNA in a typically used T7 polymerase-driven transcription system resulted in the uncoupling of transcription and translation. A strain, such as C43(DE3) (Lucigen, Middleton, WI) selected to avoid the effects of toxicity, demonstrated better yields of proteins despite a slower rate of synthesis. The requirement of a slower rate of synthesis suggests that cellular machinery, which is not overexpressed, may be required for proper folding. On the other hand, mixed results are observed in providing the cell with extra cytosolic proteins. Interestingly, one of the more productive strategies is to use seleno-methionine or selenium-free media instead of selenium-containing media. Although this scenario is more likely, it is certainly not true in all cases. As mentioned above, other factors may be necessary for proper folding. Furthermore, a protein that is part of a complex may not express in a stable or properly folded manner in the absence of its partners or assembly factors.

Although the optimization of growth conditions for eukaryotic cells is limited, the growth conditions for bacterial systems can be varied widely, and they have a large influence on the expression of the protein target. Many “tricks” investigators use are anecdotal and may be specific to their protein of interest. Several publications detail the approaches and often many different conditions must be tested (see References 33 and 38 for extensive reviews). We will not attempt a thorough listing of such methods, but a few are worth noting. First, medium strength promoters are used to slow down synthesis rate. As mentioned above, the cell may respond to an overwhelming synthesis of a specific protein by activating degradation systems. Second, a lower culture temperature is used, as low as 15 °C, with expression induced for many hours to days. Third, using a defined minimal medium may also result in accumulation of considerable amounts of protein, as the cells grow much more slowly. Development of expression protocols in defined medium is always advantageous for structural approaches because the investigator hopefully will need to generate selenium-methionine derivatives for NMR. We will not attempt a thorough listing of such methods, but a few are worth noting. First, medium strength promoters are used to slow down synthesis rate. As mentioned above, the cell may respond to an overwhelming synthesis of a specific protein by activating degradation systems. Second, a lower culture temperature is used, as low as 15 °C, with expression induced for many hours to days. Third, using a defined minimal medium may also result in accumulation of considerable amounts of protein, as the cells grow much more slowly. Development of expression protocols in defined medium is always advantageous for structural approaches because the investigator hopefully will need to generate selenium-methionine derivatives for NMR. We will not attempt a thorough listing of such methods, but a few are worth noting. First, medium strength promoters are used to slow down synthesis rate. As mentioned above, the cell may respond to an overwhelming synthesis of a specific protein by activating degradation systems. Second, a lower culture temperature is used, as low as 15 °C, with expression induced for many hours to days. Third, using a defined minimal medium may also result in accumulation of considerable amounts of protein, as the cells grow much more slowly. Development of expression protocols in defined medium is always advantageous for structural approaches because the investigator hopefully will need to generate selenium-methionine derivatives for NMR.
Membrane Proteins, Properties of

Furthermore, the homogeneity of the protein can be evaluated by gel filtration. Light scattering, or sedimentation velocity (39–41). Figure 2

To characterize the protein preparation, the critical parameters are purity, which is estimated by SDS-PAGE, the fold and evaluated by secondary structure content via circular dichroism spectroscopy, as well as the oligomeric state determined by gel filtration, light scattering, or sedimentation velocity (39–41). Furthermore, the homogeneity of the protein can be evaluated by mass spectrometry. Detergents will normally interfere with mass spectrometry, but MALDI TOF approaches have been developed to avoid such problems (42). As discussed in the next section, a range of detergents must be explored for protein structural determination, and as a result, the characteristics of the protein in different detergents must be determined (see Fig. 2).

Optimization of Expression
t-vector
-cell line
-temperature
-co-expression with folding factors

Screen Extraction Detergent

Protein Purification
-activity
-affinity
-gel-filtration
-ion exchange

Screen Detergent
-solubility
-activity
-oligomeric state
-fold

Structure Determination

Figure 2: A flow chart that outlines the general strategies in preparing membrane proteins for structural studies.

If possible, the function of the protein should be assessed. Unfortunately, detergent solubilization can often make it difficult to assay protein activity, as in the case of an ion channel; however, other parameters may be assessed in such instances. Ligand binding affinity, enzymatic activity, or association with another protein may provide a convincing assessment of protein function. Other biochemical techniques will also indicate folding and homogeneity, such as limited protease digestion, reactivity with reagents such as sulfhydryl reactive compounds, or chemical cross-linking patterns. Spectroscopic approaches have also been used. The dynamics of specific regions on the protein can be evaluated by nitroxide spin labeling and electron paramagnetic spectroscopy. If the protein can be metabolically labeled with 7H-15N, then a relatively quick heteronuclear single quantum coherence NMR experiment can provide critical information on the homogeneity and fold of the protein (39).

Structural Characterization of Membrane Proteins

X-ray crystallography

High-resolution membrane protein structures are determined predominantly using X-ray crystallography. Although notoriously difficult to crystallize, several methods have been applied successfully to crystallize membrane proteins either by manipulating the detergent/lipid components or by altering the protein component. Most membrane protein structures have been determined using detergent solubilized protein in which the entire protein detergent complex is crystallized. Often, the best-quality crystals of a membrane protein may be obtained only in one or a few detergents, and extensive screening based on detergent properties is required (42).

Thirty-three detergents have been used to crystallize membrane proteins, and three of those detergents have been used to determine three NMR structures (43). Some detergents are fold-specific. For example, C8E4 is predominantly used for β-barrel proteins; whereas DDM has been mostly successful for α-helical membrane proteins. The four detergents that have been used to crystallize most membrane proteins are octyl-glucoside, lauryl dimethyl amine oxide, C8E4, and dodecyl maltoside. These four detergents vary in aliphatic chain length and shape and size of the head group; however, they are all neutral. This commonality is very significant. For three-dimensional protein crystals to form, protein molecules need to contact each other to form crystal contacts that are essential to propagate the lattice. It is likely that charged or even zwitterionic repulsive forces would hinder the association of the protein detergent complexes, which is a process that must occur at early stages of crystal nucleation.

Membrane protein crystals have significantly more solvent (64%) content than soluble proteins [47% (44)] presumably because of the detergent in the crystal. The organization of the detergent in the membrane protein crystal has been investigated in a select few cases and is different in each case. In the LH2 crystal, the detergent forms a belt around the hydrophobic surface of the protein consistent with the dimension of the OG...
detergent molecule (45). Similar to LH2, the OG detergents form a belt around the hydrophobic surface of phospholipase A in the crystal; however, the belts fuse to form a continuous three-dimensional network throughout the crystal (46). The continuous density of detergent was also observed in crystals of porin and two photoreaction centers. Snijder et al. (44) suggest the possibility that organic amphiphile additives in the crystal screen could facilitate this fusion; however, no experimental data correlate the detergent structure observed in the crystal and the physical properties of the detergent/amphiphile mixed micelle. In addition to the detergent used for solubilizing the membrane protein, the native (or synthetic) lipid concentration may have profound effects on diffraction quality as in the micelle. In addition to the detergent used for solubilizing the channel KcsA (65). NMR structure determination of sequential chemical shift assignment achieved for two polytopic strides with the first step to NMR structure determination, the remains a challenge. Several research groups are making great advances, Oberai et al. (3) estimate that if no acceleration of membrane protein structure determination occurs, then it will take more than three decades to determine at least one structural representative of 90% of the α-helical membrane protein sequence families (3).

Solution nuclear magnetic resonance spectroscopy

Although solution NMR techniques do not require crystallization, a molecular weight limit does exist (32 kD is the largest to date (59)), and optimization of detergent conditions has proven to be difficult (60). Several NMR structures of β-barrel outer membrane proteins have been determined (59, 61–63); however, a solution structure of a polytopic α-helical membrane protein remains a challenge. Several research groups are making great strides with the first step to NMR structure determination, the sequential chemical shift assignment achieved for two polytopic membrane proteins (44) and the potassium channel KcsA (65). NMR structure determination of β-barrel proteins has been more successful than α-helical membrane proteins primarily because the nuclear Overhauser effects (NOEs) between amide protons are across strands (i.e., between secondary elements), rather than within the secondary element as in α-helices, which provides valuable structural constraints. The lack of NOE data is currently being overcome with measurements from paramagnetic relaxation (66) and residual dipolar coupling experiments (67).

In addition to the limited distance restraints, the preparation of membrane protein samples has proven to be a major challenge to high-resolution NMR. Similar to X-ray crystallography, the selection of detergent strongly influences the quality of NMR spectra. No single detergent is well suited for NMR studies of membrane proteins; solubility, dynamics, the hydrophilic surface area of the protein, and other physical properties differ for each protein detergent complex, and the proper combination still needs to be determined empirically through extensive screening (39, 68).

Beyond structure determination, solution NMR can be used to investigate backbone dynamics and protein–ligand interactions. Bax and colleagues (65, 69) have characterized backbone dynamics (69) and ion binding affinity of the tetrameric KcsA potassium channel (65). These studies added additional structural insights to the crystal structure. On the ps–ns time scale, the selectivity filter is not dynamic. In SDS micelles, the intra–cellular C-terminal α-helix is dynamic on the ns–ps timescale and does not associate into a tetrameric bundle. In addition to determining the PagP NMR solution structure (62), Hwang et al. (70) characterized a two state dynamic rearrangement in which the more flexible state facilitates the entry of the substrate into the central cavity of the β-barrel.

Site-directed spin labeling

Site-directed spin labeling (SDSL) is used to investigate membrane protein structure and dynamics in lipid bilayers as well as in detergents. In SDSL, a nitroxide probe is introduced to a unique site within a protein. In most cases, a cysteine residue is introduced and subsequently reacted with a sulfhydryl-reactive nitroxide reagent. The resulting nitroxide side chain is sensitive as in detergents. In SDSL, a nitroxide probe is introduced to a unique site within a protein. In most cases, a cysteine residue is introduced and subsequently reacted with a sulfhydryl-reactive nitroxide reagent. The resulting nitroxide side chain is sensitive to the molecular environment, which allows the determination of secondary and tertiary structure (71), conformational dynamics (72), and site-specific dynamics (73). Unlike solution NMR, the technique does not have a molecular weight limit, and membrane proteins can be investigated in detergent solutions or lipid bilayers.

From the electron paramagnetic resonance (EPR) spectrum of the nitroxide side chain, four primary parameters are obtained: 1) solvent accessibility, 2) mobility of the R1 side chain, 3) a polarity index for its immediate environment, and 4) the distance between R1 and another paramagnetic center in the protein. Solvent accessibility of the side chain is determined from the collision frequency of the nitroxide with paramagnetic reagents in solution. The mobility, polarity, and distances are deduced from the EPR spectral line shape. For regular secondary structures, accessibility, mobility, and polarity are periodic functions of sequence position. The period and phase of the function reveal the type of secondary structure and its orientation within the protein, respectively (71, 74). In the case of membrane proteins, the topography of the secondary structure with respect to the membrane surface can also be described (75, 76).

When a pair of spin labels is incorporated into a protein or a protein complex, the dipolar interactions between labels
Membrane Proteins, Properties of

can be used to measure the distance and distance distribution between labels (77). If the labels are separated by 7 to 20 Å, then the dipolar interactions are sufficiently strong that they can be observed and quantified by continuous wave (CW) EPR spectroscopy (78–81). If the labels are separated by a distance greater than 20 Å, then the resulting weak dipolar interactions can be measured with newer pulse methods such as double electron-electron resonance (DEER) or double quantum coherence (82–84). These methods have been used to measure internitroxide distances and distance distributions out to 60 Å or more. An advantage of the CW EPR measurements is that they can be performed at room temperature (78), whereas the pulse measurements, such as DEER, require that the samples be frozen, typically at liquid nitrogen temperatures or lower.

Changes in any of the SDSL parameters measured can be used to detect changes in protein conformations, and most importantly, the data can be interpreted in terms of helix rigid body motions, relative domain movement, and changes in secondary structure. Recently, SDSL distance measurements were used to map ligand-induced conformational changes in BtuB (85), LacY (86), and the NhaA antiporter (87). Figure 3 shows an example of the use of SDSL and DEER to determine structural changes that accompany ligand binding to the extracellular loops in the outer membrane transporter, BtuB (88).

Solid-state nuclear magnetic spectroscopy

Solid-state NMR also can be applied to membrane proteins in lipid bilayers, and recent advancements in magic angle spinning solid-state NMR show promise for structure determination. Although the structures of small crystalline proteins (89) and membrane bound peptides (90) have been determined, the structure of a polytopic membrane protein has yet to be reported. The major necessity that is required to push the technique forward is the de novo sequential chemical shift assignment of the amino acid residues, and in the last few years, several groups have reported successful strategies (91, 92). Beyond structure determination, solid-state NMR has been used to investigate the structures of membrane bound ligands. Several recent examples are a low-resolution structure of neuretinin bound to its G protein coupled receptor (93), scorpion toxin bound to a chimeric potassium channel (94), rhodopsin and the metarhodopsin II intermediate (95), and acetylcholine bound to its receptor (96). These studies have great potential for designing and optimizing drugs-targeting membrane proteins (97).

Concluding Remarks

Several biophysical tools can provide information on the structure, dynamics, and conformational changes of membrane proteins. Many of these methods are beyond the scope of this review and were not mentioned here. However, each of these approaches has strengths and weaknesses; to investigate membrane protein structure and functional dynamics fully, a multitude of techniques is required. Static high-resolution structures are highly informative and provide a good starting point to generate hypotheses; however, they do not provide a complete understanding of protein molecular function.

References

Membrane Proteins, Properties of


Membrane Proteins, Properties of

46. Snijder HJ, Timmins R, Kalik KH, Djikstra BW. Detergent organ-
isation in crystals of monomeric outer membrane phospholipase A.

47. Guan L, Smirnova IN, Yeager M, Nissen P. Dynamic change in the
structure of bacterial outer membrane proteins visualised at atomic

detergent-lipid complex for crystallization of integral membrane

49. Liang B, Tamm LK. A defined protein-detergent-lipid complex for

50. Wang DN. Three-dimensional crystallization of the Escherichia
coli glycerol-3-phosphate transporter: a major facilitator superfamily.

51. Landa-BE, Rosenthal JH. Lipidic cubic phases: a novel concept

52. Wadsten P, Wohlfahrt-Arnfred T, Snijder A, Katona G, Hubbell WL.
Identifying conformational changes with site-directed spin-labelling.

53. Liu B, Tamm LK. Structure of outer membrane protein A transmembrane
Membrane Proteins, Properties of


Further Reading

Books

Websites
http://pdbm.enzim.hu/.
http://www.mpdb.ul.ie/.

See Also
Ion Channels
Lipid Bilayers, Properties of Membrane Assembly in Living Systems
Membrane Proteins, Properties of Protein Targeting and Transport
NAD(P)-Dependent Dehydrogenases

Robert S. Phillips, University of Georgia, Athens, Georgia

doi: 10.1002/9780470048672.wecb363

NAD(P)H is the major carrier of reducing equivalents in cells. NAD(P)-dependent dehydrogenases and reductases catalyze the reversible transfer of hydrogen equivalents between substrates and NAD(P). These enzymes include alcohol dehydrogenases, 2-hydroxyacid dehydrogenases, aldehyde dehydrogenases, amino acid dehydrogenases, secondary amine dehydrogenases, disulfide oxidoreductases, transhydrogenases, azoreductases, sulfite and nitrite reductases, and phosphite dehydrogenase. NAD(P)-dependent dehydrogenases are structurally diverse, but many contain a Rossmann fold for NAD(P) binding. Usually, alcohol dehydrogenases contain an essential metal ion cofactor that serves as a Lewis acid for substrate activation. However, most dehydrogenases use general acid/base catalysis to activate substrates for reaction. NAD(P)-dependent dehydrogenases have been used in biocatalysis for stereoselective reductions of ketones and imines to give chiral alcohols and amines, respectively. Several important drug targets (e.g., dihydrofolate reductase) are NAD(P)-dependent dehydrogenases.

Properties of NAD(P)

Two different chemical forms of NAD(P) exist, NAD+ and NADP+ (Fig. 1), which differ by the presence of a phosphate ester at C-2' of the adenosine. NAD+ is the abbreviation for nicotinamide adenine dinucleotide, and NADP+ stands for nicotinamide adenine dinucleotide 2'-phosphate. The reduced forms of NAD+ and NADP+ are called NADH and NADPH, respectively (Fig. 1). In this review, NAD and NADP are used to designate either oxidized or reduced forms of the cofactor, whereas NAD+ and NADP+ are used to represent the specific oxidized and reduced forms. NAD is often considered to be used preferentially for oxidative catabolic processes that produce energy, whereas NADP is considered a source of reducing equivalents for biosynthesis. Dehydrogenases typically show a high selectivity for either NAD or NADP. The two molecules are thus in separate pools for cellular processes. The two pools are also separated for the most part in eukaryotes, with NAD in the mitochondria and NADP in the cytoplasm. Of course, exceptions to this generalization exist, and hydrogen equivalents can be transferred between NADH and NADPH by transhydrogenases.

Chemical properties of NAD(P)

NAD(P)H is the main carrier of molecular hydrogen equivalents in biologic systems. The redox potential of NAD(P)H is –340 mV; thus, the transfer of hydrogen from NAD(P)H to substrates is generally thermodynamically favorable and is often the physiologic reaction. In principle, hydrogen transfer from NAD(P)H could occur either by a hydride or by a radical mechanism, but the preponderance of mechanistic data collected to date with dehydrogenases indicates that hydride transfer is the preferred pathway. The two hydrogens at C-4 of the dihydro- nicotinamide ring of NADH and NADPH are diastereotopic and thus stereochemically inequivalent (Fig. 1). Elegant work by Fisher et al. (1) in the 1950s demonstrated that the pro-(R) hydrogen from C-1 of ethanol is transferred to the re-face of the nicotinamide ring of NADH and NADPH are diastereotopic and thus stereochemically inequivalent (Fig. 1). Elegant work by Fisher et al. (1) in the 1950s demonstrated that the pro-(R) hydrogen from C-1 of ethanol is transferred to the re-face of the nicotinamide ring of NAD(P)H.
NAD(P) Dependent Dehydrogenases

Figure 1: Structures of NAD⁺ (top) and NADH (bottom).

C-4 of the nicotinamide ring of NAD⁺ by yeast alcohol dehydrogenase, where it becomes HR of NADH (1), although the transfer was thought incorrectly at the time to take place at C-6 rather than C-4 of the nicotinamide. The correct structure for NADH, called DPNH (for Diphospho Pyridine Nucleotide) at the time, was determined by Pullman et al. in 1954 (2). A particular enzyme will be stereospecific for one or the other of these diastereotopic hydrogens, either HHR or HHS, on the dihydronicotinamide. In older literature, the re-face of the nicotinamide ring is called the "A-face" and the si-face is called the "B-face" of the nicotinamide ring.

Biologic properties of NAD(P)

NAD(P) is synthesized de novo from L-tryptophan in animals (3) and in some bacteria (4) by the kynurenine pathway through quinolinate (pyridine-2,3-dicarboxylate). In plants and in most bacteria, NAD(P) is biosynthesized from quinolinate by a different pathway, which starts from L-aspartate. Quinolinate is then converted to nicotinate mononucleotide by an unusual reaction with phosphoribosylpyrophosphate catalyzed by quinolinate phosphoribosyl transferase, with concomitant decarboxylation of the C-2 carboxylate. Amidation of the nicotinate and reaction with ATP to form the AMP dinucleotide completes the biosynthesis of NAD⁺. Subsequent phosphorylation at the 2'-OH of the adenosyl gives NADP⁺. Nicotinic acid (niacin) can also be converted to NADP⁺ by a salvage pathway in both procaryotes and eucaryotes. In addition to its use as a redox cofactor for dehydrogenases, NADP⁺ is a substrate for other biologically important reactions such as ADP-ribosylations and deacetylation of histones (5).

Reactions Catalyzed by NAD(P)-Dependent Dehydrogenases

NAD(P)-dependent dehydrogenases catalyze the oxidation/reduction of a wide range of organic substrates, which include alcohols, amines, alkanes, and thiols, as well as some inorganic substrates, such as nitrite, hydrogen, and phosphite (Table 1).

Alcohol dehydrogenases

Alddehyde dehydrogenases catalyze the oxidation of aldehydes to carboxylic acids (Eq. 2). Under physiologic conditions, this

Table 1: Reactions performed by NAD(P)-dependent dehydrogenases

<table>
<thead>
<tr>
<th>Reaction</th>
<th>EC number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol dehydrogenases</td>
<td>1.1.1.1 to 1.1.1.292</td>
</tr>
<tr>
<td>Alddehyde dehydrogenases</td>
<td>1.2.1.1 to 1.2.1.72</td>
</tr>
<tr>
<td>CH dehydrogenases and reductases</td>
<td>1.3.1.1 to 1.3.1.79</td>
</tr>
<tr>
<td>Primary amine dehydrogenases and reductases</td>
<td>1.4.1.1 to 1.4.1.21</td>
</tr>
<tr>
<td>Secondary amine dehydrogenases and reductases</td>
<td>1.5.1.1 to 1.5.1.35</td>
</tr>
<tr>
<td>NAD(P)H dehydrogenases</td>
<td>1.6.1.1 to 1.6.1.2</td>
</tr>
<tr>
<td>Other N-dehydrogenases and reductases</td>
<td>1.7.1.1 to 1.7.1.13</td>
</tr>
<tr>
<td>S-dehydrogenases and reductases</td>
<td>1.8.1.1 to 1.8.1.15</td>
</tr>
<tr>
<td>Other miscellaneous dehydrogenases and reductases</td>
<td>1.10.1.x, 1.12.1.x, 1.20.1.x</td>
</tr>
<tr>
<td>NAD(P) dependent enzymes without net redox</td>
<td>various</td>
</tr>
</tbody>
</table>
NAD(P) Dependent Dehydrogenases

reaction is irreversible because of the subsequent ionization of the carboxylic acid.

\[
R \text{H} + \text{NAD(P)H} + \text{H}_2\text{O} \rightarrow R \text{O}^{-} + \text{NAD(P)H} + 2\text{H}^+ \quad (2)
\]

CH dehydrogenases and reductases

CH dehydrogenases form carbon–carbon double bonds by dehydrogenation (Eq. 3). Generally, these enzymes act on CH bonds activated by adjacent carbonyl or similar groups.

\[
R \text{H}_2 \text{O} \rightarrow R \text{O}^{-} + \text{NAD(P)H} + \text{H}^+ \quad \text{R}
\]

Primary amine dehydrogenases

These enzymes oxidize primary amines to form imines (Eq. 4). The most common substrates are alpha-amino acids.

\[
R \text{H} + \text{NAD(P)H} + \text{H}_2\text{O} \rightarrow R \text{O}^{-} + \text{NAD(P)H} + \text{H}^+ \quad \text{R}
\]

Secondary amine dehydrogenases and reductases

These enzymes oxidize secondary amines to imines (Eq. 5).

\[
R \text{H}_2 \text{O} \rightarrow R \text{O}^{-} + \text{NAD(P)H} + \text{H}_2\text{O} \quad \text{R}
\]

NAD(P)H dehydrogenases

These enzymes transfer H(2) from NAD(P)H to NADP+ (Eq. 6).

\[
\text{NAD(P)H} + \text{NAD(P)H}^{-} \rightarrow \text{NADP}^+ + \text{NAD(P)H} \quad (6)
\]

Other N-dehydrogenases and reductases

Most enzymes remove an oxygen from a nitrogen (Eq. 7), either sp(2), as in nitrate reductase, or sp(3), as in hydroxylamine reductase, or reduce an N-N double bond. Because the physiologic direction is reduction, most enzymes are called reductases.

\[
R \text{O}^{-} + \text{H}_2\text{O} + \text{NAD(P)H} \rightarrow R \text{O}^{-} + \text{NAD(P)H} + \text{H}^+ \quad \text{R}
\]

S-dehydrogenases and reductases

These enzymes catalyze the reversible oxidation of sulfur containing compounds. Most enzymes catalyze the interconversion of thiols and disulfides (Eq. 8), but a few enzymes can form or remove sulfur-oxygen or sulfur-carbon bonds.

\[
2R\text{SH} + \text{NAD(P)H} \rightarrow 2R\text{S}^{-} + \text{NAD(P)H} + \text{H}^+ \quad (8)
\]

Other miscellaneous dehydrogenases and reductases

NAD(P)-dependent dehydrogenases act on other types of substrates, such as H(2) (Eq. 9), diphenols, or phosphite.

\[
\text{H}_2 + \text{NAD(P)H} \rightarrow \text{NAD(P)H} + \text{H}^+ \quad (9)
\]

NAD(P)-dependent enzymes without net redox

Some NAD(P)-dependent enzymes require NAD(P) as a co-factor but they do not result in net oxidation or reduction in the product. In some cases, a reversible dehydrogenase reaction exists to form a catalytic intermediate (e.g., S-adenosyl homocysteine hydrolase or UDP-galactose epimerase). In catalase, the reduction occurs of tightly bound NADP+ by NADPH in solution without ligand exchange (6). The NADPH seems to be present to protect catalase against oxidation during turnover.
Structural Features of NAD(P)-Dependent Dehydrogenases

NAD(P)-dependent dehydrogenases exhibit a wide range of structures (Table 2). However, many have a common structural motif, called the Rossmann fold (7). This structure consists of a six-stranded parallel beta-sheet flanked by two alpha helices on each side of the sheet with the nucleotide bound at the COOH terminal end of the beta-sheet. A typical dehydrogenase structure that shows a Rossmann fold is presented in Fig. 2. The dehydrogenases can be grouped according to the presence of the Rossmann fold (Table 2). Most dehydrogenases are specific for either NAD or NAD(P), but not both, as coenzyme. The structural determinants for coenzyme specificity have been examined in many of these systems by site-directed mutagenesis. It has generally been found that an acidic residue exists, either aspartate or glutamate, in NAD-specific enzymes, and a basic residue exists, either lysine or arginine, in NADP-specific enzymes. These structures provide electrostatic interactions either to reduce NADP binding (aspartate, glutamate) or to increase NADP binding (lysine, arginine) (8, 9). In addition, NADP-specific enzymes have hydrogen bond donors like serine or tyrosine that can donate hydrogen bonds to the 2'-phosphate.

Alcohol dehydrogenases

Alcohol dehydrogenases are found in three major superfamilies that have no sequence or structural homology. The short chain dehydrogenases/reductases (SDRs) are the largest superfamily of alcohol dehydrogenases, with over 1600 members known to date (10). The SDRs have about 250 residues and are metal-ion independent. The medium chain dehydrogenases/reductases (MDRs) have about 350 residues and require a metal ion for activity, usually zinc (11). The MDRs form dimers or tetramers, and they contain one or two zinc ions per subunit. One zinc ion is always in the active site, whereas the other ion, if present, is held in a "structural site," remote from the active site. The active site zinc is coordinated by one or more cysteine thiolates, a histidine imidazole, carboxylates from glutamate or aspartate, and a water. Usually, the structural sites coordinate zinc with four cysteine thiolates. The MDRs include the canonical mammalian liver alcohol dehydrogenases. The structure of one subunit of a tetrameric alcohol dehydrogenase from a thermophilic bacterium, Thermoaerobium brockii, is shown in Fig. 2. The NADP⁺ can be seen as the ball and stick model between the Rossmann fold and the catalytic domain, which has the catalytic zinc shown as the large sphere. The third family of alcohol dehydrogenases is the smallest and is composed of ferrous iron-dependent dehydrogenases.

The 2-hydroxyacid dehydrogenases are found in two different superfamilies, which are stereospecific for their substrates. The D-2-hydroxyacid dehydrogenase superfamily (12) is larger and more diverse in reactions catalyzed than the smaller L-malate/lactate dehydrogenase superfamily. Neither of these 2-hydroxyacid dehydrogenases requires metal ions for activity.

Another group of alcohol dehydrogenases is the oxidative dehydrogenases, which include malic enzyme and isocitrate dehydrogenase. Malic enzyme, which catalyzes the reversible oxidation of L-malate to pyruvate and CO₂, is a tetrameric structure.

Table 2. Families of NAD(P)-dependent dehydrogenases

<table>
<thead>
<tr>
<th>Reaction catalyzed</th>
<th>Superfamily</th>
<th>Rossmann clan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol dehydrogenase</td>
<td>Short chain dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Alcohol dehydrogenase</td>
<td>Medium chain dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Alcohol dehydrogenase</td>
<td>Fe-dependent dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Hydroxyacid dehydrogenase</td>
<td>D-2-hydroxyacid dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Hydroxyacid dehydrogenase</td>
<td>L-2-hydroxyacid dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Malic enzyme</td>
<td>Malic enzyme</td>
<td>yes</td>
</tr>
<tr>
<td>Isocitrate dehydrogenase</td>
<td>Isocitrate/soropyrrol malate dehydrogenase</td>
<td>no</td>
</tr>
<tr>
<td>Aldehyde dehydrogenase</td>
<td>Aldehyde dehydrogenase</td>
<td>no</td>
</tr>
<tr>
<td>CH dehydrogenase</td>
<td>various</td>
<td>some</td>
</tr>
<tr>
<td>Amino acid dehydrogenase</td>
<td>Glu/Leu/Val dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Amino acid dehydrogenase</td>
<td>Ala dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Amino acid dehydrogenase</td>
<td>Diamino-pimelate dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Secondary amine dehydrogenase</td>
<td>Dihydrofolate reductase</td>
<td>no</td>
</tr>
<tr>
<td>Secondary amine dehydrogenase</td>
<td>Saccharopine dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Secondary amine dehydrogenase</td>
<td>1-Pyrroline-5-carboxylate reductase</td>
<td>yes</td>
</tr>
<tr>
<td>Transhydrogenase</td>
<td>Ala dehydrogenase</td>
<td>yes</td>
</tr>
<tr>
<td>Azo-reductase</td>
<td>Flavodoxin</td>
<td>no</td>
</tr>
<tr>
<td>IMP dehydrogenase</td>
<td>Triose phosphate isomerase</td>
<td>no</td>
</tr>
<tr>
<td>Disulphide oxidoreductase</td>
<td>Dithiol oxidoreductase</td>
<td>yes</td>
</tr>
<tr>
<td>Sulphite reductase</td>
<td>Ferrodoxin, Sulfitenitrile reductase</td>
<td>no</td>
</tr>
<tr>
<td>NADH dehydrogenase</td>
<td>NADH dehydrogenase</td>
<td>no</td>
</tr>
</tbody>
</table>
NAD(P) Dependent Dehydrogenases

Figure 2. Structure of one subunit of *Thermoanaerobium brockii* secondary alcohol dehydrogenase (RCSB PDB entry 1YKF). The Rossmann fold can be seen as the seven-stranded beta sheet capped with three alpha helices underneath the NADP⁺, shown as a ball and stick model. The catalytic zinc is the large sphere behind the NADP⁺.

Primary amine dehydrogenases

These enzymes are members of the Rossmann fold clan, but are distributed in several superfamilies within the clan. The glutamate/asparagine/asparaginase dehydrogenases form a family (12) that is related structurally and functionally, but alanine dehydrogenase is in a separate family. Another enzyme, diamino propionate dehydrogenase, is a member of the semialdehyde dehydrogenase family.

Secondary amine dehydrogenases

The secondary amine dehydrogenases are a small group of enzymes, most of which are involved in folate metabolism, such as dihydrofolate reductase, which has its own superfamily outside the Rossmann fold clan. Typically, the dihydrofolate reductases are monomers or dimers. The other members of the group, saccharopine dehydrogenase and 3-pyrroline-5-carboxylate reductase, have their own superfamilies within the Rossmann clan. 1-Pyrroline-5-carboxylate reductase has a dimeric catalytic unit but forms multimers as large as decamers.

NAD(P)H transhydrogenases

The transhydrogenases are members of the Rossmann fold clan and show similarity in the NAD(P) binding domain with alanine dehydrogenase. These enzymes are membrane bound and transfer reducing equivalents and protons across the membrane. Three subunits exist, an NADH binding subunit, an NADPH binding subunit, and a proton translocating subunit.

Other N-dehydrogenases and reductases

Another diverse group of enzymes includes the N-dehydrogenases and reductases. The reductases are dimeric flavoproteins that contain a Rossmann fold which transfer reducing equivalents from NAD(P)H. IMP dehydrogenase is a tetrameric enzyme in the GTP biosynthetic pathway that is a member of the triose phosphate isomerase (TIM) clan. Nitrate reductase contains a molybdenum cofactor. Nitrite reductase is a complex structure related to sulfitreductase (see below).

S-dehydrogenases and reductases

Several disulfide oxidoreductases form a separate superfamily within the Rossmann fold clan (16), which includes dihydrolipoamide dehydrogenase, glutathione reductase, mercuric reductase, and thioredoxin reductase. Dihydrolipoamide dehydrogenases are components of large enzyme complexes involved in oxidation of ketoacids, such as pyruvate and 2-oxoglutarate dehydrogenase. They contain FAD and NAD⁺ binding domains and transfer the reducing equivalents formed in the oxidative decarboxylation to NAD⁺. The disulfide reductases are in the same superfamily, and contain FAD, but they reduce disulfides with the consumption of NADPH. Mercapturic reductase, also in this family, is an NADPH-dependent flavoprotein that catalyzes the reduction of tertiary methyl mercapturates to sulfides.

CH dehydrogenases

These enzymes do not fall into a single superfamily, but they have members distributed throughout the other superfamilies of dehydrogenases. Some enzymes contain a catalytic flavin that mediates the transfer of hydrogen between substrate and NAD⁺.

Aldehyde dehydrogenases

Aldehyde dehydrogenases are found in a large superfamily distributed in many bacteria and archa as well as eucaryotes (14). Because of the distinctive mode of NAD(P) binding, the aldehyde dehydrogenases are not included in the Rossmann fold clan. These enzymes include highly substrate-specific members such as glyceraldehyde-3-phosphate dehydrogenase and non-specific aldehyde dehydrogenases for detoxification of toxic aldehydes. They may form either dimers or tetramers.

Primary amine dehydrogenases

These enzymes are members of the Rossmann fold clan, but are distributed in several superfamilies within the clan. The glutamate/asparagine/asparaginase dehydrogenases form a family (12) that is related structurally and functionally, but alanine dehydrogenase is in a separate family. Another enzyme, diamino propionate dehydrogenase, is a member of the semialdehyde dehydrogenase family.

Secondary amine dehydrogenases

The secondary amine dehydrogenases are a small group of enzymes, most of which are involved in folate metabolism, such as dihydrofolate reductase, which has its own superfamily outside the Rossmann fold clan. Typically, the dihydrofolate reductases are monomers or dimers. The other members of the group, saccharopine dehydrogenase and 3-pyrroline-5-carboxylate reductase, have their own superfamilies within the Rossmann clan. 1-Pyrroline-5-carboxylate reductase has a dimeric catalytic unit but forms multimers as large as decamers.

NAD(P)H transhydrogenases

The transhydrogenases are members of the Rossmann fold clan and show similarity in the NAD(P) binding domain with alanine dehydrogenase. These enzymes are membrane bound and transfer reducing equivalents and protons across the membrane. Three subunits exist, an NADH binding subunit, an NADPH binding subunit, and a proton translocating subunit.

Other N-dehydrogenases and reductases

Another diverse group of enzymes includes the N-dehydrogenases and reductases. The reductases are dimeric flavoproteins that contain a Rossmann fold which transfer reducing equivalents from NAD(P)H. IMP dehydrogenase is a tetrameric enzyme in the GTP biosynthetic pathway that is a member of the triose phosphate isomerase (TIM) clan. Nitrate reductase contains a molybdenum cofactor. Nitrite reductase is a complex structure related to sulfitreductase (see below).

S-dehydrogenases and reductases

Several disulfide oxidoreductases form a separate superfamily within the Rossmann fold clan (16), which includes dihydrolipoamide dehydrogenase, glutathione reductase, mercuric reductase, and thioredoxin reductase. Dihydrolipoamide dehydrogenases are components of large enzyme complexes involved in oxidation of ketoacids, such as pyruvate and 2-oxoglutarate dehydrogenase. They contain FAD and NAD⁺ binding domains and transfer the reducing equivalents formed in the oxidative decarboxylation to NAD⁺. The disulfide reductases are in the same superfamily, and contain FAD, but they reduce disulfides with the consumption of NADPH. Mercapturic reductase, also in this family, is an NADPH-dependent flavoprotein that catalyzes
the reduction of Hg₂⁺ to Hg₂⁺. Thioredoxin reductase also contains a catalytically essential selenocysteine residue. The sulfite reductases have a complex architecture, with an α4β4 quaternary structure in which the alpha subunit contains both FAD and FMN and the beta subunit contains an Fe₄S₄ cluster connected to a siroheme. An elemental sulfur reductase also exists from Pyrococcus furiosus, which is a dimeric flavoprotein that uses NADPH to reduce 50% to 72%.5

Other dehydrogenases
The NADH dehydrogenases are components of complex membrane-bound respiratory systems sometimes called Complex 1, found in mitochondria and chloroplasts. The NADH is used by an NADH-ubiquinone reductase subunit, which in turn transfers electrons from the reduced ubiquinone to the electron transport system. A similar enzyme is found associated with a NiFe hydrogenase in bacteria.

Chemical Mechanisms of NAD(P)-Dependent Dehydrogenases
As we have seen in the structural section, dehydrogenases may contain catalytic metal ions, most frequently zinc, but also iron, magnesium, or manganese. These metal ions may play a role as Lewis acids to facilitate the redox chemistry, or they may mediate electron transfer between coenzyme and substrate. Dehydrogenases without catalytic metal ions use general acid/base catalysis and/or other organic cofactors like FAD to facilitate the redox reactions.

Mechanisms of metal ion dependent dehydrogenases
The best-studied metal ion dependent dehydrogenase is horse liver alcohol dehydrogenase (H.LADH). Horse liver has a relatively high content of alcohol dehydrogenase, so it was a convenient source of alcohol dehydrogenase before recombinant technology became available. Two isozymes of horse liver ADH exist, the E-form that reacts rapidly with ethanol and the S-form that reacts faster with steroids. The reaction mechanism of horse liver ADH and related alcohol dehydrogenases is shown in Fig. 3 (17). The substrate oxygen becomes coordinated to the catalytic zinc, either by displacing the water or expansion of the coordination environment of the zinc. The polariza-tion of the substrate oxygen by the zinc-Lewis acid assists in alcohol deprotonation. The conserved serine-48/histidine-51 pair, together with the Zn²⁺ of the ribose of NAD, is believed to play a role in shuttling protons into and out of the active site. Mutation of serine-48 to alanine results in loss of activity, whereas mutation of histidine-51 to glutamine results in an 11-fold decrease in activity (18). Formation of the zinc alkoxide activates it for the hydride transfer to NAD⁺. Structural studies have shown that the dihydronicotinamide ring is puckered in some complexes (19). This finding may increase the reactivity of the nicotinamide for hydride transfer. The hydride transfer in the mechanism of H.LADH and related alcohol dehydrogenases has been studied in detail. Evidence exists to suggest that the hydride transfer proceeds with a significant amount of quantum mechanical tunneling in these enzymes (20).

In the case of the oxidative decarboxylase, malic enzyme, coordination of the substrate to the catalytic divalent metal ion M⁰₂⁺ helps to reduce the alcohol pKₐ so that the catalytic lysine-183 can form the alkoxide to initiate hydride transfer. The divalent metal ion then polarizes the carbonyl of the beta-ketoacid intermediate to facilitate decarboxylation (13).

Mechanisms of dehydrogenases without catalytic metal ions
The 2-hydroxy acid dehydrogenases such as lactate dehydrogenase generally carry a catalytic metal. The lactate dehydrogenase has an essential histidine residue that abstracts the proton from the alcohol concomitant with hydride transfer (21). The carboxylate of the substrate forms an ion pair with an essential active site arginine. The mechanisms of D-lactate dehydrogenase and malate dehydrogenase are very similar.

A dehyde dehydrogenases contain an essential active site cysteine thiol, which forms a hemithioacetal by nucleophilic attack on the aldehyde carbonyl of the substrate. The electron rich hemithioacetal is then activated for hydride transfer from the alkoxide to NAD⁺, giving a covalent thioester acyl intermediate, which undergoes hydrolysis, or phosphorolysis in the case of G3PDH, to provide the carboxylic acid product (22).

Amino acid dehydrogenases, such as glutamate dehydrogenase from bovine liver, use general acid/base catalysis to transfer hydride from an alpha-amino acid to NAD⁺. Glutamate dehydrogenase is one of the few dehydrogenases that use NAD⁺ and NADP⁺ equally. The formation of the free amino group by deprotonation of the ammonium of the amino acid zwitterion activates the substrate and has been shown to occur prior to hydride transfer (23).

Dihydrofolate reductase (DHFR) is the best studied of the dehydrogenases/reductases that act on secondary amines. The rate constants for each individual step in the mechanism have been determined by stopped-flow and rapid-quench kinetic analysis. Dynamics are important in the mechanism of hydride transfer in DHFR, as residues 25 A away from the active site can have a large influence on catalysis (24). As a result of both experimental and computational studies, the hydride transfer in DHFR has been proposed to involve long-range coupled motions driving hydrophobic tunneling.

In the reactions of enzymes with small inorganic substrates, such as sulfite or nitrite reductase, the reducing equivalents are not transferred directly from the NAD(P)H to the substrate. These structurally complex enzymes contain additional redox cofactors like flavins or iron-sulfur clusters, which are reduced by the NAD(P)H and interact with the substrate directly. Since NAD(P)H prefers to transfer a hydride ion with a pair of electrons, rather than stepwise single electron transfer, the flavin is necessary to split the hydride into single electrons. Some enzymes use an NADP⁺ as a cofactor to catalyze reactions without an overall redox change. In the case of
S-adenosylhomocysteine hydrolase, the bound NAD$^+$ converts the 3'-OH of the adenosine to a keto group, which increases the acidity of the CH bond at C-4', allowing for facile beta-elimination of the homocysteine (25). The 3'-ketoadenosine is reduced subsequently to adenosine by the bound NADH before release of product.

NAD(P)-Dependent Dehydrogenases—Applications in Biocatalysis

The ability of dehydrogenases to catalyze stereoselective transformations of carbonyl and related compounds has made them the object of considerable interest as biocatalysts. The major problem in the biocatalytic application of dehydrogenases is the regeneration of the NAD(P)H cofactor. It can be done in two ways. One way is to use a cheap sacrificial substrate to recycle the cofactor using the same enzyme. The other way is to add a second enzyme that uses a different substrate, for example, formate dehydrogenase for NADH, and glucose-6-phosphate dehydrogenase for NADPH. Mutant forms of phosphite dehydrogenase have been developed recently that can be used for recycling of both NADH and NADPH (8).

Alcohol dehydrogenases

The asymmetric reduction of ketones by horse liver ADH to produce optically active secondary alcohols was studied by Prelog (26), and a general rule was formulated, now commonly called “Prelog’s Rule” (Fig. 4). This rule states that reduction will usually occur by hydride transfer to the re-face of the carbonyl, resulting in the (S)-alcohol. Because of the commercial availability of horse liver ADH, the substrate activity and stereoselectivity for a wide range of ketones have been examined (27). However, several problems exist with the use of horse liver ADH in ketone reduction, including low activity with acyclic ketones, low stability, and sensitivity to organic solvents. Later work identified alcohol dehydrogenases with high activity for secondary alcohols and ketones in thermophilic bacteria, particularly *Thermoanaerobium brockii* and *Thermoanaerobacter ethanolicus*. These enzymes have high activity for ketone reduction, high stability, and are active in the presence of high concentrations of organic solvents. The secondary ADH from *Thermoanaerobacter ethanolicus* has been shown to be particularly effective in the asymmetric reduction of ketones.
T. ethanolicus exhibits low enantiospecificity for small alcohols like 2-butanol and 2-pentanol. A strong temperature dependence of the enantiomeric ratio, $E = R/S$, exists, with the preferred configuration for 2-butanol inverting from (S) to (R) as the temperature increases, with a racemic temperature at about 300 K (28). These results show that a large entropic contribution to the stereospecificity occurs, possibly because of differences in active site solvation for the enantiomeric substrates. Wild-type secondary ADH has no significant activity toward aromatic alcohols and ketones, but a mutant form, W110A, has high activity and stereoselectivity for reduction of 4-phenyl-2-butanone and 1-phenyl-2-propanone (29). The ketone substrates of interest are often sparingly soluble in water, which limits the scale of the reaction severity. However, recent studies that use biphasic systems and sol-gel immobilized alcohol dehydrogenases in organic media have shown high activity for ketone reduction, with comparable or increased stereoselectivity (30).

**Hydroxyacid dehydrogenases**

Hydroxyacid dehydrogenases are attractive catalysts for synthesis of 2-hydroxyacids because families are stereospecific for either D- or L-2-hydroxyacids. Despite this, few applications of these enzymes have occurred in synthesis. A microbial 2-hydroxyacid dehydrogenase has been used to prepare (S)-2-hydroxy-4-phenylbutanoic acid and (S)-phenyllactic acid in enantiomerically pure form (31).

**Amino acid dehydrogenases**

Amino acid dehydrogenases have been exploited to synthesize amino acids and analogs by reductive amination of α-ketocarboxylic acids. Glutamate dehydrogenase has been used extensively because of its commercial availability. Several glutamate analogs, such as 3-hydroxy and 4-methylglutamate, have been prepared with glutamate dehydrogenase. In addition, 6-OH-L-norleucine, which is an intermediate in the synthesis of the antihypertensive drug Omapatrilat, was prepared using glutamate dehydrogenase (32). Phenylalanine dehydrogenase has been exploited similarly in the synthesis of ring-substituted L-phenylalanine analogs. Some mutant phenylalanine dehydrogenases were found to be more tolerant of ring substituents than the wild-type enzyme. Leucine dehydrogenase has been used in oxidative resolution of DL-tert-leucine to provide enantiomerically pure D-tert-leucine (33).

**NAD(P)-Dependent Dehydrogenases as Drug Targets**

**Dihydropolate reductase**

Dihydropolate reductase (DHFR) has been the target for anticancer, antibacterial, and antiparasitic drugs because of its role in folate metabolism and in DNA biosynthesis (34). Methotrexate, a potent inhibitor of DHFR, has been used for many years to treat various cancers including leukemia, breast and lung cancers, as well as to treat arthritis and psoriasis. Trimethoprim is a DHFR inhibitor used for treatment of urinary infections, and it is sometimes used with sulfamethoxazole, a sulfonamide. Pyrimethamine with a sulfonamide is also given as an antimalarial. It is also interesting that a covalent adduct of NAD with isoniazid, which is the standard antituberculosis drug, is a potent inhibitor of DHFR in Mycobacterium tuberculosis. However, the main target of isoniazid in M. tuberculosis is the enoyl-acyl carrier protein reductase, which is strongly inhibited by the (4S)-adduct of isoniazid with NAD, whereas DHFR is inhibited by the (4R)-diastereomer (35).

**Inosine monophosphate dehydrogenase**

Inosine monophosphate dehydrogenase (IMPDH), which is an essential enzyme in the purine biosynthetic pathway, has been investigated as a potential target for anticancer drugs and immunosuppressants. A series of NAD mimetic drugs has been examined as inhibitors of IMPDH. Although it might be thought that NAD analogs would be rather non-selective for dehydrogenases in general, it was found that cofactor analogs with high affinity and specificity for IMPDH can be prepared by replacement of the nicotinamide ring with a heterocyclic ring that contains a carboxamide group. Tiazofurin is an analog that contains a thiazole ring, and it has been evaluated for antitumor activity (36). Unfortunately, high toxicity in clinical trials prevented tiazofurin from additional development. Non-nucleoside inhibitors of IMPDH include mycophenolic acid, which is a fungal natural product and is used as an immunosuppressant.

**17β-hydroxysteroid dehydrogenase**

Inhibitors of 17β-hydroxysteroid dehydrogenase (17β-HSD) have been of interest as potential anticancer agents because of the critical role of this enzyme in androgen and estrogen formation (37). Many common cancers, such as breast and prostate, exhibit a steroid dependency. 17β-HSD exists in several isoforms, which have different substrate specificity and tissue distribution. Isoform-specific inhibitors have been developed.
Aldose reductase/sorbitol dehydrogenase

The complications of diabetes include neuropathy, retinopathy, and nephropathy. These complications are caused by accumulation of sorbitol in cells from the reduction of glucose by aldose reductase, which produces osmotic stress on the cell and results in cell death. Thus, inhibitors of aldose reductase have been examined as potential therapeutic agents for diabetes (38). Aldose reductase and a similar enzyme, sorbinol dehydrogenase, which converts sorbitol to fructose, are also increased as a result of ischemia from myocardial infarction; thus, inhibitors of these enzymes may help in the treatment of heart attacks.

References

NAD(P) Dependent Dehydrogenases


Further Reading


To obtain the coordinates for enzymes discussed in this article, go to the RCSB (http://home.rcsb.org/).

More information on enzyme families and superfamilies can be obtained at the PFAM database (http://pfam.sanger.ac.uk/).

See Also

Cofactor Biosynthesis
Common Drug Target
proteins: Computational Analysis of Structure, Function and Stability
Enzyme Cofactors
Enantioselectivity of Enzymes
Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of

Zhihua Liu and Victor J. Hruby, Department of Chemistry, The University of Arizona, Tucson, Arizona
doi: 10.1002/9780470048672.wecb658

How to synthesize biologically active polypeptides chemically has long been an important challenge for many fields in chemistry and biology. In the past, the development of solution-phase and solid-phase synthesis techniques has made it possible to synthesize large molecules like peptides with molecular weights up to 10,000. The purpose of this review is to provide an overview of how to prepare natural bioactive peptides synthetically. Because of space limitations, this review concentrates on peptide hormones and neurotransmitters.

Abbreviations: Standard abbreviations as recommended by IUPAC-1:UB Commission are used in this review (1). Other abbreviations include: Acm, acetamidomethyl; Bn, benzyl; Boc, t-butyloxycarbonyl; Bom, benzyloxymethyl; BHA, benzhydrylamine; Cbz, benzyloxycarbonyl; cHx, cyclohexyl; DCC, N,N-dicyclohexylcarbodiimide; DCM, dichloromethane; DIC, N,N-disopropylcarbodiimide; DMF, N,N-dimethylformamide; DMSO, dimethyl sulfoxide; For, formyl; Fmoc, 9-fluorenylmethyloxycarbonyl; HMP, 4-hydroxymethylphenoxyacetyl; HOAt, 1-hydroxy-7-azabenzotriazole; HOBt, 1-hydroxybenzotriazole-benzotriazole; MBHA, 4-methylbenzhydrylamine; Mtr, 4-methoxy-2,3,6-trimethylphenylsulphonyl; Np, nitrophenyl; NSu, N-succinimide; ONsu, N-hydroxysuccinimide; Trt, triphenylmethyl; Pam, 4-(hydroxymethyl) phenylacetamidomethyl; PAL, 5-[4-(9-fluorenylmethyloxycarbonyl)aminomethyl-3,5-dimethoxyphenoxoy]valeric acid; Pcp, pentachlorophenyl; Pmc, 2,2,5,7,8-pentamethyl-chroman-6-sulphonyl; Tcp, trichlorophenyl; TFA, trifluoroacetic acid; TFAA, trifluoroacetic anhydride; Tos, p-toluenesulfonyl.

Peptide hormones and neurotransmitters are among the most important classes of natural products in living systems, including in humans. They modulate and regulate most major biological functions and behaviors that are necessary for the maintenance of life and health, including feeding behavior, reproduction, response to stress, pain, cardiovascular function, kidney function, energy homeostasis, aggression, maternal behavior, learning behavior, and many others. In addition, these hormones often are intimately involved in virtually all major diseases, including hypertension and other cardiovascular dysfunctions, diabetes, obesity, cancer, CFS diseases, and many others. The recognition that polypeptides played such a central role in health and disease is a relatively recent scientific development. The discovery of insulin as the principal modulator (along with glucagon—a later discovery) of glucose levels in the blood and the determination of the structure and total synthesis of the peptide hormone and neurotransmitter oxytocin, which controls uterine contractions, milk ejection, and later maternal behavior; socialization; and aspects of sexual behavior are milestones not only in peptide and protein chemistry, biology, and medicine, but also in modern drug discovery and design. Especially in the case of oxytocin immediately after its structure determination and total
synthesis by du Vigneaud and co-workers in the early 1950s, he and many others began to modify the "functional groups" and structure of oxytocin to determine those features important to biological activity. Of central importance in this regard was the development of increasingly efficient synthetic methods (if you cannot make it you cannot study it) and analytical methods for the preparation and purification of polypeptides. In this review, we will outline synthetic methods that have been developed for the synthesis of polypeptide hormones and neurotransmitters, and in the process we also point out the key analytical procedures that have been used in the synthesis and purification of these important biological natural products. See Table 1 for an overview of this article and its contents.

### Table 1 Peptides discussed in this article

<table>
<thead>
<tr>
<th>Name</th>
<th>Target</th>
<th>Structure features</th>
<th>Major biological activities</th>
<th>Synthesis techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atrial natriuretic peptide (ANP)</td>
<td>ANP-R1 to R3 receptors in the peripheral system</td>
<td>28-residue peptide with a disulfide bridge</td>
<td>Modulate blood pressure</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Cholecystokinin (CCK)</td>
<td>CCK-A receptor, CCK-B receptor</td>
<td>33-residue peptide amide with a Tyr sulfonic acid</td>
<td>Modulate digestion system</td>
<td>Solution-phase and solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Gastrin</td>
<td>CCK-B receptor</td>
<td>14- and 17-residue peptide amide with a Glu rich region</td>
<td>Stimulate gastric acid secretion</td>
<td>Solution-phase and solid-phase Nα-Fmoc chemistry</td>
</tr>
<tr>
<td>Glucagon-like peptide 1 (GLP1)</td>
<td>GLP-1 receptor</td>
<td>30- and 31-residue peptide amide differed by C-terminal Gly</td>
<td>Stimulate insulin secretion</td>
<td>Solid-phase Nα-Fmoc chemistry</td>
</tr>
<tr>
<td>Glucose-dependent insulinotropic polypeptide (GIP)</td>
<td>GIP receptor</td>
<td>42-residue peptide</td>
<td>Inhibit gastric acid and stimulate insulin secretion</td>
<td>Solution-phase chemistry</td>
</tr>
<tr>
<td>Motilin</td>
<td>Motilin receptor</td>
<td>22-residue peptide</td>
<td>Stimulate migrating motor complex</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Neuropeptide Tyrosine (NPY)</td>
<td>NPY Y1 to Y5 receptors</td>
<td>36-residue peptide amide</td>
<td>Increase food intake</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Peptide Tyrosine (PY)</td>
<td>PY Y receptor</td>
<td>36-residue peptide amide</td>
<td>Inhibit food digestion</td>
<td>Solid-phase Nα-Fmoc chemistry</td>
</tr>
<tr>
<td>Secretin</td>
<td>Secretin receptor</td>
<td>27-residue peptide amide</td>
<td>Regulate pH of duodenal contents</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Somatostatin</td>
<td>Somatostatin 1 to 5 receptors</td>
<td>14- and 28-residue peptide with a disulfide bridge</td>
<td>Inhibit gut peptides and growth hormone</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Substance P</td>
<td>Neurokinin 1 receptor</td>
<td>11-residue peptide amide</td>
<td>Involved in pain, vomiting, cellular growth, mood regulation</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Vasoactive intestinal peptide (VIP)</td>
<td>VIPC1 and VIPC2 receptors</td>
<td>28-residue peptide amide, highly basic</td>
<td>Inhibit gastric acid and stimulate water and electrolytes secretion</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Name</th>
<th>Target</th>
<th>Structure features</th>
<th>Major biological activities</th>
<th>Synthesis techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corticotropin-releasing factor (CRF or CRH)</td>
<td>CRH-R1 and CRH-R2 receptors</td>
<td>41-residue peptide amide</td>
<td>Involved in stress response</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Gonadotropin-releasing factor (GnRF or GnRH)</td>
<td>GnRH I and II receptor</td>
<td>10-residue peptide amide</td>
<td>Stimulate the release of luteinizing hormone and follicle-stimulating hormone</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Growth hormone releasing factor (GRF or GRH)</td>
<td>GRH receptor</td>
<td>40–44-residue peptide amide</td>
<td>Stimulates growth hormone secretion</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Thyrotropin-releasing factor</td>
<td>TRF receptor</td>
<td>3-residue peptide amide</td>
<td>Stimulates thyroid-stimulating hormone and prolactin secretion</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Angiotensin II</td>
<td>Angiotensin II AT1 and AT2 receptor</td>
<td>8-residue peptide</td>
<td>Increase blood pressure</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Amylin</td>
<td>Amylin receptor</td>
<td>37-residue peptide amide</td>
<td>Regulate blood glucose level</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Glucagon</td>
<td>Glucagon receptor</td>
<td>29-residue peptide</td>
<td>Increase blood glucose level</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Insulin</td>
<td>Insulin receptor</td>
<td>Contains 21-residue A chain with an intra disulfide bridge and 30-residue B chain, linked by two disulfide bridges</td>
<td>Lower blood glucose level</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Pancreatic polypeptide</td>
<td>Pancreatic polypeptide receptor, NP Y4 receptor</td>
<td>36-residue peptide amide</td>
<td>Inhibit pancreatic secretion</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Adrenocorticotropic hormone (ACTH)</td>
<td>Melanocortin 2 receptor</td>
<td>39-residue peptide</td>
<td>Stimulate the cortex of adrenal gland</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Melanocyte-stimulating hormone (MSH)</td>
<td>Melanocortin 1 to 5 receptors</td>
<td>12–22-residue peptide or peptide amide, with His-Phe-Ang-Trp as the major pharmacophor</td>
<td>Darken skin and hair</td>
<td>Solid-phase N(^\alpha)-Boc chemistry</td>
</tr>
<tr>
<td>Name</td>
<td>Target</td>
<td>Structure features</td>
<td>Major biological activities</td>
<td>Synthesis techniques</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td><strong>Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Table 1 (Continued)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oxytocin</td>
<td>Oxytocin receptor</td>
<td>9-residue peptide amide with a disulfide bridge</td>
<td>Facilitate birth and breastfeeding</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Vasopressin</td>
<td>Arginine vasopressin receptor 1A, 1B, 2</td>
<td>9-residue peptide amide with a disulfide bridge</td>
<td>Regulate the retention of water</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Relaxin</td>
<td>Relaxin/insulin-like family peptide receptor 1 to 4</td>
<td>Contains 24-residue A chain with an intra disulfide bridge and 29-residue B chain, linked by two disulfide bridges</td>
<td>Facilitate birth</td>
<td>Solid-phase Nα-Boc and Nα-Boc chemistry</td>
</tr>
<tr>
<td>Calcitonin</td>
<td>Calcitonin receptor</td>
<td>32-residue peptide amide with a disulfide bridge</td>
<td>Lower blood calcium and phosphate levels</td>
<td>Solid-phase Nα-Fmoc chemistry</td>
</tr>
<tr>
<td>Calcitonin gene-related peptide (CGRP)</td>
<td>CGRP1 and 2 receptors</td>
<td>37-residue peptide amide with a disulfide bridge</td>
<td>Lower blood pressure</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Dynorphin</td>
<td>κ-opioid receptor</td>
<td>13- or 17-residue peptides</td>
<td>Endogenous κ-opioid receptor agonist effect</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Endorphin</td>
<td>δ-opioid receptors</td>
<td>16- to 31-residue peptides</td>
<td>Have analgesic effect</td>
<td>Solid-phase Nα-Boc and Nα-Boc chemistry</td>
</tr>
<tr>
<td>Enkephalin</td>
<td>δ-opioid receptors</td>
<td>5-residue peptides different at the C-termini sequence</td>
<td>Have analgesic effect</td>
<td>Enzymatic synthesis</td>
</tr>
<tr>
<td>Leumorphin</td>
<td>κ-opioid receptor</td>
<td>29-residue peptide</td>
<td>Endogenous κ-opioid receptor agonist</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Neurokinin A &amp; B</td>
<td>Neurokinin 1 to 3 receptors</td>
<td>10-residue peptide amide</td>
<td>Similar activities as substance P</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Eledoisin, Physalaemin and Kassinin (GRP)</td>
<td>Neurokinin 1 to 3 receptors</td>
<td>10- or 11-residue peptide amide</td>
<td>Similar activities as substance P</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Gastrin releasing peptide receptor</td>
<td>Gastrin-releasing peptide receptor</td>
<td>27-residue peptide amide</td>
<td>Stimulate gastrin release</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Neurotensin</td>
<td>Neurotensin 1, 2 receptors, Sortilin 1 receptor</td>
<td>13-residue peptide</td>
<td>Lower blood pressure, increase blood glucose levels</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
<tr>
<td>Bradykinin</td>
<td>Bradykinin 1, 2 receptors</td>
<td>9-residue peptide</td>
<td>Lower blood pressure, pain related</td>
<td>Solid-phase Nα-Boc chemistry</td>
</tr>
</tbody>
</table>
Peptide Hormones

Cardiac hormones

Atrial natriuretic peptide (ANP)

ANP is a 28-amino-acid peptide first discovered by de Bold et al. (2). It is released from heart atrial myocytes in response to a local arterial wall stretch. ANP acts on outer adrenal cells to decrease aldosterone production and blood pressure, increase salt and water excretion, and transudate plasma water to the interstitial (3).

The sequence of ANP is as follows: Ser-Leu-Arg-Arg-Ser-Arg-Ser-Leu-Arg-Arg-Ser-Val-Ser-Leu-Gly-Gly-OH.

The structure of ANP shows a disulfide ring at position 7 to 23, which is shared by the other two natriuretic peptides (brain natriuretic peptide and C-type natriuretic peptide). ANP has been prepared synthetically via solid-phase peptide synthesis using the N'-Boc strategy, von Geldern et al. (4) used a Biosearch Model 9500 automated synthesizer (Biosearch Technologies, Novato, CA) to assemble the amino acid sequence on a Merrifield resin preloaded with N'-Boc-Tyr. A standard protocol for solid-phase synthesis (DCM/CH2Cl2) with only a 5% yield from the fully protected peptide. The coupling of protected peptides with L-amino acid activated amino acid residues (WSCI-HOBT method) (6). The coupling of protected peptides with L-amino acid residues at the C-terminus was carried out using WSCI together with 2,4-dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazaine (HOOBT), which is known to minimize racemization of activated amino acid residues (WSCI-HOBT method) (6). The side-chain protections were as follows: Asp was protected by a cHx group; Arg and His by Tos groups; Lys by a Cl-Cbz group; Tyr by a Boc group; Ser by a Pha group; Trp by a For group. The fully protected peptide thus obtained was deprotected by HF in the presence of arsine and MeEt. The Pha and For groups were not removed under these conditions. The Tyr residue was then sulfonated with 40 equivalents of sodium phenylsulfinate in TFA in an ice bath for 2 hours. After precipitation by adding water, the product was treated with 40 equivalents of sodium carbonate and 10% NaOH in DMSO in an ice bath for 10 minutes to remove the remaining Pha and For groups. The crude product was purified by CM-cellulose chromatography and then by reversed-phase HPLC to produce a homogeneous preparation of the final product with only a 5% yield from the fully protected peptide. The scheme for the synthesis is shown in Fig. 1.

Penke’s group has reported two successful solid-phase syntheses of this peptide. The first study was using the N’-Boc chemistry published in 1987, with an overall yield of 1%. In 1991, they announced another synthesis using the N’-Fmoc strategy, and the protected Tyr(SO3H)-OH was introduced into the peptide sequence during the chain assembly (8). A new solid support system, 4-succinylamido-2,2′,4′-trimethoxy-4-acetamido-benzophenone (SAMBA)-polystyrene resin, was also made for the synthesis to accommodate the acid-labile sulfate ester.

The anchor SAMBA was prepared through a five-step synthesis (Fig. 2). First, 2,4-dimethoxy-acetanilide benzenoic acid (0.03 mol) at room temperature for 25 minutes, followed by adding ice and ethylacetate and 10% sodium carbonate solution to neutralize the substance. The organic phase was washed three times with sodium carbonate solution and once with H2O and dried over sodium sulfate. After removing the solvents in vacuo, the oily residue was crystallized by triturating with ether, obtaining 2,4-dimethoxy-acetanilide benzene. The product was then dissolved in a mixture of 4N NaOH (30 mL) and methanol (30 mL) and refluxed for 1 hour. The mixture was then treated with 40 equivalents of ethyl acetate and 10% sodium carbonate solution at 0°C for 20 minutes. After washing with methanol (10 mL) and 2,4-dimethoxybenzene (0.035 mol), the solution was concentrated in vacuo, and 2,4-dimethoxybenzylidenecarbonylamine (SAMBA)-polystyrene resin, was also made for the synthesis to accommodate the acid-labile sulfate ester.

The anchor SAMBA was prepared through a five-step synthesis (Fig. 2). First, 2,4-dimethoxy-acetanilide benzenoic acid (0.03 mol) was dissolved in an ice bath for 2 hours. After precipitation by adding water, the product was treated with 40 equivalents of sodium carbonate and 10% NaOH in DMSO in an ice bath for 10 minutes to remove the remaining Pha and For groups. The crude product was purified by CM-cellulose chromatography and then by reversed-phase HPLC to produce a homogeneous preparation of the final product with only a 5% yield from the fully protected peptide. The scheme for the synthesis is shown in Fig. 1.

The first study was using the N’-Boc chemistry published in 1987, with an overall yield of 1%. In 1991, they announced another synthesis using the N’-Fmoc strategy, and the protected Tyr(SO3H)-OH was introduced into the peptide sequence during the chain assembly (8). A new solid support system, 4-succinylamido-2,2′,4′-trimethoxy-4-acetamido-benzophenone (SAMBA)-polystyrene resin, was also made for the synthesis to accommodate the acid-labile sulfate ester.

The anchor SAMBA was prepared through a five-step synthesis (Fig. 2). First, 2,4-dimethoxy-acetanilide benzenoic acid (0.03 mol) was dissolved in an ice bath for 2 hours. After precipitation by adding water, the product was treated with 40 equivalents of sodium carbonate and 10% NaOH in DMSO in an ice bath for 10 minutes to remove the remaining Pha and For groups. The crude product was purified by CM-cellulose chromatography and then by reversed-phase HPLC to produce a homogeneous preparation of the final product with only a 5% yield from the fully protected peptide. The scheme for the synthesis is shown in Fig. 1.
Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of

Figure 1  Scheme for the synthesis of porcine CCK-33. Peptide coupling was done by the WSCI-HOBT method unless otherwise indicated. (a) Catalytic hydrogenolysis. (b) TFA. (c) Zn-acetic acid. (d) phenoxacyclic anhydride. (e) WSCI-HOBt. (f) WSCI-HOBUt. OPac = phenacyl. Taken from Kurano et al. (6).

dissolved in a mixture of pyridine (7.5 mL) and ethanol (50 mL), and it was refluxed for 45 minutes with hydroxyamine hydrochloride (0.067 mol). The solvent was then removed in vacuo, and the oily residue was triturated with 5% aqueous KHSO₄ solution. The resulting solid was filtered, washed with H₂O, and dried, obtaining 2,2',4,4'-trimethoxy-4-succinylamido-benzophenone oxime. The oxime (5 mmol) was suspended in glacial acetic acid (200 mL) and dissolved by warming. Then, 1 mL concentrated HCl and 0.2 g freshly prepared Pd-Catalyst were added, and the oxime was hydrogenated under pressure (1 M Pascal) for 5 hours. After filtering the catalyst and removing the solvents in vacuo, the oily residue was triturated with ether; the final product SAMBHA hydrochloride was crystallized as white solid. Four-mmol SAMBHA hydrochloride was then reacted overnight with Fmoc-OSu (4.4 mmol) in a mixture of H₂O (5 mL) and dioxane (15 mL) in the presence of sodium bicarbonate (670 mg) to get Fmoc-SAMBHA, which was used to couple with aminomethyl polystyrene suspended in DCM in the presence of DCC. The Nα-Fmoc group was cleaved by 10 minutes of treatment with 20% piperidine in DMF, and the SAMBHA resin was obtained with 0.33 mmol/g loading.

Nα-Fmoc-Tyr(SO₃Na)OH was prepared by reacting tyrosine with chlorosulfonic acid in TFA at −20°C for 5 minutes. Then Nα-Fmoc group was added by reacting with Fmoc-Cl in a mixture of 10% sodium carbonate solution and dioxane. The peptide was assembled using an Applied Biosystems 430A automatic peptide synthesizer (Applied Biosystems, Inc., Foster City, CA), and the SAMBHA-polystyrene resin was used as solid support. Amino acids were coupled as symmetric anhydrides using DCC in a 1:1 mixture of DCM-DMF or in pure DMF as coupling reagent, except that Nα-Fmoc-Asp(Ot-Bu) and Nα-Fmoc-Glu(Ot-Bu) were coupled as HOBt active esters and Nα-Fmoc-Tyr(SO₃Na) as a pentafluorophenyl ester. The side-chain protections were as follows: Asp and Ser were protected by t-Bu groups; His by a Fmoc group; Lys by a Boc group; Arg by a Pmc group. Coupling was repeated if a ninhydrin test was positive. After completion of the synthesis, the peptide resin was washed with a 1:1 mixture of DMF and DCM, and the coupling reaction was repeated with Fmoc-SAMBHA and DCC.
Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of

Figure 2

Synthesis of 4-succinylamido-2,2',4'-trimethoxybenzhydrylamine (SAMBHA).

that contained 0.4 g of NaOH for 1 minute, methanol for 2 minutes, and ether (4 × 2 minutes). Then the resin was dried. The protected pCCK-33 sulfate ester resin was washed with DCM (4 × 2 min) and was treated two times with 12 mL of 50% TFA solution in DCM that contained 5% ethanedithiol and 2% dimethyl sulfoxide (1 × 1 min, 1 × 15 min, 1 × 1 min). The TFA-containing solutions were pooled, and solvents were evaporated in vacuo. pCCK-33 was precipitated by adding peroxide-free ether.

Crude pCCK-33 was dissolved in the mixture of acetonitrile and 0.05 M (pH = 6.5) ammonium acetate buffer and was purified by prep-HPLC on a C18 column. The synthetic peptide was identified by using the native pCCK-33 standard. Pure fractions were pooled and lyophilized three times.

Gastrin

In 1905, Edkins discovered a hormone from the stomach and named it gastrin. Different forms and fragments of gastrin have been reported since then. The characteristic forms include a 17-residue amide called little gastrin and a 14-residue amide called minigastrin, which were known for their abilities of stimulating acid, pepsin, and pancreatic secretions in response to food (9).

The sequence of human little gastrin is as follows: pyroGlu-Gly-Pro-Trp-Leu-Glu-Glu-Glu-Glu-Glu-Ala-Tyr-Gly-Trp-Met-Asp-Phe-NH2.

The sequence of human minigastrin is as follows: Trp-Leu-Glu-Glu-Glu-Glu-Glu-Ala-Tyr-Trp-Met-Asp-Phe-NH2.

Gastrins and gastrin analogs have drawn great interest from chemists, and many total synthesis studies have been reported. Solution-phase synthesis was achieved by Anderson et al. in the 1960s (10). In 1980, Brown and coworkers reported a solid-phase synthesis of these two peptide amides (11). Methoxycarbonyl-functionalized poly(dimethylacrylamide) resin was chosen as the solid support. It was converted into a primary amine by treating it with ethylenediamine and then the addition of the reversible-linkage agent [2,4,5-trichlorophenyl 4-(hydroxymethyl)benzoate] in the presence of HOBT. The peptide sequence was then assembled using initially N-Boc-Phe anhydride in the presence of catalyst (dimethylaminopyridine) to establish the peptide–resin ester bond, followed by 16 cycles of N-Boc amino-acid coupling. The symmetric anhydride method was used for coupling in all cases except for the last Gln, which was introduced as its p-nitrophenyl ester in the presence of HOBT. Deprotection of the N-Boc groups was done by treating with HCl-AcOH, whereas the Nα-Fmoc groups were removed by 20% piperidine in DMF. Gln was chosen as the N-terminal residue rather than pyroglutamic acid, which is present in little gastrin, so as to permit possible future extension into the prohormone series. Side-chain protections were as follows: Asp, Tyr, and Glu were protected by tert-butyl groups. After completing the synthesis, side-chain tert-butyl derivatives were removed by treatment with 90% aqueous TFA, and the free peptide was released from the resin by ammonolysis in saturated methanolic ammonia for 22 hours. Residual resin analysis indicated the removal of 91% of the peptide. At this stage, the product consisted of a mixture of the 17-residue N-terminal glutamine peptide amide and its cyclized pyroglutamyl analog (5:1). Cyclization was brought nearly to completion by treatment with 20% acetic acid (30°C, 64 hours under Argon).

HPLC analysis showed the crude peptide has three major impurity peaks (Fig. 3a). The residual glutaminyl peptide
GLP1 is a peptide hormone produced by the posttranslational processing of proglucagon secreted from L-cells in the lower gut. It has two forms: GLP1(7-36), which is a 30-residue glycine-extension peptide. It stimulates insulin secretion in response to the plasma glucose concentration change, and it is considered to have great potential in treating Type 2 diabetes.

Motilin is a 22-amino acid peptide first isolated from hog duodenal mucosa by Brown and colleagues in 1972 (18, 19). It is secreted from the duodenum and proximal jejunum in response to food. Two major physiological effects of GIP are inhibition of gastric acid secretion and stimulation of insulin release.

The sequence of porcine GIP is as follows: Tyr-Ala-Glu-Gly-Thr-Phe-Ile-Pro-Ile-Phe-Thr-Ser-Asp-Val-Ser-Ser-Tyr-Leu-Glu-Gly-Gln-Ala-Ala-lys-Glu-Phe-Ile-Ala-Trp-Leu-Val-Lys-Glu-Arg-NH₂.

The sequence of motilin is as follows: Phe-Val-Pro-Ile-Phe-Thr-Tyr-Gly-Glu-Leu-Gln-Arg-Met-Glu-Glu-Glu-Leu-Leu-Ala-Gln-Leu-Lys-Gly-Lys-Lys-Lys-Gln-

The total synthesis of GLP-1(7-36) was reported by A delhorrst et al. (12) in 1994 using the solid-phase peptide synthesis method. The peptide was synthesized on an ABI MED 422 multiple synthesizer. An N-\(\alpha\)-Fmoc strategy, which was modified according to Gauselv et al. (13), was used starting from a Rink-resin. The side-chain protections were as follows: Asn, Gin, and His were protected by Tt groups; Arg by a Pmc group; Asp, Glu, Tyr, Ser, and Thr by t-buty1 groups; and Trp and Lys by Boc groups. The peptides were cleaved from the residue and deprotected by a treatment of TFA/trimethylsilane/water (92.5:5:2.5) for 120 minutes and precipitated in t-butylmethyl ether and lyophilized from 10% acetic acid. The crude peptides were purified by reverse phase HPLC with Superpak Pep. S C2C18 column, obtaining a purity of >95%.

Glucose-dependent insulinotropic polypeptide (GIP)

Glucose-dependent insulinotropic polypeptide is originally known as gastric inhibitory polypeptide (GIP), which is a 42-residue peptide first isolated by Brown and Dryburgh (14). It is secreted from the duodenum and proximal jejunum in response to food. Two major physiological effects of GIP are inhibition of gastric acid secretion and stimulation of insulin release.


Gluconeogenic amino acids can be directly formed on HF cleavage of the final peptide residue 25 was introduced exceptionally by itself to the 26-42 or 26-43 peptide intermediate by the active ester method. The Cbz and Bn groups were removed by catalytic hydrogenation and the Boc and tBu groups by TFA treatment during the course of fragment condensation. After completing the synthesis, the crude protected peptide was treated with sodium in liquid ammonia to remove N-Cbz and N-Tos groups. The product was first purified by gel filtration on Sephadex G-25, followed by ion-exchange chromatography on CM-cellulose and HPLC with TSK-Gel LS-420K C-18 column.

Motilin was the first identified gut hormone to be a hormone that caused smooth muscle contraction. It was found to be released from enteroendocrine cells of the duodenal-jejunal mucosa to control gastrointestinal muscles by controlling the migrating motor complex.

The sequence of motilin is as follows: Phe-Val-Pro-Ile-Phe-Thr-Tyr-Gly-Glu-Leu-Gln-Arg-Met-Glu-Glu-Glu-Leu-Leu-Ala-Gln-Leu-Lys-Gly-Lys-Lys-Lys-Gln-

Ikeda (20) reported the first solid-phase synthesis of motilin in 1980, and the overall yield was 1.8%. Coy et al. (21) improved the synthesis and got a 12% yield. Coy’s synthesis was carried out on a Beckman model 990 automatic peptide synthesizer using N-\(\alpha\)-Boc chemistry. Motilin was assembled stepwise on a 1% cross-linked BHA resin with Boc-a-benzyl-Glu as the first amino acid to be incorporated. In this manner, C-terminal Gin could be directly formed on HF cleavage of the final peptide from the resin. The detailed coupling schedule is shown in Table 2 (22). The symmetric anhydride method was used for the synthesis of the peptide.
Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of

**Table 2** Schedule of events for assembling the peptide BHA resin

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagent/solvent and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCM wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>33% TFA-DCM</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>33% TFA-DCM</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>DCM wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>E7H3 wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CHCl3 wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>10% Et$_3$N-CHCl$_3$ (&gt; 2)</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>CHCl3 wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>DCM wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>N$:^\alpha$-Boc-amino acid (3 eq) + DIC (3 eq) in DCM*</td>
<td>60</td>
</tr>
<tr>
<td>11</td>
<td>DCM wash (&gt; 3)</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>E7H3 wash (&gt; 3)</td>
<td>1</td>
</tr>
</tbody>
</table>

*Asn and Gln were dissolved in DMF, N$:^\alpha$-Boc-Arg(Tos) in 10% DMF-DCM. The extent of couplings were monitored by the ninhydrin test. Couplings, which were incomplete after 1 hour, were recoupled by using the appropriate symmetric anhydride procedure at room temperature in DMF.

After the completion of the solid-phase synthesis, 0.25 mmol of peptide-resin was treated with HF (30 mL) that contained 10% anisole for 30 minutes at 0$^\circ$C. After rapid removal of HF under a stream of nitrogen, free, deprotected peptide amide was precipitated by addition of ether, and then it was partially purified on Sephadex G-25. A further purification was done on an octadecylsilica column (5 x 45 cm) by elution with a linear gradient formed from 200 mL each of 15 and 30% 1-propanol in 10% acetic acid at a flow rate of 2 mL/min. Examination of the fractions by TLC and HPLC located the peptide in the third peak. Fractions were pooled for maximum purity to give 70 mg (10%) of motilin.

**Neuropeptide Y (NPY)**

NPY is a 36-amino acid peptide amide first isolated from porcine brain by Tatemoto et al. (23, 24) in 1982. It is found distributed in the central nervous system where it is involved in the control of blood pressure and appetite, and also in the
peripheral nervous system, where it is a potent vasoconstrictor and presynaptic inhibitor of neurotransmission.

The sequence of porcine NPY is as follows: Tyr-Pro-Ser-Lys-Pro-Asp-Asn-Pro-Gly-Glu-Asp-Ala-Pro-Ala-Glu-Asp-Leu-Arg-Tyr-Tyr-Ser-Ala-Leu-Arg-His-Tyr-Ile-Ala-Leu-Ile-Thr-Arg-Gln-Arg-Tyr-NH₂.

Krstenansky et al. (25) successfully prepared porcine NPY in 1987. It was synthesized by solid-phase techniques with an Applied Biosystems Inc. Model 430-A peptide synthesizer using MBHA resin. All residues were double coupled as the preformed symmetric anhydrides of the Nα-Boc amino acid derivatives except for Asn and Gln, which were coupled with the DCC/HOBT methodology (refer to the section on "Corticotropin-releasing factor synthesis" for details). The side-chain protections were as follows: Arg and His were protected by Tos groups; Asp and Glu by cHx groups; Ser and Thr by Bn groups; Lys by a 2,6-Cl-Cbz group; Tyr by a 2-Br-Cbz group. After completing the synthesis, deprotection and the resin cleavage were performed by a treatment with HF that contained 5% anisole for 50 minutes at 0°C. The HF was removed in vacuo at 0°C, and the crude peptide was precipitated with ether and extracted with 30% acetic acid. The extract was lyophilized. First-round purification was carried out on a 2.6 × 92-cm Sephadex G-15 column at 30 mL/h and 20 minutes per fraction (Fig. 6, top) (25). The fractions that contained the major peak detected by UV at 254 nm were combined and lyophilized. The residue was then purified for the second round by preparative HPLC on a 21.4 × 250 mm Rainin Dynamax C-18 column at 10 mL/min with 34% MeCN in 0.1% TFA at 0°C. The peptide was precipitated with ether and extracted with 30% acetic acid. The lyophilized residue was then purified using a reverse-phase HPLC column (MCI GEL ODS-1H4, 10 × 300 mm, Mitsubishi Kasei, Japan) with a linear gradient of 0.1% TFA/water and 0.1% TFA/acetonitrile in DMF at 0.1 mL/min. Taken from Krstenansky et al. (23).

Peptide tyrosine tyrosine (PYY)
PYY is a 36-amino-acid peptide originally isolated from porcine upper intestine by Tatemoto et al. (26, 27). It contains an N-terminal tyrosine and a C-terminal tyrosine amide and therefore is named peptide YY. It has a high degree of sequence homology (70%) with neuropeptide Y, and it strongly inhibits pancreatic exocrine secretion and jejunal and colonic motility as well as causes vasoconstriction.

The sequence of human PYY is as follows: Tyr-Pro-Ile-Lys-Pro-Glu-Ala-Pro-Gly-Glu-Asp-Ala-Ser-Pro-Glu-Glu-Leu-Asn-Arg-Tyr-Tyr-Ala-Ser-Leu-Arg-His-Tyr-Leu-Asn-Leu-V al-Thr-Arg-Gln-Arg-Tyr-NH₂.

Tatemoto et al. (28) also reported a solid-phase synthesis of PYY. Human PYY was prepared manually using Nα-Fmoc protection strategy on a 4-aminomethyl-3,5-dimethylphenoxyl resin (PAL resin). The coupling reaction was carried out using a five-fold excess of Nα-Fmoc amino acid and DIC in DMF. The u-amino group of the growing peptide was deprotected with 20% piperidine in DMF for 10 minutes. The side-chain protections were as follows: Asp, Glu, Ser, Thr, and Tyr were protected by t-Bu groups; Lys by a Boc group; His by a Trt group; Arg by a Mtr group. The peptide was deprotected and cleaved from the resin by a treatment with TFA/thioanisole/ethanedithiol/anisole (90/5/3/2) for 8 hours at room temperature. After filtration, TFA was removed in vacuo, and the peptide was precipitated by ether. The precipitate was redissolved in 0.1 M acetic acid and lyophilized. The crude peptide was purified using an reverse-phase HPLC column (MCI GEL ODS-1H4, 10 × 300 mm, Mitsubishi Kasei, Japan) with a linear gradient of 0.1% TFA/water and 0.1% TFA/acetonitrile. The synthetic

Figure 5 Purification of crude synthetic motilin on column of LRP-1 ODS silica. Taken from Coy et al. (21).

Figure 6 HPLC of synthetic porcine NPY. Top: after HF cleavage and Sephadex G-15 chromatography. Bottom: after preparative HPLC purification. Condition: Vydac 218TP54 (4.6 × 250mm) C-18 column with a linear gradient of 25–100% acetonitrile in 0.1% TFA over 25 minutes at 2 mL/min. Taken from Krstenansky et al. (25).
peptide obtained was found to coelute in HPLC with the natural peptide, and the results of amino acid and sequence analyses indicated that it was identical to human PYY.

Secretin
Secretin is a 27-residue peptide amide hormone produced by 5 cells of the duodenum. The primary effect of secretin is stimulating the release of bicarbonate from liver, pancreas, and duodenal tissues to inhibit gastrin-induced gastric acid release. It also enhances the effects of cholecystokinin and promotes normal growth and maintenance of the pancreas.


Several classic syntheses of secretin have been reported using either fragment condensation or repetitive coupling of single residues (29); one solid-phase synthesis has also been reported (30). However, a peptide-stability problem arose, and it caused difficulty in developing an efficient solid-phase synthesis of the peptide. Coy and Gardner (31) developed a two-stage rapid purification method to avoid traditional long purification routes, and this approach solved potential degradation problem of the peptide.

Coy’s synthesis was carried out on a Beckman model 990 automatic peptide synthesizer (Beckman Instruments, Fullerton, CA) using N-Boc chemistry. Secretin was assembled stepwise on a 1.0-mmol of 1% cross-linked BHA resin support using the synthesis schedule described in Table 2. The side-chain protections were as follows: Arg, Glu, Ser, and Thr were protected by Bn groups; Arg and His by Tos groups. Each amino acid was successively coupled in the presence of DIC and, in the case of Asn or Gln, 1 equivalent of HOBt was added. Couplings were monitored at each step by using the ninhydrin test. Couplings that were incomplete within 1 hour were recoupled by using the appropriate symmetric anhydride preformed at room temperature in DMF.

After the completion of the solid-phase synthesis, the peptide resin was treated with HF that contained 10% anisole for 30 minutes at 0°C. After rapid removal of HF under a stream of nitrogen, free, deprotected peptide was precipitated by addition of ether, and it was washed and extracted into 2 M acetic acid. The protected somatostatin tetradecapeptide was synthesized in a stepwise manner on chloromethylated resin prepared in a stepwise manner on a column (2.5 × 95 cm) of ODS silica LRP-1 (Whatman) (13–24 µm) with a linear gradient of 15 and 35% 1-propanol in 0.1 M ammonium acetate (pH 4) at a flow rate of 5 mL/min and pressure about 80 psi. Taken from Coy et al. (21).

Figure 7 Chromatography of crude synthetic secretin eluted on a column (2.5 × 45 cm) of ODS silica LRP-1 (Whatman) (13–24 µm) with a linear gradient of 15 and 35% 1-propanol in 0.1 M ammonium acetate (pH 4) at a flow rate of 5 mL/min and pressure about 80 psi. Taken from Coy et al. (21).

are secreted not only by cells of the hypothalamus but also by delta cells of the stomach, intestine, and pancreas. They inhibit the release of numerous gut peptides, like CCK, gastrin, secretin, motilin, GIP, and they also inhibit insulin and glucagon secretion from the pancreas.

The sequence of hypothalamic somatostatin 14 is as follows: Ala-Gly-c(Cys-Lys-Asp-Pro-Phe-Thr-Cys)-OH.

The sequence of hypothalamic somatostatin 28 is as follows: Ser-Ala-A s n-Ser-Ala-Asp-Pro-Ala-Met-Ala-Pro-Arg-Glu-Arg-Lys-

Somatostatin 14 was isolated, characterized, and synthesized by Guillemin’s group in 1973 (32, 33). Several independent studies of the peptide’s synthesis were reported immediately after that (34–37). Among those studies, River reported in detail the first solid-phase synthesis of the peptide in a highly purified form. The protected somatostatin tetradecapeptide was synthesized in a stepwise manner on chloromethylated resin prepared according to Stewart and Young (38). N-Boc chemistry was used for the synthesis: all amino acids were coupled as symmetric anhydrides using DCC as the coupling reagent except for Asn, which was coupled as Np active ester. The side-chain protections were as follows: Thr and Ser were protected by Bn groups; Lys by a 2-CI-Chz group. N'-Boc-Cys(p-OMe-Bn) was used because it is easily removed by HF. Ala and Gly were introduced as a Cbz-protected dipeptide (Cbz-Ala-Gly-OH) to have a reliable internal standard to evaluate the amino-acid analyses. TFA was used for cleavage of the N’-Boc protecting groups, and 1, 2-ethanediol/thiol was added for protection of the Thr residue from oxidation, which is a problem that has long been recognized in solid-phase synthesis. The completion of coupling was monitored by the ninhydrin test. The detailed schemes used for the synthesis are reported in Tables 3 and 4.

Cleavage and concomitant deprotection of the tetradecapeptide were achieved by HF in presence of anisole. After removing HF in vacuo and work-up, a white fluffy material that accounted for 60% of the calculated yield was obtained. A second HF treatment was applied to the resin, and 25% more of the peptide was obtained. The crude material was subjected to gel filtration. The main peak (Fig. 8, profile I) after lyophilization and...
by mixing the resin and peptide synthesis techniques. The first amino acid was loaded on chloromethyl/1% divinylbenzene resin by using solid-phase WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY

al. (39) in 1986.

Table 3 Schedule for DCC Coupling in Solid-phase Synthesis of Somatostatin-14

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagents or solvents and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCM wash, 80 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>MeOH wash, 30 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>DCM wash, 80 mL (× 3)</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>50% TFA + 5% 1,2-ethanediol in DCM, 70 mL (× 2)</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>DCM wash, 80 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Et,N 12.5% in DMF, 70 mL (× 2)</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>MeOH wash, 40 mL (× 2)</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>DCM wash, 80 mL (× 3)</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>Boc-amino acid (10 mmol) in 30 mL of DMF (× 1) + DCC (10 mmol) in DMF</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>MeOH wash, 40 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>Et,N 12.5% in DMF, 70 mL (× 1)</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>MeOH wash, 30 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>DCM wash, 80 mL (× 2)</td>
<td>3</td>
</tr>
</tbody>
</table>

* Aliquots taken for ninhydrin test: if negative, go back to step 1, if positive or slightly positive, go back to steps 9 → 13.

Table 4 Solid-phase synthesis of somatostatin schedule for Nα-Boc-Asn-PNP coupling

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagents or solvents and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>DMF wash, 60 mL (× 3)</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>Boc-Asn-PNP (15 mmol) in 20 mL of DMF (× 1)</td>
<td>800</td>
</tr>
<tr>
<td>11</td>
<td>MeOH wash, 30 mL (× 4)</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>Et,N 12.5% in DMF, 30 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>MeOH wash, 30 mL (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>DCM wash, 80 mL (× 3)</td>
<td>3</td>
</tr>
</tbody>
</table>

*Same as lines 1-8 in Table 3; followed by changes shown. Aliquots taken for ninhydrin test: if negative, go back to step 1, if positive or slightly positive, go back to steps 9 → 13.

routine handling was reapplied on the same column under the same conditions in an attempt to obtain one single symmetrical pattern (Fig. 8, profile I). A total yield of 28% in purified linear somatostatin was obtained.

Synthesis of somatostatin-28 has been reported by Nicolas et al. (39) in 1986. Nα-Boc protected amino acids were assembled on chlorosulfonyl/1% divinylbenzene resin by using solid-phase peptide synthesis techniques. The first amino acid was loaded by mixing the resin and Nα-Boc-Cys(3,4-dimethylbenzyl)-OH in DMF for 18 hours at 50°C in the presence of KI. The side-chain protections were as follows: Asp, Thr, Ser, and Glu were protected by Bn groups; Lys by a 2-Cl-Cbz group; Tyr by a 2-Br-Cbz group; Arg by a Tos group; Cys by a 3, 4-dimethyl-Bn group; Trp by a For group. The coupling was done using the symmetric anhydride method according to the protocol shown in Table 4 (40), except Asn was first coupled for 15 hours as its p-nitrophenyl ester in DMF and then by the use of HOBT as reported (41). After the coupling was completed, the peptide-resin was subjected to steps 1–10, washed with EtOH, and dried. The cleavage of the peptide was done by treating the peptide-resin with liquid HF in the presence of anisole for 30 minutes at −20°C and then 60 minutes at 0°C. After removal of HF by a stream of nitrogen, the residue was stirred in EidA.c, and the solid material was filtered off. The filtrate was purified on a Sephadex G-25 column (2.6 cm × 100 cm) equilibrated and run with degassed 10% AcOH saturated with nitrogen. The main peak was lyophilized, which yielded 665 mg product. Deformylation and formation of the disulfide bond were carried out by dissolving 100 mg material in water (25 mL) that contained 75% hydrazine and taken to pH 11.5 with 1 M NaOH. After 2 minutes, the solution was diluted to 1000 mL with distilled water, adjusted to pH 8 with glacial acetic acid, and allowed to stand at 24°C. After 22 hours, less than 1% free thiols remained as assessed by the Ellman’s method (42). After lyophilization, the material was dissolved in 5 mL of 10% AcOH and filtered through a Sephadex G-50F column (2.5 cm × 91 cm) in 10% AcOH. Isolation of material in the major peak gave 63 mg. This material was subjected to chromatography on CM-cellulose in a 1.5 cm × 35-cm column initially equilibrated with 10 mM NH₄OAc, pH 4.2. A gradient was applied through a 200 mL constant volume mixing chamber that contained the initial buffer by introducing 0.45 M NH₄OAc, pH 7. Material in the major peak was isolated to give 16.5 mg (a 6% overall yield) of somatostatin-28.
670–740 mL, 500 mg, 40%; profile II (employing the solid-phase technique using overnight with an excess of hydrochloride was neutralized with triethylamine and stirred with a mixture of Ac₂O and triethylamine in DMF overnight. The remaining unreacted amino groups of the resin were acetylated with 25% triethylamine-DCM. The remaining amino acid was used for coupling with a reaction time of 6 hours except for Gin, which was incorporated as a fivefold excess of its Np active ester with an 11-hour coupling time. Dry DCM was used as the coupling solvent except for Arg and Gin, which required purified DMF. N-Boc deprotection was accomplished by 50% TFA in DCM with neutralization by 10% Et₃N in DCM. The completion of coupling was monitored by the ninhydrin test, and recoupling was done when needed (refer to the sections on “Möllin synthesis” and “Secretin synthesis” for details). Cleavage of the peptide from the resin, with simultaneous removal of the protecting groups and formation of the carboxy terminal amide, was affected with 20 mL of dry HF in the presence of 2 mL of anisole for 1 hour at 0 °C. After removal of the excess HF in vacuo, the resin was washed with EtOH to remove anisole, followed by 0.5N H₂OAc to extract the peptide; 424 mg crude product was obtained and lyophilized.

Purification was done by gel filtration on a 102 × 2.5 cm column of Bio-Gel P2 eluted with 1.1% AcOH, with detection of the peptide peaks by UV at 256 nm. The main fraction (253 mg) was partitioned on a 100 × 1.5 cm column of Sephadex G-25 eluted with the system 0.1% AcOH-n-BuOH-Pyr (11:5:3) with detection of the peptide peaks by the Folin-Lowry procedure (45) at 700 nm, giving 90 mg (16.7%) of pure substance P. N-Fmoc synthesis of substance P also has been reported (46).

**Figure 8** Gel filtration of crude and purified synthetic reduced somatostatin. Column: 2.5 × 100 cm Sephadex G-25. 2 N AcOH, 10-2 M (mercaptoethanol). V₀ = hold-up volume; profile I (katharanth); 500 mg from I, yield 660–726 mg, 350 mg, 28%. Taken from Boxer et al. (57).

### Substance P

Substance P is an 11-amino-acid peptide amide hormone discovered by von Euler and Gaddum in 1931 (43) from certain tissue extracts, especially from intestinal plain muscles and brains of horses. It is involved in the transmission of pain impulses from peripheral receptors to the central nervous system. It is also involved in the vomit reflex, stimulates salivary secretions, and induces vasodilation. Antagonists seem to have antidepressant properties.

The sequence of substance P is as follows: Arg-Pro-Lys-Pro-Gln-Gln-Phe-Phe-Gly-Leu-Met-NH₂

This peptide was synthesized by Fisher et al. (44) in 1974 employing the solid-phase technique using N-Boc chemistry. BHA type resin was used for the synthesis on a Beckman Model 990 solid-phase peptide synthesizer. Dried BHA-resin hydrochloride was neutralized with triethylamine and stirred overnight with an excess of N-Boc-Met and DCI in DCM. The remaining unreacted amino groups of the resin were acetylated with Ac₂O and triethylamine in DCM overnight. Amino-acid analysis of the resin, after hydrolysis in 6 N HCl-propionic acid (1:1), gave a value of 0.2 mM/g of Met. The side-chain protections were as follows: Lys was protected by a Cbz group; Arg by a Tos group. A 2.5-fold excess of each amino acid was used for coupling with a reaction time of 6 hours except for Gin, which was incorporated as a fivefold excess of its Np active ester with an 11-hour coupling time. Dry DCM was used as the coupling solvent except for Arg and Gin, which required purified DMF. N-Boc deprotection was accomplished by 50% TFA in DCM with neutralization by 10% Et₃N in DCM. The completion of coupling was monitored by the ninhydrin test, and recoupling was done when needed (refer to the sections on “Möllin synthesis” and “Secretin synthesis” for details). Cleavage of the peptide from the resin, with simultaneous removal of the protecting groups and formation of the carboxy terminal amide, was affected with 20 mL of dry HF in the presence of 2 mL of anisole for 1 hour at 0 °C. After removal of the excess HF in vacuo, the resin was washed with EtOH to remove anisole, followed by 0.5N H₂OAc to extract the peptide; 424 mg crude product was obtained and lyophilized.

Purification was done by gel filtration on a 102 × 2.5 cm column of Bio-Gel P2 eluted with 1.1% AcOH, with detection of the peptide peaks by UV at 256 nm. The main fraction (253 mg) was partitioned on a 100 × 1.5 cm column of Sephadex G-25 eluted with the system 0.1% AcOH-n-BuOH-Pyr (11:5:3) with detection of the peptide peaks by the Folin-Lowry procedure (45) at 700 nm, giving 90 mg (16.7%) of pure substance P. N-Fmoc synthesis of substance P also has been reported (46).

### Vasointestinal peptide (VIP)

VIP is a 28-amino-acid residue peptide amide isolated in 1970 by Said and Mutt in the course of the purification of secretin from porcine duodenum (47). It was characterized 2 years later as a highly basic peptide related to gut peptides. It inhibits acid and pepsin secretion, acts as a neurotransmitter in peripheral autonomic nervous system, and increases secretion of water and electrolytes from pancreas and gut.

The sequence of porcine VIP is as follows: His-Ser-Asp-Ala-Val-Phe-Thr-Asp-Asn-Tyr-Thr-Arg-Leu-Arg-Lys-Gln-Met-Ala-Lys-Lys-Tyr-Leu-Asn-Ser-Ile-Leu-A sn-NH₂

Bolansky reported the first synthesis of VIP using solution-phase peptide synthesis (48). Several other syntheses, both solid-phase and solution-phase, also have been reported (49–52). Among these syntheses, Fourrier’s study is representative. N-Boc solid-phase synthesis techniques were used. Peptide syntheses were carried out with a Vega peptide synthesizer (Vega Biotech, Tucson, AZ) in the automatic mode following operations as described (53) (similar to Tables 3 and 4) except that 1% D,L-methionine was used in TFA instead of 5% thioanisole. The first amino acid was loaded to the BHA resin (0.21 meq/g) via the DCC-HOBt procedure after neutralization with triethylamine and thioanisole. The first amino acid was loaded to the BHA resin (0.21 meq/g) via the DCC-HOBt procedure after neutralization with triethylamine and thioanisole. The first amino acid was loaded to the BHA resin (0.21 meq/g) via the DCC-HOBt procedure after neutralization with triethylamine and thioanisole.
Naturally Occurring Peptide Hormones and Neurotransmitters. Synthesis of somatostatin-28

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagents or solvents and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCM wash, 15 mL ((\times 3))</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>55% TFA-DCM wash, 15 mL ((\times 1))</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>55% TFA-DCM wash, 15 mL ((\times 3))</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>DCM wash, 15 mL ((\times 2))</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>25% Dioxane-DCM wash, 15 mL ((\times 1))</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>DCM wash, 15 mL ((\times 1))</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>5% DIEA-DCM wash, 15 mL ((\times 2))</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>DCM wash, 15 mL ((\times 6))</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>Symmetric anhydride of Boc-amino acid (1.5 mmole) in DCM 13 mL</td>
<td>60*</td>
</tr>
<tr>
<td>10</td>
<td>DCM wash, 15 mL ((\times 1))</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Symmetric anhydride of Boc-amino acid (1.5 mmole) in DCM 13 mL</td>
<td>60*</td>
</tr>
<tr>
<td>12</td>
<td>DCM wash, 15 mL ((\times 2))</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>5% DIEA-DCM wash, 15 mL ((\times 2))</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>DCM wash, 15 mL ((\times 6))</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>Symmetric anhydride of Boc-amino acid (1.5 mmole) in DCM 13 mL</td>
<td>60*</td>
</tr>
<tr>
<td>16</td>
<td>DCM wash, 15 mL ((\times 3))</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>33% EtOH-DCM wash, 15 mL ((\times 3))</td>
<td>3</td>
</tr>
</tbody>
</table>

*In the case of N\(^\alpha\)-Boc-Val, -Leu, -Thr, 90 minutes of reaction time.

Bn groups; His by a Boc group; Tyr by a 2-Br-Cbz; Arg by a Tos group; Lys by a 2-Cl-Cbz group. Every coupling was monitored for completion using the ninhydrin test (refer to the sections on “Motilin synthesis” and “Secretin synthesis” for details). Excess amino groups were acetylated when the test showed slightly positive results (−1%). After the last amino acid was introduced, the cleavage and deprotection were executed in the K-F reaction vessels of a liquid HF apparatus. HF (8–10 mL/g) and a 5-molar excess D,L-methionine were added to the vessel that already contains anisole. The reaction proceeded for 30 minutes at \(−20\) °C and for another 30 minutes at 0 °C. The HF was rapidly evaporated in vacuo, and the resin was washed with ether. The crude peptide was extracted with 30% acetic acid \((\sim 200 \text{ mL})\), and the solution was lyophilized.

The crude peptide was dissolved in a minimal volume of 0.005-M ammonium acetate buffer at pH 6.0, and then it was applied to a column of Whatman CM-23 ion exchange resin equilibrated with the same buffer. The peptide was eluted with a pH and ionic gradient of adding 0.20 M pH 7.5 ammonium acetate buffer to the initial one at a flow rate of 700 mL/h. The elution was monitored by UV at 280 nm. Frations (12 mL) that corresponded to the major peak were collected and lyophilized twice to eliminate the volatile salt. A second chromatographic step on CM-23 resin was achieved with a pH and ionic gradient obtained by the addition of 0.12 M pH 7.5 to 0.005 M pH 5.2 at a flow rate of 700–1000 mL/h. In the third chromatographic step on CM-23 resin was achieved with a pH and ionic gradient obtained by the addition of 0.12 M pH 7.5 to 0.005 M pH 5.9 at a flow rate of 700–1000 mL/h. In the third chromatographic step, the semi-purified VIP was injected on a partition chromatography column packed with Sephadex G-50 resin and equilibrated with the solvent system 1-butanol, acetic acid, and 0.5%-pyridine (5:10:3). The flow rate was set to 20 mL/h, and 5-mL fractions were collected. Final purification (Fig. 9) was carried out on HPLC using a semi-preparative Waters µBondapak C-18 column under isocratic conditions: methanol ammonium acetate 0.10 M, pH 5.2 (70:30). The flow rate was adjusted to 1.2 mL/min, and the peptide was detected at 246 nm by UV. Two-hundred fifty µL of 40 mg/mL peptide solution were injected. The main peak was collected and lyophilized, which gave an overall yield of 4%.

![Figure 9](image-url)
Hypothalamic peptide hormones

Corticotropin-releasing factor (CRF or CRH)

CRF is a 41-amino-acid peptide first isolated from ovine hypothalamic extracts and characterized by Vale et al. (54) in 1981. It has a high potency for stimulating the secretion of corticotropin-like and β-endorphin-like immunoreactive substances.


Vale et al. (54) also synthesized the peptide, but they did not provide details of the synthesis. In 1982, Sueiras-Diaz and Coy (55) published the first detailed synthesis of ovine CRF using N-Boc solid-phase synthesis techniques. The amino acids were assembled on a 1% cross-linked BHA resin (0.1 g: 0.49 mmol amino groups) using a Beckman 990 automatic synthesizer (Beckman Instruments, Fullerton, CA). The symmetric anhydride procedure in DMF was used for the coupling of amino acids, and DIC was the coupling reagent. In the case of Asn and Gin, they were coupled with an equimolar amount of HOBt.

N-Boc protection was removed at each stage by two treatments with 39% TFA in DCM for 1 minute and 25 minutes. Reactive side-chains protections were as follows: Ser and Thr were protected by Bn groups, Gln by a Np group, Arg by a 2-Cl-Cbz group, His and Tyr by a 2-Ch-Orn group, and Lys by a 2-Cl-Cbz group.

The final purification was achieved by a preparative RP-HPLC procedure using a C-18, 10 µm, 300 Å (25 × 25 cm) Synchropack column loaded with 11.7 mg and eluted with a gradient of 25–35% isopropanol in 0.1% TFA developed over 30 minutes. Flow rate 1 mL/min, absorption at 215 nm, taken from Sueiras-Diaz et al. (55).


Gonadotropin-releasing factor (GnRF or GnRH)

GnRH, which was originally known as luteinizing hormone (LH) and follicle-stimulating hormone (FSH) releasing hormone, is a 10-residue polypeptide isolated from porcine hypophysis by Schally et al. (56) in 1971. It acts on the gonadotrope to control the release of LH and FSH.


Matuo et al. (57) also have prepared this peptide via solid-phase peptide synthesis. The protected decapeptide resin ester (N-Boc-Glu-His-Trp-Ser-D-Pro-Gly-Glu-Lys-Asp-Tyr-Gly-Leu-Arg-N(SO₂)-Pro-Gly-Pyr-H) corresponds to the amino acid sequence of LH-RH/FSH-RH was synthesized using N-Boc chemistry by the method described by Stewart and Young (38), which starts with N-Boc-Gly-resin ester (1.0 g; 0.35 mmol). The side-chain protections were as follows: Ser and Tyr were protected by Bn groups, Gin by a Np group, Arg by a Boc group.
NOC group. Coupling was achieved with DCC with the single exception that Gin was coupled by means of its p-nitrobenzyl ester. Stepwise synthesis was carried out in the DCM and/or DMF using a glass shaker at room temperature. The removal of the N-Boc groups in the first seven steps was performed by treatment with 50% TFA in DCM for 20 minutes. After the incorporation of the Trp residue, 1N HCl in acetic acid that contained 1% 2-mercaptoethanol was used for the removal of the N-Boc group as described by Marshall (58). The neutralization was carried out by shaking with 10% triethylamine in CHC13. In every DCC coupling step, 4 equivalents of N-Boc-amino acid (4 × 0.35 mmole) for every equivalent of the starting glycine resin ester was used in the presence of 4 equivalents of DCC for 5 hours. An additional reaction with the same reagents was performed for another 5 hours. For the incorporation of His, 5 equivalents of N-Boc-His were used, and the same reaction was carried out twice as above. Nitosopropyl ester coupling of N-Boc-Gln was performed with 10 equivalents of the active ester for 5 hours, followed by an additional treatment for 5 hours in the presence of 5 equivalents of imidazole. The yield of the protected decapeptide resin ester was 1.38g (about 78% based on dry weight and amino acid analysis).

To achieve cyclization of N-terminal glutaminyl group on the resin to the pyroglutamyl ring after removal of N-Boc group, the N-Boc-glutaminyl-peptide resin ester was treated with 1N HCl in acetic acid that contained 1% of 2-mercaptoethanol for 1 hour at room temperature. The peptide was cleaved by stirring in 20mL of absolute methanol saturated with ammonia for 3 days at room temperature. After filtration, evaporation of the methanolic filtrate yielded 368 mg of the corresponding amide as a yellow semi-solid material. All protecting groups were removed before 1g of the peptide-resin conjugate was treated with a mixture of 14mL of HF, 1.5mL of anisole, and 0.25mL of methylethyl sulfide at 0°C for 0.5 hour and at 0°C for 0.5 hour. The HF was removed in vacuo at 0°C, and the resulting peptide and resin mixture was washed twice with ether and twice with CHCl3 and ether alternately. The peptide was extracted five times with 2M CH3COOH, and the extract was lyophilized.

The first purification was done by loading the lyophilized product on a column of Sephadex G-50 developed in 30% CH3COOH to remove the truncated fragments and salt. The next step of purification was by CM-32 carboxymethyl cellulose cation-exchange chromatography developed with a gradient generated by adding 2.5 volumes of 0.4M NH4OAC at pH 6.5 to 1 volume of 0.01M NH4OAC at pH 4.5. Final purification was achieved by partition chromatography on Sephadex G-50 using the solvent system 1-butanol:ethanol:acetic acid:water (15:10:3:12). 

Growth hormone releasing factor (GRF or GRH)

GRF has been isolated and characterized from a human tumor of the pancreas (61) as well as from rat (62), porcine (63), bovine (64), caprine (65), ovine (65), and human (65) hypothalamic stalk-median eminence. It is a 40-44-amino-acid peptide hormone produced in the arcuate nucleus of the hypothalamus and released from neurosecretory nerve terminals of these arcuate neurons, and it is carried by the hypothalamo-hypophyseal portal circulation to the anterior pituitary gland where it stimulates growth hormone (GH) secretion (66).


Ling et al. (65) have described the synthesis of hGRH in 1984 after they isolated and characterized the peptide from human hypothalamic tissues. The peptide was prepared by solid-phase methodology on a Beckman model 990 peptide synthesizer (Beckman Instruments, Fullerton, CA) using N-Boc chemistry. The MβHA resin (6g, 0.6mmol/g) was used, and the side-chain protections were as follows: Ser, Glu, Asp, and Thr were protected by Bn groups; Arg by a Tos group; Lys by a 2-Cl-Cbz group; Tyr by a 2,6-Cl-Bn group. The N-terminal amino acid was loaded to the MβHA resin using the symmetric anhydride method, and DCC was the coupling reagent. The subsequent amino acids were coupled according to the schedule in Table 6. After the last amino acid had been incorporated, the N-Boc protecting groups were removed before 1g of the peptide-resin conjugate was treated with a mixture of 14mL of HF, 1.5mL of anisole, and 0.25mL of methyl sulfide at 0°C for 0.5 hour and at 0°C for 0.5 hour. The HF was removed in vacuo at 0°C, and the resulting peptide and resin mixture was washed twice with ether and twice with CHCl3 and ether alternately. The peptide was extracted five times with 2M CH3COOH, and the extract was lyophilized. The first purification was done by loading the lyophilized product on a column of Sephadex G-50 developed in 30% CH3COOH to remove the truncated fragments and salt. The next step of purification was by CM-32 carboxymethyl cellulose cation-exchange chromatography developed with a gradient generated by adding 2.5 volumes of 0.4M NH4OAC at pH 6.5 to 1 volume of 0.01M NH4OAC at pH 4.5. Final purification was achieved by partition chromatography on Sephadex G-50 using the solvent system 1-butanol:ethanol:acetic acid:water (15:10:3:12). The chromatographic fractions were monitored by UV at 280nm and TLC on 0.25-mm-thick precoated silica gel 60 plates with the solvent system 1-butanol:pyridine:CH3COOH:water (6:6:12:4:8), and the spots were detected with ninhydrin spray. Four hundred twenty-six milligrams of hGRF was obtained after final purification with an overall yield of 2.3%.

Thyrotropin-releasing factor

TRF was isolated from porcine hypothalamus by Schally et al. (67, 68) and from ovine hypothalamus by Galli et al. (69) It is a tripeptide amide with the structure of pyroGlu-His-Pro-NH2.
Classic synthesis of the peptide has been reported in 1970 (70); solid-phase synthesis was envisaged 1 year later, and Rivaille et al. claimed they successfully prepared the target with a better yield (71). N'-Boc chemistry was used in the synthesis, and N'-Boc was introduced into the amino acid by reacting amino acids with Boc-azide at a constant basic pH. The His side chain was protected with an α-N-Boc chemistry was used in the precipitation method. A 300-mg portion of the crude octapeptide was purged through the fritted disk of the reaction vessel at 25 °C. The suspension was filtered, and the resin was washed 3 times with 10 mL portions of TFA. The filtrates were evaporated in vacuo. The syrupy product was then dissolved in acetic acid and lyophilized to give 1.47 g.

Liver peptide hormones

Angiotensin II

Angiotensin II is an octapeptide derived from biologically inactive decapeptide angiotensin I, which comes from angiotensinogen through the cleavage by the kidney enzyme renin. It is responsible for essential hypertension through stimulated synthesis and for the release of aldosterone from adrenal cells. The sequence of horse Angiotensin II is as follows: Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-OH.

Several syntheses of the peptide have been reported by solution methods (74–76). After the introduction of solid-phase peptide synthesis, Marshall and Merrifield conducted the first study of the synthesis of the peptide by using the new technique (77). N'-Boc chemistry was used, and Merrifield resin was selected as the solid support. The side chain protections were as follows: His, Arg, and Asp were protected by Bn groups; Arg by a NO2 group. The Phe was esterified onto the resin in ethanol with the presence of 1 equivalent of triethylamine. The symmetric anhydride method was used for the coupling of the amino acids, and DCC was the coupling reagent. The following cycle of reactions was used to introduce each new residue (Table 7): Amino acid analysis showed the average value of the eight amino acid residues to be 0.13 mmol/g of peptide resin or 0.16 mmol/g of unsubstituted copolymer. Removing the protecting groups was achieved by suspending the peptide resin in 20 mL anhydrous TFA and bubbling a slow stream of HBr through the fritted disk of the reaction vessel at 25 °C. The suspension was filtered, and the resin was washed 3 times with 10 mL portions of TFA. The filtrates were evaporated in vacuo. The product was redissolved in TFA and re-evaporated. The syrupy product was then dissolved in acetic acid and lyophilized to give 1.47 g.

Purification of the peptide was done by countercurrent distribution method. A 300-mg portion of the crude octapeptide was

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagents or solvents and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCM wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>55% TFA/5% 1,2-ethanediol in DCM wash (× 1)</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>DCM wash (× 3)</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>CH3OH wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>CH3OH wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>CH3OH wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>DMF wash (× 2)</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>DMF wash (× 1)</td>
<td>0.5</td>
</tr>
<tr>
<td>9</td>
<td>50% DMF-DCM wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>DMF wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Acetate unreacted Nα-amino group with 25% acetic anhydride in DMF</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>DMF wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>CH3OH wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>DMF wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>DMF wash (× 2)</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>DMF wash (× 2)</td>
<td>1</td>
</tr>
</tbody>
</table>

* The symmetric anhydride was formed with one equivalent amount of DCC. ** In the case of Asn and Gln, a 1.2 equivalents of HOBt were included.

Table 6 Schedule of events for assembling the peptide p-MBHA resin
purified by 100 transfers in a 1-butanol-acetic acid:water (4:1:5) system. Over 80% of the Sakaguchi-positive material was located in one peak which matched closely a theoretical curve with distribution constant, k = 0.30. The material in the peak was collected and the organic phase was removed by evaporation. The residual aqueous phase was removed by lyophilization; 193 mg of product was obtained, which was equivalent to a yield of 56%.

Pancreatic peptide hormones

Amylin

Amylin is a 37-residue peptide hormone first discovered independently by two research groups in 1987 (78, 79). Amylin is secreted by pancreatic β-cells at the same time as insulin (in a roughly 100:1 ratio), and it is the major component of diabetes-associated islet amyloid deposits. It inhibits basal and meal-stimulated glucose uptake as well as glycogen synthesis by skeletal muscles (80). Thus, amylin is also known as a diabetes-associated peptide.

The sequence of human amylin is as follows: Lys-c(Cys-A sn-Thr-Ala-Thr-Cys)-Ala-Thr-Gln-Ala-Leu-Ala-Asp-Phe-Leu-Val-His-Ser-Ser-Ser-Leu-Ala-Asp-Phe-Gly-Ala-Val-Ser-Ala-Asn-Thr-Tyr-NH₂.

One solid-phase synthesis of amylin has been reported in 1991 by Balasubramaniam et al. (81). Standard N-Boc chemistry was used for the synthesis, and MBHA resin was selected as the solid support. The coupling was done by using the symmetric anhydride method except for Arg, Asp, and Gln, which were coupled as their HOBt esters (refer to the section on "Amylin synthesis" for details). After completing the chain elongation, the peptide was cleaved from the resin using HF at 0°C. The residue was then oxidized with K₂Fe(CN)₆ to form the disulfide bond, followed by purification on semipreparative reversed phase column. The overall yield of the synthesis was between 10–20%.

Glucagon

Glucagon is a 29-residue polypeptide secreted by pancreas α-cells. It increases lipid mobilization and glycogenolysis to increase blood glucose levels. It was first discovered by K. I. M. Martin and Murlin in 1923 (82) when studying pancreatic extracts, but the sequence was not determined until 1957 (83).

The sequence of mammalian glucagons is as follows: His-Ser-Gln-Gly-Thr-Phe-Thr-Ser-Asp-Tyr-Ser-Leu-Asp-Ser-Arg-Ala-Gla-Gln-Glu-Val-Leu-Met-Ala-Ser-Thr-OH.

Early attempts to synthesize glucagons proved to be difficult because of the unusual structure, and successful examples were not reported until Wünsch's classic solution-phase synthesis (84). M. B. Browning and M. Wünsch reported the first stepwise solid-phase synthesis of glucagon in 1981, in which they used bifunctional (Boc) amino acid derivatives and t-Bu based side chain protections, which were unstable. They improved the method by using a native Boc strategy and benzyl based side chain protections by using their newly developed PAM resin in 1984, which gave a rapid, convenient synthesis (86). The aminomethyl-resin was prepared according to Mitchel et al. (87, 88). The first amino acid Thr was loaded to the resin through 2 steps: 1) N-Boc-Thr(Bzl)-OH was treated with powdered K₂Fe(CN)₆ and [4-(bromomethyl)phenyl]acetic acid phenacyl ester to produce N-Boc-Thr(Bzl)-OCH₂C₆H₄CH₂COOH and 2) product was then dissolved in DCM and allowed to couple with aminomethyl-resin at the presence of 1 equivalent of DCC, and unreacted amino groups were blocked by acetylation. The peptide chain was then elongated by the symmetric anhydride coupling method. In all, 8 eq. of amino acid and 4 eq. of DCC were used for each coupling, except for Leu and Gly in 1.6 eq. and 1.2 eq., respectively. A side chain protection was coupled with HOBt: Arg at positions 17 and 18 were coupled twice each. The side chain protections were as follows: Seer and Thr were protected by Bn groups. His and Arg by Tos groups; Asp by a cBz group; Tyr by a Bn-Cbz group; Lys by a Cbz-Cbz group; Trp by a Fmoc group. One synthetic cycle consisted of the following steps: 1) DCM wash,

### Table 7: Scheme of peptide assembling for angiotensin II

<table>
<thead>
<tr>
<th>Step</th>
<th>Reagents or solvents and operations</th>
<th>Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glacial acetic acid wash (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Deprotect with 1N HCl in glacial acetic acid</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Glacial acetic acid wash (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>C₂H₅OH wash, (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>DMF wash, (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Neutralize the HCl with 3mL of triethylamine in DMF</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>DMF wash, (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>DCM wash, (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Add 3.83 mmol of the appropriate BOC amino acid in 20mL of DCM and allowed to mix</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>Add 3.83 mmol of DCC to couple*</td>
<td>120</td>
</tr>
<tr>
<td>11</td>
<td>DCM wash, (× 3)</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>C₂H₅OH wash, (× 3)</td>
<td>1</td>
</tr>
</tbody>
</table>

* For the im-benzyl-L-His and nitro-L-arginine cycles, step 8 was deleted, and DMF was substituted for DCM in steps 9–11.
1 minute, × 3; 2) deprotection with 50% TFA/DCM wash, 1 minute prewash + 20 minutes; 3) DCM wash, 1 minute, × 6; 4) neutralization with 5% iPr2EtN/DCM wash, 1 minute, × 3; 5) DCM wash, 1 minute, × 6; 6) performed symmetric anhydride coupling in DCM (4 eq.), 60 minutes; 7) DCM wash, 1 minute, × 6; 8) neutralization with 5% iPr2EtN/DCM wash, 1 minute, × 3; 9) DCM wash, 1 minute, × 6; and 10) repeat steps 6 and 7).

After the completion of the peptide assemblies, the final deprotection and cleavage of the peptide was done in a Teflon vessel using the HF procedure (80). The crude peptide was then purified by preparative HPLC using C-18 column, the eluant from the major peak was collected and desalted on Sephadex G-10, thereby obtaining pure synthetic glucagon with 48% overall yield.

The following 29 cycles of coupling were completed by using the HF procedure (94). Three grams of N-\text{Boc}-A\text{la esterified to 8 grams of the supporting cross-linked polystyrene resin. The shorter time was adequate for complete debenzylation of His and Cys and complete detosylation of Arg. The deprotected triacontapeptide was converted to the S-sulfonate. On electrophoresis, a major Pauly-positive spot existed with the same mobility as that of the B-chain S-sulfonate (BSSO₃) obtained by sulfidolysis of natural bovine insulin, and a minor contaminant probably caused by B₁-27. The overall yield from the first A₁a residue was 21%.

The A₁ chain was synthesized by the automated solid-phase procedure (94). Three grams of N-\text{Boc}-A₁a resin was carried through 20 reaction cycles as described for the B chain except that the reagents used for all deprotection and neutralization steps were 4 M HCl in dry dioxane and triethylamine in CHCl₃, respectively (95). The peptide was cleaved from the resin as described above (yield, 69%). The total time required was 8 days. The S-benzyl protecting groups were removed by treatment with sodium in liquid ammonia as described by Niu et al. (93), except that the stable light blue end point was limited to exactly 15 seconds to prevent excessive cleavage of the Thr–Pro bond. Under these conditions, this cleavage was only 20–25%, whereas 80% was lost during a 60-second treatment. The shorter time was adequate for complete debenzylation of His and Cys and complete detosylation of Arg. The deprotected triacontapeptide was converted to the S-sulfonate. On electrophoresis, a major Pauly-positive spot existed with the same mobility as that of the B-chain S-sulfonate (BSSO₃) obtained by sulfidolysis of natural bovine insulin, and a minor contaminant probably caused by B₁-27. The overall yield from the first A₁a residue was 21%.

After the completion of the peptide assembly, the final deprotection and cleavage of the peptide was done in a Teflon vessel using the HF procedure (80). The crude peptide was then purified by preparative HPLC using C-18 column, the eluant from the major peak was collected and desalted on Sephadex G-10, thereby obtaining pure synthetic glucagon with 48% overall yield.

Insulin
Insulin is a polypeptide hormone produced by β-cells of the pancreas; it contains two disulfide bonded peptide chains: an A chain of 21 residues and an intra-disulfide bridge and a B chain of 30 amino acids. Insulin is well known for regulating carbohydrate metabolism as well as increasing glucose uptake and utilization.

The structure of insulin is shown in Fig. 11. Insulin became one of the most demanded drugs after finding its power to treat Type 1 diabetes in the 1920s. However, it was thought almost impossible to be chemically synthesized at that time after the structure was elucidated by Sanger in 1955 because it involved three disulfide bridges, complicated higher-level structures, and so on. But these complications did not stop it from becoming one of the hottest targets for synthetic chemists. Several independent syntheses of insulin have been reported in the 1960s (89–91) using classic peptide-synthesis methods, which usually required several months to finish, and yields were low. Merrifield et al. (92) reported a fast and efficient synthesis later on by employing solid-phase peptide synthesis technique that only took several days with good yields. N-\text{Boc} chemistry was used for the synthesis. The assembling of the fully protected insulin B chain started with 1.9 mmol of N-\text{Boc}-A₁a esterified to 8 grams of the supporting cross-linked polystyrene resin. The following 29 cycles of coupling were completed by using the symmetric anhydride method and DCC as the coupling reagent. The side-chain protections were as follows: Glu, Cys, Ser, Tyr, and His were protected by Bn groups; Lys by a Cbz group; Arg by a Tos group. Cleavage of the peptide from the resin was done by firstly bubbling HBr through a liquid ammonia as described above (yield, 69%). The total time required was 8 days. The S-benzyl protecting groups were removed by treatment with sodium in liquid ammonia (96), and the four cysteine residues were converted to the S-sulfonate (A5SO₃⁺) (over all yield, 37%).

The synthetic A- and B-chain sulfonates (ASSO₃ and BSSO₃) were combined with each other and with the complementary natural chains. The mixtures were first reduced to the thiol derivatives with thioglycolic acid at 25 °C by a modification of the method of Du and colleagues (93, 97, 98). The reduced chains were precipitated together at pH 3.8, washed, and then oxidized in air at pH 10.0 to form insulin. The molar ratio of ASSO₃:BSSO₃ at the beginning of reductions was 4:1. The reduced insulin preparations appeared as Pauly-positive spots with the same mobility as standard bovine insulin, and both synthetic insulin preparations appeared as Pauly-positive spots with the same mobility as bovine insulin.

Pancreatic polypeptide
Pancreatic polypeptide is a peptide hormone secreted by pancreatic polypeptide-producing cells in the islets of Langerhans.

![Figure 11](image-url) The structure of insulin.
in the pancreas. It consists of 36 amino acids and has a molecular weight about 4200 Da. It suppresses the pancreatic secretion and stimulates gastric secretion.

The sequence of human pancreatic polypeptide (hPP) is as follows: Ala-Leu-Glu-Pro-Pro-Tyr-Pro-Gly-Asp-Asn-Ala-Thr-Pro-Glu-Gln-Met-Ala-Gln-Tyr-Ala-Ala-Asp-Leu-Arg-Arg-Tyr-Ile-Asp-Met-Leu-Thr-Arg-Pro-Arg-Tyr-NH₂

Regular solid-phase synthesis of hPP has been reported by Meyers and Czy (99). The peptide was assembled on BHA resin using N-BOC chemistry. Amino acids were coupled as symmetric anhydrides with DIC except for Asn and Gln, which were coupled as HOBT ester. The side-chain protections were as follows: Ser and Thr were protected by Bn groups; Glu and Asp by C1-Bn groups; Arg by a Tos group; Tyr by a 2-Br-Cbz group.

Deprotection of the N-BOC groups was done via treatment with 25% and 50% TFA in DCM. Couplings were monitored by the ninhydrin test, and they were repeated if not complete after 1 hour (refer to the sections on "Motilin synthesis" and "Secretin synthesis" for details). Free amino group found after double couplings were acetylated by acetylimidazole in DCM. After completing the synthesis, the peptide resin was treated with HF in the presence of 37% TFA to deprotect the resin and precipitate the peptide. The resulting peptide material was purified on a column of Sephadex G-50, which was followed by a column of CM-cellulose; fractions from the major peak were collected and lyophilized. The resulting peptide material was purified on a column of Sephadex G-25, which was followed by a column of CM-cellulose; fractions from the major peak were collected and lyophilized. The resulting peptide material was purified on a column of Dowex 1-X4 (acetate form), lyophilized, followed by chromatography on Sephadex G-10 in 1 N acetic acid in which only one peak of 346 mg peptide material was obtained. It was then applied on a column of Dowex 1-X4 (acetate form), lyophilized, followed by chromatography on Sephadex G-10 in 1 N acetic acid in which only one peak of 346 mg peptide material was obtained. It was then

Tyr-Ile-Asp-Met-Leu-Thr-Arg-Pro-Arg-Tyr-NH₂

Naturally Occurring Peptide Hormones and Neurotransmitters. Synthesis of

Adrenocorticotropic hormone (ACTH)

ACTH is a 39-amino acids polypeptide hormone cleaved from a precursor peptide called pro-opiomelanocortin, which is released from corticotropic cells of the anterior pituitary gland. ACTH acts through the stimulation of cell surface ACTH receptors, which are primarily located on the adrenocortical cells. It stimulates the cortex of the adrenal gland and increases the synthesis of corticosteroids. ACTH is also related to the circadian rhythm in many organisms.

The sequence of human ACTH is as follows: Ser-Tyr-Ser-Met-Glu-His-Phe-Arg-Trp-Gly-Lys-Pro-Val-Gly-Lys-Lys-Arg-Pro-Val-Lys-Val-Tyr-Pro-Asn-Gly-Ala-Glu-Asp-Glu-Ser-Ala-Glu-Ala-Phe-Pro-Leu-Glu-Phe-OH.

The total synthesis of ACTH attracted many researchers in the 1960s and 1970s. Schwyzer and Sieber (100) first synthesized porcine ACTH by using a conventional solution-phase peptide synthesis method. After the development of solid-phase synthesis by Merrifield (92), Yamashiro and Li (101) extended the scope of how to prepare this important biomolecule. N-BOC chemistry was used throughout Yamashiro and Li’s synthesis. The remaining 37 cycles of couplings were accomplished by using a regular DCC coupling procedure, except that the deprotection of the Trp N-BOC group was achieved by treating with 25% TFA in DCM for 30 minutes followed by treating with 50% reagent for 6 minutes. The side-chain protections were as follows: Asp, Ser and Glu were protected by Bn groups; His by a Boc group; Arg by a Tos group; Lys by a 2-Br-Cbz group; Tyr by a 2,6-Cl-Bn group (refer to the sections on “Somatostatin synthesis” and “Angiotensin II synthesis" for details). After completing the synthesis, a portion (1.00 g) of the peptide resin was cleaved and deprotected with HF in the presence of anisole. The resulting peptide material was purified on a column of Dowex 1-X4 (acetate form), lyophilized, followed by chromatography on Sephadex G-10 in 1 N acetic acid in which only one peak of 346 mg peptide material was obtained. It was then

Figure 12 Carboxymethylcellulose chromatography of crude synthetic ACTH. Taken from Yamashiro et al. (101).
subjected to gel filtration on Sephadex G-25, which gave a major peak of 202-mg peptide material. The material was treated with dithiothreitol in 0.1 N acetic acid for 21 hours at 50°C to convert any Met-sulfoxide to Met and then chromatographed on carboxymethylcellulose (Fig. 12) to give a major peak of 68-mg peptide material, which was very similar to that of the natural ACTH. This 32 mg of the purified material was purified even more by partition chromatography on Sephadex G-50 with a solvent system of 1-butanol/pyridine:0.1% aqueous acetic acid (5:3:11) to give 12-mg highly purified synthetic hACTH with almost the same RF as natural hACTH (Fig. 13).

Melanocyte-stimulating hormones (melanotropins, MSHs)

MSHs consist of three peptide hormones: α-MSH, β-MSH, and γ-MSH, which are secreted by intermediate lobe of the pituitary gland. They are cleaved from the same precursor peptide as ACTH. Their basic function is stimulation of melanocytes to darken skin and stimulation of melanin synthesis to darken the skin and hair. They also have been found to be released in the brain affecting appetite, sexual arousal, and many other functions.

The sequence of α-MSH is as follows: N\(^{-}\)Acetyl-Ser-Tyr-Ser-Met-Glu-His-Phe-Arg-Trp-Gly-Lys-Pro-Val NH\(_2\).

The sequence of human β-MSH is as follows: Ala-Glu-Lys-Asp-Glu-Gly-Lys-Pro-Tyr-Arg-Met-Glu-His-Phe-Arg-Trp-Gly-Ser-Pro-Pro-Lys-Acetyl-OH.

The sequence of porcine β-MSH is as follows: Asp-Glu-Gly-Pro-Tyr-Lys-Met-Glu-His-Phe-Arg-Trp-Gly-Ser-Pro-Pro-Lys-Acetyl-OH.

The sequence of γ-MSH is as follows: Tyr-Val-Met-Gly-His-Phe-Arg-Trp-Asp-Arg-Phe-Gly-OH.

Solid-phase synthesis of α-β-MSH were achieved in our group in 1980 (103), both using N\(^{-}\)Boc chemistry. In the preparation of α-MSH, amino acids were coupled successively to valine-\(\text{BHA}\) resin with threefold excess of N\(^{-}\)Boc amino acid and 2.4-fold excess of DCC in 1-15 hours. Removal of the N\(^{-}\)Boc protecting group was achieved by a treatment of 45% TFA in DCM that contained 2% anisole. The side-chain protections were as follows: Ser and Asp were protected by Bn group, Tyr by a 2,6-Cl-Bn group; Lys by a 2,6-Cl-Cl2 group; Arg and His by Tos groups; Trp by a For group. After completing all coupling cycles, the amino terminal end of the peptide was acetylated with threefold excess of N\(^{-}\) acetylimidazole (refer to the sections on "Motilin synthesis" and "Secretin synthesis" for details). The peptide was cleaved from the resin, and all protecting groups were removed with anhydrous liquid HF with the exception of the For group on Trp. The residue was purified by gel filtration on Sephadex G-35, followed by deacylation with 4N NaOH at pH 11.5 for 3 minutes. The reaction was quenched by additional glacial AcOH to a final pH of 4.5. Cation exchange chromatography was used for the purification of α-MSH, followed by gel filtration on Sephadex G-25. Porcine β-MSH was prepared similarly but on Asp-Merrifield resin.

The total synthesis of γ-MSH was reported by Ling et al. (104). N\(^{-}\)Boc chemistry was used for the synthesis. The coupling was started from loading N\(^{-}\)Boc-Gly to the Merrifield resin. The rest of the amino acids were coupled subsequently to the resin using DCC as coupling reagent for 1 or 2 hours, except 8 hours for Asn, which was coupled as Np ester. After each coupling was done, the unreacted amino group was blocked by acetylation with Ac\(_2\)O. Peptide was cleaved and deprotected by a treatment of HF with presence of anisole and p-methoxyethyl sulfide. The residue was purified by cation-exchange chromatography on CM-32 carboxymethyl cellulose and followed by gel filtration on Sephadex G-25. The overall yield of the synthesis was 10.3%.

Oxytocin

Oxytocin holds a special place in the development of modern hormone and neurotransmitter chemistry and biology. It was the first peptide hormone (neurotransmitter) to be isolated, its structure determined, and prepared by total synthesis. Oxytocin is a 9-amino-acid poly peptide made in magnocellular neurosecretory cells of the hypothalamus, and is released into the blood from the posterior lobe of the pituitary gland. It causes uterine contraction and milk ejection in lactating women as well as facilitates birth and breastfeeding. It is also involved in social recognition and bonding, aspects of sexual function, and might be involved in the formation of trust between people.
The sequence of oxytocin is as follows: c(Cys-Tyr-Ile-Glu-Asp-Cys)-Pro-Leu-Gly-NH₂. The synthesis of oxytocin has drawn significant interest to chemists. It was first made by du Vigneaud et al. (105, 106), and many other syntheses have been reported since then in 1950s, 1960s, and beyond. One of the most high-yield and rapid syntheses was done by Manning (107). The protected nonapeptide was synthesized in a stepwise manner using N-Boc chemistry. The symmetric anhydride method was used for the couplings, and DCC was the coupling reagent, except that Asn and Gin were coupled as their p-nitrophenyl esters. The side-chain protections were as follows: Trp and Cys were protected by Bn groups. Final Cys was coupled as N-benzylcarbonyl-S-benzyl derivative. After completing the synthesis, the dried resin was weighed, giving a yield of 81%. The protected nonapeptide was cleaved from the resin by ammonolysis, which was achieved by suspending the resin in anhydrous methanol and bubbling with a stream of ammonia from a refluxing solution of dry ammonia for 2.5 hours, at 4°C for overnight, and at 23°C for 2 hours. The overall yield of the protected nonapeptide was 59% based on the amount of Gly originally esterified to the resin. Sodium and liquid-ammonia reduction was used to remove the Bn protecting groups on Cys and Tyr and the disulfide bond was formed by K₂Fe(CN)₆ oxidation (108). The residue was then purified by gel filtration on Sephadex G-15, and the pure oxytocin was obtained in an overall yield of 27%.

Vasopressin

The presence of vasopressin has been known for more than 100 years when people found pressor activity in extracts of the posterior lobe of the pituitary gland (109). However, it took almost a half-century before du Vigneaud and his colleagues figured out what it is (110, 111). Aagine vasopressin is a 9-amino-acid human hormone that is responsible of reducing plasma volume and increasing plasma osmolality by inducing the kidneys to conserve water.

The sequence of arginine vasopressin (porcine) is as follows: c(Cys-Tyr-Phe-Glu-Asp-Cys)-Pro-Arg-Gly-NH₂. The sequence of oxytocin is as follows: c(Cys-Tyr-Ile-Glu-Asp-Cys)-Pro-Leu-Gly-NH₂. The structure of human relaxin is shown in Fig. 14. Tregear and colleagues (115–119) have done many studies on the synthesis of relaxin, but results were mainly published in meeting papers, which do not provide synthetic details. An interesting total synthesis of human relaxin was reported by Bülesbach and Schwabe (120). The highlight of this synthesis was their use of cysteine-protecting groups in the selective synthesis of the three disulfide links. The most stable disulfide link, the intrachain disulfide loop between A10 and A15, was synthesized first. The liberation of the cysteine side chain in position A24 and the subsequent thiolysis of the S-activated cysteine in position B23 directed the formation of the interchain disulfide bond A24-B23. The third disulfide bond was formed by oxidative removal of the corresponding protecting groups in positions A11 and B11.

N-\textsuperscript{Fmoc} solid-phase chemistry was used for the synthesis of the K chain. HM P resin was selected as solid support, and standard DCC/HOBt coupling protocol was used to elongate the peptide chain. The side-chain protections were as follows: Glu, Asp, Tyr, Ser, and Thr were protected by iBu groups; His by a Trt group; Lys by a Boc group; Arg by a Fmoc group. The Cys side chains were protected by S-Trt groups in positions A10 and A15, by an Acm group in A11, and by a methyl-Bn group in A24. Deprotection of the N-\textsuperscript{Fmoc} groups was done in 20% piperidine in DMF. After finishing the peptide chains synthesis, the peptide was cleaved from resin and deprotected with TFA, using thiophenol as a scavenger to yield an A chain with two free thiol groups (A10, A15) and two differently...
protected cysteines. The formation of the intrachain disulfide loop was achieved by titration with iodine in 50% acetic acid. After purification with HPLC, the chain was stored in this form and the methyl-Bn group was removed prior to reaction with the corresponding B chain by HF, and the resulting A chain that contains one thiol group (A24), one disulfide link (A10/A15), and one Ac group (A12) was about 95% pure. The B chain was synthesized using N-Boc chemistry. PM resin was used as the solid support, and standard DCC/HOBt coupling protocol was used to elongate the peptide chain. The side chain protections were as follows: Glu, Ser, Thr, and Asp were protected by Bn groups; Arg by a Tos group, Lys by a Cl-Cbz group, Trp by a For group; Met was protected as a sulfoxide (refer to the section on "Glucagon’s synthesis"

details). Cys was protected by an Acm group in position B11 and by a 2-nitropyridinesulphonyl group in position B23. After completing the synthesis, the peptide was cleaved and partially deprotected with HF, followed by purification on Sephadex G-50. Trp(formyl) and Cys(Nou) remained untouched.

The formation of the interchain disulfide link A24/B23 was achieved in 8 M guanidinium chloride at pH 4.5 using monothiol-A chain and thiol-activated B chain in a molar ratio of 1:1.3 for 24 hours at 37 °C. The third disulfide bond was synthesized by oxidation with iodine in aqueous AOH. Final deprotection of the partially protected relaxin was achieved by reducing Met-sulfoxide in aqueous TFA that contained NH₄I for 24 hours at 37 °C. Met sulfoxide was also reduced during the cleavage (as Met was oxidized during the synthesis). The deprotected crude peptide was then oxidized with trans-[P(η5-C₅H₅)₂Cr₂]²⁺ to form the intramolecular disulfide bond and to obtain human calcitonin.

Calcitonin gene-related peptide (CGRP)

Human CGRP (hCGRP) is a 37-residue polypeptide isolated by Morris et al. (127) from human medullary thyroid carcinoma. It acts as a vasodilator and stimulates the formation of cAMP. The sequence of hCGRP is as follows: Ala-c(Cys-Asp-Thr-AlaThr-Cys)-Val-Thr-His-Arg-Leu-Ala-Gly-Leu-Leu-Ser-Arg-Ser-Gly-Gly-Val-Val-Lys-Ala-Asn-Phe-Val-Pro-Thr-Asn-Val-Gly-Ser-Ser-Lys-Ala-Phe-NH₂.

The peptide has been made using a conventional solution-phase method (128). Improved synthesis on solid support has been reported later by Smith et al. (129), who have prepared hCGRP using N-Boc chemistry and MBHA resin as solid support. All amino acids were coupled as HOBT active ester by using DCC/HOBt, the deprotection of the N-Boc groups was done by treating with 30% TFA in DCM. The side-chain protections were as follows: Asp, Thr, and Ser were protected by Bn groups; His by a Bom group; Arg by a mesitylene-2-sulfonyl group; Lys by a 2-CI-Cbz group. When the coupling yield was greater than 99%, the unreacted amino groups were acetylated with Ac₂O; when the yield was lower than 99%, another coupling cycle was performed before acetylation (refer to the section on "Corticotropin-releasing factor synthesis" for details). After completing the synthesis, the peptide was cleaved with TFA, and the disulfide bond was formed by oxidation with K₂FeCN₅ in NH₄OAc solution. The crude hCGRP was purified by gel filtration on a column of Bio-Gel P-6, followed by ion-exchange column chromatography on CM-Sephadex C25. Fractions that correspond to the hCGRP were collected and lyophilized, and final purification was achieved on HPLC using a Vydac C18 semi preparative column, obtaining pure hCGRP with 4.4% overall yield.

Peptide Neurotransmitters

Opioids

Dynorphin

Dynorphins are a class of endogenous opioid peptides produced by many different populations of neurons. They function primarily as µ- and δ-opioid receptor agonists; therefore, they may act as antidepressants in the brain. However, they also have µ- and δ-opioid receptor agonist activities. Dynorphin A, which is a 17-residue peptide, and Dynorphin B, which is a 13-residue peptide, are found in nature.

The sequence of dynorphin A (1-17) is as follows: Tyr-Gly-Gly-Phe-Leu-Arg-Leu-Arg-Pro-Lys-Leu-Lys-Tyr-Arg-Asp-Asn-Gln-OH.

The sequence of dynorphin B (1-13) is as follows: Tyr-Gly-Gly-Phe-Leu-Arg-Leu-Arg-Pro-Lys-Leu-Lys-Tyr-Arg-Asp-Asn-Gln.

Dynorphin A with a diiodotyrosine at position 1 was synthesized by using N-Boc solid-phase chemistry (130), and many other syntheses of the peptide and its analogs also were reported. An optimized N-Boc chemistry synthesis method by Lemr (131) achieved a 94% yield.
of [8-Ala]-dynorphin, which is a dynorphin-A analog particularly potent in an opiate receptor-binding assay (131), was reported by Solé and Barany (132). Fmoc-PAL-Nle-MBHA resin was selected as the solid support, which has a tris(alkoxy)-benzylamide linker that acts as side-chain anchor to prevent alkylation of the deprotected protecting groups to the peptide residues such as Cys, Met, Tyr, and Trp. Coupling was achieved by using DIC/HOBt as the coupling reagent, and the deprotection of the N'-Fmoc groups was done by treatment with 20% piperidine in DME. After completing the synthesis, several cleavage cocktails were tried, and the best one turned out to be TFA with phenol, water, and trisopropylsilane (88:5:5:2), which gave more than 95% cleavage yield. The crude peptide was then purified by RP-HPLC, giving pure [8-Ala]-dynorphin A in 58% overall yield.

Synthesis of dynorphin B also has been reported by using fragment condensation in solution-phase (133).

Endorphin

Endorphins are endogenous opioid peptides produced by the pituitary gland and the hypothalamus in vertebrates. They are known as "natural pain killers" because of their analgesic effect. The best-known endorphins are α-, β-, and γ-endorphin, of which β-endorphin seems to be most implicated in pain relief.

The sequence of β-endorphin is as follows: Tyr-Gly-Gly-Phe-Leu-OH.

The sequence of human α-endorphin is as follows: Tyr-Gly-Phe-Met-Thr-Ser-Glu-Ser-Gln-Thr-Pro-Leu-Val-Thr-Leu-OH.

The sequence of γ-endorphin is as follows: Tyr-Gly-Phe-Met-Thr-Ser-Glu-Lys-Ser-Gln-Thr-Pro-Leu-Val-Thr-Leu-Phe-Lys-Ala-Ile-Ile-Lys-Ala-Tyr-Lys-Gly-Glu-OH.

Many syntheses of β-endorphin have been achieved in the past, either in solution or on solid-phase. For example, an N'-Fmoc synthesis on solid-phase was reported by Atherton et al. (134). p-Alkoxybenzyl ester resin was selected as the solid support; coupling was done by using the symmetric anhydride and HOBt active ester method, except that Asn and Gin were coupled as p-nitrophényl esters. The N'-Fmoc groups were deprotected by 20% piperidine in DMF. After completing the synthesis, the peptide was cleaved from the resin, and the side-chain protection groups were removed with anhydrous TFA in the presence of excess Met. The peptide was purified by CM-S2 chromatography, with an overall yield of 43%.

α-γ-Ly-Endorphin have been synthetically prepared by Ling (135). An N'-Boc strategy was used for the synthesis. In the case of α-endorphin, N'-Boc-Thr(ΩBn) was loaded on a chloromethyl-resin. The rest of the amino acids were coupled to the resin by using DCC as coupling reagent, and the deprotection of the N'-Boc groups was done using 50% TFA in DCM. After incorporation of the Met, the TFA deprotection step was modified with the inclusion of 5% 1,2-ethanediol. The side-chain protections were as follows: Thr, Ser, and Gin were protected by Bn groups; Tyr by a 2,6-C1-Bn group, Lys by a 2-C1-Cbz group. After the completion of the chain elongation, the peptide was cleaved by HF with 10% anisole at -5°C. The crude peptide was purified by using a CM-32 carboxamidomethyl cellulose column first, followed by gel filtration on Sephadex G-25F column and partition chromatography on a Sephadex G-25F column. The pure α-endorphin was obtained with an 18% overall yield. γ-Endorphin was prepared in the same manner with a 30% overall yield.

Enkephalin

Enkephalins are endogenous opioid peptides that have analgesic effects. Met-enkephalin and Lys-enkephalin are found in nature, and they are pentapeptides that share the same first four residues.

The sequence of Leu-enkephalin is as follows: Tyr-Gly-Gly-Phe-Leu.

The sequence of human α-ENDorphin is as follows: Tyr-Gly-Phe-Met-Thr-Ser-Glu-Lys-Ser-Gln-Thr-Pro-Leu-Val-Thr-Leu-Phe-Lys-Ala-Ile-Ile-Lys-Ala-Tyr-Lys-Gly-Glu-OH.

Many syntheses of α- and β-endorphin have been reported. In addition, syntheses of these peptides have employed enzymatic methods, which could be done in low-water content systems on a preparative scale (138). N'-Cbz and N'-Boc protected Met-enkephalin and Leu-enkephalin were prepared by means of α-chymotrypsin, papain, thermolysin, and bromelain adsorbed on Celite. α-Chymotrypsin has a primary specificity for bulky and hydrophobic amino acids (Phe, Tyr, ...
Tripeptide Phe-Leu-Met, and it was selected for making the Tyr-Gly-Phe-Leu, and Phe-Met peptide bonds. Papain was selected for Gly-Gly and Gly-Phe peptide bonds. Bromelain is very similar to papain as far as its specificity. All these proteases are serine or cysteine type, and the peptide bond formation can be done under kinetically controlled conditions. Thermolysin is an aspartyl protease, and it was selected mainly for Phe-Leu and Gly-Phe bonds. MECN, EDTA, and metal caprate were used as solvents with a controlled amount of buffer or at fixed water amount.

The synthetic scheme is shown in Fig. 15 (138), which was based on 4 + 1 enzymatic fragment condensation. N -Cbz-Tyr-Gly-Gly-Phe-Leu-NH₂ and N -Boc-Tyr-Gly-Gly-Phe-Met-NH₂ were prepared in good yield.

**Leumorphin**

Leumorphin is a 29-amino-acid peptide first predicted by Yamanoto et al. (139, 140) when sequencing the cloned DNA complementary to porcine hypothalamic mRNA. They synthesized this peptide and found that it has potent opioid activity as predicted (140), and they also proved the existence of this peptide in humans (141).

The sequence of porcine leumorphin is as follows: Tyr-Gly-Gly-Phe-Leu-Arg-Ang-Gln-Phe-Lys-Val-Thr-Arg-Ser-Gln-Glu-Asp-Pro-Asn-Ala-Tyr-Tyr-Glu-Glu-Leu-Phe-Asp-Val-OH.

Portcine leumorphin was synthesized by a solid-phase technique with N -Boc strategy. The peptide chain was elongated on a chloromethylated resin with use of a peptide synthesizer (Beckman Model 990B). The side-chain protections were as follows: Asp, Glu, Thr, and Ser were protected by Bn groups; Tyr by a 2,6-Cl-Bn group; Lys by a 2-CI-Bz group; Arg by a Tos group. After the completion of the synthesis, the peptide was deprotected and cleaved from the resin with HF in the presence of anisole and ethane dithiol (refer to the sections on "Somatostatin synthesis" and "Angiotensin II synthesis" for details). The crude peptide was purified by gel chromatography on Sephadex G-10, followed by reverse-phase HPLC.

**Tachykinin**

Tachykinins are peptides characterized by the C -terminal amidated pentapeptide Phe-Gly-Leu-Met-NH₂

**Neurokinin A and B**

Neurokinin A and B are both decapeptides discovered by Kimura et al. (142) from the porcine spinal cord. Both have very similar amino acid composition and sequence homology to the mammalian mammalian bombesin as it has an identical C- terminal decapeptide with the exception of one residue as bombesin, which gives GRP similar biological activity as bombesin. It lowers body temperature and increases plasma levels of epinephrine and glucose (148).

The sequence of GRP is as follows: Ala-Pro-Val-Ser-Val-Gly-Gly-Gly-Thr-Leu-Ala-Lys-Tyr-Pro-Arg-Gly-Asn-His-Trp-Ala-Val-Gly-His-Leu-Met-NH₂.

Synthesis of this peptide has been achieved by Matki et al. (149). N -Boc solid-phase synthesis techniques were employed, and MBHA resin was selected as the solid support. All the amino acids were coupled as symmetric anhydrides using DCC as the coupling reagent except for Asn, which was coupled as N -Boc-Asn-OPfp ester with HOBt. After completing each residue for one or two coupling cycles, unreacted amino groups were acetylated. The side-chain protections were as follows: His and Arg were protected by Tos groups, Thr and Ser by Bn groups; Tyr by a 2,6-Cl-Bz group; Lys by a Cl-Cbz group. After the completion of the synthesis, the crude peptide was cleaved from the resin and the side-chain protection groups were deprotected by treating with HF (refer to the section on "Corticotropin-releasing factor synthesis" for details). Gel filtration on Sephaose G-25 was applied as the first purification of the peptide, followed by semi-preparative reverse phase HPLC and lyophilization to give the TFA-peptide salt in a 5% overall yield.

**Neurotensin**

Neurotensin is a tripeptide discovered by Carraway and Leeman (150) from bovine hypothalam in 1973. Neurotensin is distributed mainly in the central nervous system and in some regions of the digestive tract in various mammals. It has a broad spectrum of biological activities, including kinin activities such as a solid support, and the amino acids were coupled to the resin using the symmetric anhydride method; DCC was the coupling reagent. The side-chain protections were as follows: Asp, Ser, and Thr were protected by Bz groups; His by a Tos group; Lys by a C-Cbz group (refer to the sections on "Motilin synthesis" and "Secretin synthesis" for details). After the completion of the synthesis, the peptide was cleaved and deprotected by HF, and the crude peptide was purified by gel filtration on Sephaose G-25. HPLC elution was applied as the second purification, and followed by desalting on a column of Sephaose G-10, obtaining more than 98% pure peptides.

**Eledoisin, physalaemin, and kassinin**

These peptides have only been reported and prepared by traditional solution-phase synthesis (144–146). However, because their structures are not complicated, they can be readily prepared by modern solid-phase techniques.

**Others**

**Gastrin-releasing peptide (GRP)**

Gastrin-releasing peptide is a 27-amino acid-containing peptide first isolated and sequenced by McDonald et al. (147) from porcine gastric mucosa. As the name itself implies, it stimulates the release of gastrin from gastric mucosa. It is also called mammalian bombesin as it has a similar biological activity as bombein, which gives GRP similar biological activity as bombesin. It lowers body temperature and increases plasma levels of epinephrine and glucose (148).

The sequence of GRP is as follows: Ala-Pro-Val-Ser-Val-Gly-Gly-Gly-Thr-Leu-Ala-Lys-Tyr-Pro-Arg-Gly-Asn-His-Trp-Ala-Val-Gly-His-Leu-Met-NH₂.

Synthesis of this peptide has been achieved by Matki et al. (149). N -Boc solid-phase synthesis techniques were employed, and MBHA resin was selected as the solid support. All the amino acids were coupled as symmetric anhydrides using DCC as the coupling reagent except for Asn, which was coupled as N -Boc-Asn-OPfp ester with HOBt. After completing each residue for one or two coupling cycles, unreacted amino groups were acetylated. The side-chain protections were as follows: His and Arg were protected by Tos groups, Thr and Ser by Bn groups; Tyr by a 2,6-Cl-Bn group; Lys by a Cl-Cbz group. After the completion of the synthesis, the crude peptide was cleaved from the resin and the side-chain protection groups were deprotected by treating with HF (refer to the section on "Corticotropin-releasing factor synthesis" for details). Gel filtration on Sephaose G-25 was applied as the first purification of the peptide, followed by semi-preparative reverse phase HPLC and lyophilization to give the TFA-peptide salt in a 5% overall yield.

**Neurotensin**

Neurotensin is a tripeptide discovered by Carraway and Leeman (150) from bovine hypothalam in 1973. Neurotensin is distributed mainly in the central nervous system and in some regions of the digestive tract in various mammals. It has a broad spectrum of biological activities, including kinin activities such
as inducing hypotension, contraction of ileum and uterus, and relaxation of the duodenum; it also can cause a rapid increase of glucose levels in blood.

The sequence of neurotensin is as follows: Glu-Leu-Tyr-Glu-Asn-Lys-Pro-Arg-Pro-Tyr-Ile-Leu-OH.

Carrawy and Leeman also sequenced (151) and synthesized (152) this peptide. Standard solid-phase synthesis using N-\textit{Boc} chemistry was employed for the synthesis, and the Merrifield chloromethylated resin was selected as the solid support. Amino acids were coupled as symmetric anhydrides using DCC as the coupling reagent, except that Asn and Gln were coupled as \textit{p}-nitrophenyl esters. The side-chain protections were as follows: Glu and Tyr were protected by Bn groups; Lys by a \textit{Boc} group; Arg by a \textit{NO2} group. The deprotection of the N-\textit{Boc} groups was achieved by 50% TFA in DCM. After completing the synthesis, the peptide was cleaved and partially deprotected by treatment with HBr-TFA, and the remaining NO2 group was removed by hydrogenation (refer to the sections on "Somatostatin synthesis" and "Angiotensin II synthesis" for details). The crude peptide was first purified on Sephadex LH-20 column, followed by cation exchange chromatography on sulfoethyl-Sephadex column. Pure neurotensin was obtained with 7% overall yield.

**Bradykinin**

Bradykinin is a 9-residue peptide discovered by Silva et al. (153). It is found to be a potent endothelium-dependent vasodilator, to cause contraction of nonvascular smooth muscle, and to increase vascular permeability; it also is involved in the mechanism of pain.

The sequence of bradykinin is as follows: Arg-Pro-Pro-Gly-Asp-Val-Pro-Phe-Arg-Asp. Bradykinin has been prepared synthetically many times in the 1960s (154-156). The first solid-phase synthesis was achieved by Merrifield (157) using N-\textit{Boc} chemistry. More recent modifications of bradykinin synthesis were reported by Khar et al. (158), who also used N-\textit{Boc} solid-phase synthesis technique. Amino acids were coupled as active esters to the preloaded N-\textit{Boc}-Arg(\textit{NO2})-OCH2-resin. Phe and Gly were coupled as O-Tcp esters, Pro was coupled as O-Pro ester, and Ser was coupled as an O-N\textit{Ns}u ester. The side-chain protections were as follows: Ser was protected by a Bn group, Arg by a \textit{NO2} group. After completing the synthesis and before removing the last N-\textit{Boc} protecting group, the peptide resin was submitted to hydrogenation using Pd(OAc)2 as catalyst, which resulted in only the N-\textit{Boc} protected crude peptide. Deprotection of the N-\textit{Boc} groups with TFA followed by purification on carboxymethyl-cellulose column gave bradykinin in 20% overall yield.

**Summary and Conclusions**

The total synthesis of peptide natural products, especially hormones and neurotransmitters, is a highly developed area of organic synthesis. A plethora of peptides with numerous \(\beta\)-substituted amino acids, highly constrained peptides, highly hydrophobic peptides, and \(\text{N}-\text{alkyl substituted amino acids can present synthetically challenging, numerous synthetic strategies have been developed to overcome these problems. Furthermore, despite all the mythology about the applications of peptides as drugs for a wide variety of diseases, peptide drugs are becoming increasingly important in medicine because of their general lack of toxicity and their high efficacy especially for hormones and neurotransmitters. These trends are likely to accelerate into the future as numerous alternative methods of drug delivery are developed and optimized even more. This trend will be especially true for hormones and neurotransmitters, in which oral delivery is often a very poor choice for numerous reasons. Furthermore, the cost of peptide synthesis has decreased dramatically in recent years, and several companies have been developing increasingly efficient syntheses of peptides for use as pharmaceutical reagents in up to metric ton quantities.**

**As outlined in this article, several very effective solution- and solid-phase synthetic methods have been developed for the total synthesis of peptides. We have covered most synthetic strategies and tactics that have been developed and found to be useful. In general, successful peptide synthesis depends on careful attention to the details of the synthesis and careful analytical monitoring of the progress of the synthesis. If this care is taken, then successful synthesis of most peptides can readily be accomplished by a competent synthetic chemist.**

A detailed readings about peptide synthesis and its practical and mechanistic considerations can be found in many books, chapters in books and reviews. A good starting point for the interested person includes the items listed within the Further Reading section.

**References**


43. Frey WH & Co. San Francisco, CA.


121. Trifile Sl, Carey S. Calcitonin from the parathyroid which lowers the calcium level of the blood. Nature 1962;193:381–382.


Naturally Occurring Peptide Hormones and Neurotransmitters, Synthesis of


Further Reading


Springer-Verlag, New York.


Oxygen-Activating Enzymes, Chemistry of

Corinna R. Hess*, Richard W. D. Welford* and Judith P. Klinman†, Departments of Chemistry and Molecular Cell Biology, University of California, Berkeley, California
doi: 10.1002/9780470048672.wecb431

Aerobic organisms derive the energy required for cellular processes from the conversion of dioxygen to water, which highlights the importance of dioxygen chemistry in biologic systems. The biochemistry of dioxygen is far from simple and has been the subject of intense study in a range of chemical and biologic disciplines. The reduction of dioxygen is energetically favorable; however, dioxygen is a ground state triplet, kinetically unreactive with singlet organic molecules. Nature has developed a diverse array of catalysts to overcome this kinetic barrier. These dioxygen-activating enzymes are divided into two classes: oxygenases and oxidases. Oxygenases incorporate directly at least one atom from dioxygen into the organic products of their reaction. Oxidases couple the reduction of dioxygen with the oxidation of substrate. Typically, enzymes that react with dioxygen contain transition metal ions and/or conjugated organic molecules as cofactors. The reaction with dioxygen is initiated by electron transfer from the cofactor to O₂. The subsequent chemistry varies depending on both the nature of the cofactor and the protein scaffold. Here we review the fascinating chemistry of the dioxygen-activating enzymes and identify some of the common strategies and themes that have emerged from over half a century of research.

Biological Background

The introduction of an oxygen-rich atmosphere led to the evolution of enzymes capable of exploiting this diatomic gas for a variety of different chemistries. Viewing life today, we can conclude that the use of dioxygen ultimately turned out to be advantageous. Dioxygen is required by all aerobic forms of life for a variety of chemical transformations and biological processes, which are catalyzed by a rich and diverse set of protein catalysts (1). In addition to its crucial function in the respiratory chain, dioxygen also is involved in biosynthesis, signaling, xeno-biotic metabolism, DNA repair, and biodegradation. However, a darker side exists to dioxygen biochemistry. The reactive oxygen species (ROS) formed by reduction of dioxygen species have been linked to several detrimental processes such as aging and cancer (2). These ROS can damage proteins, cell walls, and DNA. One key area of interest is how dioxygen-activating enzymes use ROS without damaging their own peptide backbone in the process.

Dioxygen Chemistry

The full four-electron, four-proton reduction of O₂ to two molecules of H₂O is strongly exothermic (Table 1). However, most redox processes involve sequential one or two-electron transfer pathways. The first electron transfer to dioxygen is the most difficult step (Table 1) because of the loss in O–O bond strength associated with formation of superoxide (3). Most organic molecules do not possess enough reducing power to facilitate this initial reduction, and as a consequence, their one-electron oxidation by dioxygen is thermodynamically unfavorable. Moreover, dioxygen is a triplet in its ground state, whereas most organic molecules possess singlet ground states. Consequently, the direct reaction of dioxygen with organic molecules necessitates a spin conversion to the singlet excited state of O₂, which has 22.5 kcal/mol more energy.
strength upon reduction of O$_2$. The rich spectroscopy associated with metal–oxygen bonds compensate partly for the loss of bond torsion present in dioxygen and its many reduced forms. The resultant active sites, such as Mn, have been documented. Metallocofactor electrons to yield the final enzymatic products. Dioxygen products are highly reactive and readily accept additional electrons to yield the final enzymatic products.

Cofactors used in Dioxygen Activation

Practically all known oxygen-activating enzymes employ a cofactor in the form of a metal center, a highly conjugated organic molecule, or both. Ultimately, the reactions involve the reduction of O$_2$ by two or four electrons to generate hydrogen peroxide or water. Metallocofactors and flavin molecules are powerful redox centers, able to overcome the large potential associated with the formation of superoxide (Table 1), and with comparatively stable one-electron oxidized forms. The activated dioxygen products are highly reactive and readily accept additional electrons to yield the final enzymatic products.

Dioxygen activation by metalloenzymes is dominated by Fe and Cu, although a few examples of other transition metal active sites, such as Mn, have been documented. Metallocofactors bind dioxygen and its many reduced forms. The resultant metal-oxygen bonds compensate partly for the loss of bond strength upon reduction of O$_2$. The rich spectroscopy associated with metal sites has advanced our understanding of the chemistry of dioxygen-activating enzymes. Methods such as electron paramagnetic resonance spectroscopy (EPR), X-ray absorption spectroscopy (XAS), and M$\text{O}$$\text{S}$bau have been used to obtain structural information regarding the metallocofactors at various stages in the catalytic mechanisms (4).

Dioxygen-activating enzymes can be categorized according to the type of cofactor(s) employed. Although similar types of cofactors can catalyze different reactions, common structural features and reaction intermediates often exist for a given type of cofactor, as discussed below.

Iron-containing motifs for dioxygen activation

Iron-containing proteins are classified as heme, mononuclear non-heme, and binuclear non-heme enzymes. The Fe center in heme proteins is ligated by four nitrogens of a porphyrin molecule, in addition to one or two axial ligands (Fig. 1a). The reaction of Fe$^{II}$ with O$_2$ generates an Fe$^{III}$-superoxide, in which the reduced O$_2$ is coordinated at an axial position of the heme unit in a bent, end-on fashion.

Mononuclear non-heme iron centers with several different types of coordination environments have been identified (5–6). The most prevalent coordination is an octahedral or square pyramidal site with a 2His-1carboxylate ligand set coordinated to one face of an Fe$^{II}$ center (Fig. 2a). In the resting state, the other positions are occupied by water molecules, which can be readily substituted during the reaction cycle by the substrate, the co-substrate, or O$_2$. A notable group of mononuclear non-heme iron enzymes activate their organic substrate for direct reaction with dioxygen. These enzymes contain an Fe$^{II}$, in which the coordination environment differs by subfamily. For example, intradiol dioxygenases contain a trigonal bipyramidal Fe$^{III}$ ligated by 2His, 2Tyr, and a hydroxide in the resting state (Fig. 3a). The active sites of binuclear non-heme enzymes consist of two Fe atoms, separated by 3–4 Å and bridged by O-atoms derived from hydroxide or carbonylate residues (Fig. 3b). The iron centers can adopt 4-, 5-, or 6-coordinate geometries, with the bridging ligands bound via one or two O-atoms. The remaining coordination sites are occupied by His and Asp/Glu residues.

The reactions of iron-containing enzymes with O$_2$ often involve high oxidation states of the metal. Generally, the initial reaction of dioxygen with both heme and mononuclear non-heme ferrous enzymes results in the formation of Fe$^{III}$-superoxide intermediates. Highly reactive Fe$^{IV}=O$ intermediates often are employed often for C–H activation. The mechanism of substrate oxidation by binuclear non-heme enzymes involves high valent, o xo-bridged species, with Fe in the +3 or +4 oxidation state.

Copper-containing motifs for dioxygen activation

The copper-containing enzymes can be classified as mononuclear, binuclear, or multicopper proteins (9). Unlike iron, copper cannot readily access high oxidation states and the formation of Cu$^{III}$ in biological systems remains controversial. Generally, Cu$^{II}$ centers undergo one-electron oxidations to activate dioxygen. Thus, the Cu-containing enzymes tend to employ multiple copper atoms or an additional cofactor for the final two or four-electron oxidation of their substrates.

Mononuclear copper enzymes capable of dioxygen activation contain a Type 2 copper center (Fig. 1b). Generally, Type 2 copper sites have a tetragonal geometry in the Cu$^{II}$ state and can be identified by their EPR spectra (10). Reduction of the Cu site is accompanied by a loss of water molecules and a change in coordination geometry; trigonal or tetrahedral geometries are common for Cu$^{II}$ (11). Dioxygen is thought to bind to the reduced copper site in an end-on or side-on fashion, to yield a Cu$^{II}$-superoxide intermediate.

A Type 3 active site, which consists of two antiferromagnetically coupled Cu atoms, ~3 Å apart, is found in binuclear copper enzymes (10). Each copper is coordinated by three histidine residues and by one or two water molecules that serve...
as $\eta^2$ bridging ligands (Fig. 1f). These enzymes all bind to O$_2$ in an $\eta^2$ fashion, where each O atom is bound by the two copper atoms. Each copper atom transfers one electron to dioxygen to yield a CuII$_2$-peroxide intermediate (for example, “oxy state” in Fig. 3g). Proteins with Type 3 copper centers include dioxygen transport and dioxygen-activating enzymes.

The multicopper enzymes contain a binuclear metal cluster that consists of a Type 2 copper site and a binuclear, Type 3 copper site (10). The Type 3 site of multicopper enzymes is distinct from the active site of the coupled binuclear copper enzymes described above. Although the Type 3 Cu centers are antiferromagnetically coupled, the centers are separated by ~ 5 Å and are bridged by a hydroxide ligand. The multicopper oxidases also contain at least one mononuclear, Type 1, blue copper site in addition to the binuclear Cu cluster. The Type 1 Cu site is ligated by two histidines, a cysteine, and often a weakly coordinating axial 4th ligand. Blue copper sites also are found commonly in electron transfer proteins and are defined by their intense Cys-to-metal charge-transfer transition at ~ 600 nm (11).

**Dioxygen activation by flavins**

A large group of enzymes react with O$_2$ by using an organic flavin cofactor (Fig. 1g) (12, 13). The reactions can be divided into two half-reactions. First the flavin is reduced by substrate, and then the reduced flavin reacts with an electron acceptor, such as dioxygen. When O$_2$ is the electron acceptor, the first step in the oxidative half-reaction is the rate-limiting electron transfer, which leads to the formation of a caged radical pair of superoxide anion and flavin semiquinone. The fate of the radical pair depends on whether the enzyme is an oxidase or an oxygenase and will be discussed below. The second-order reaction of a reduced free flavin with O$_2$ in solution proceeds at a rate of $2.5 \times 10^2$ M$^{-1}$ s$^{-1}$. For an enzyme-bound flavin, the rate can vary between 2 M$^{-1}$ s$^{-1}$ and 10$^6$ M$^{-1}$ s$^{-1}$. It is not fully understood how enzymes with extremely similar active sites can access such a range of rates for reaction with O$_2$.

The following sections describe specific examples of oxygenase and oxidase chemistry catalyzed by the cofactors described above. It is by no means an extensive list, but it should offer the reader a flavor of the diverse mechanisms of dioxygen activation by enzymes.
Oxygenases

The enzymes in this section can incorporate either one or two atoms from dioxygen into the organic product(s) of their reaction. It is of note that many oxygenases also can catalyze oxidation-like reactions, such as deamination and ring closure. Figure 2 illustrates the reactions catalyzed by some of these enzymes, whereas Figure 3 shows partial reaction mechanisms, including key intermediates.

Iron oxygenases

Heme-Fe: The cytochromes P450

The cytochromes P450 are some of the most well-studied oxygenases enzymes; their oxidation reactions, typically hydroxylation, are important for xenobiotic metabolism and biosynthesis (18, 19). Several notable differences exist between P450s, which catalyze oxygenations, and the O2 transport globins, which simply bind O2 reversibly. P450s have an axial cysteine thiolate iron ligand on the proximal side of the heme, whereas the transport globins have a histidine residue (20). Additionally, P450s have a conserved GX(E/D)T sequence motif on the distal side of the heme, which is thought to be involved in the proton donation required for cleavage of O2. In the resting state, the P450s contain a heme-ligated FeII. Typically, the electrons for the reduction of O2 are supplied by NADPH via protein redox partners. Extensive spectroscopic work has led to the observation of several of the peroxo intermediates involved in the conversion of the Fe–oxygen complex, FeIII–O–O, to the compound I + H2O (Fig. 3a). Most recently, elegant one-electron cryo-reduction and EPR spectroscopy allowed the detection of the FeII–O–O− and FeII–O–OH− intermediates. The loss of water from the latter generates the highly oxidizing compound I, which is believed to be an FeIV=O species with radical character localized on either the porphyrin ring or thiol ligand. Compound I can abstract a hydrogen atom from an organic substrate. The substrate radical combines rapidly with the iron-ligated hydroxyl in a "radical rebound" mechanism (21). Several facets of P450 catalysis are underlying themes for some other oxygenase enzymes.

Mononuclear non-heme oxygenases

The mononuclear, non-heme iron enzymes can be divided into subfamilies, which differ in the cosubstrate that provides the electrons for the reduction of O2 and the relative positions of the 2His-1Carboxylate residues in the protein sequence (7). Even among members of the same subfamily, the only conserved residues are the 2His-1Carboxylate, which indicates the use of this motif for activation of O2. Despite the conserved ligand set, the chemistry can vary greatly. In the reaction cycle of the a-ketoglutarate (a-KG) dependent subfamily, it is the a-KG co-substrate that ligates the Fe, whereas the prime substrate occupies a second sphere position. The counter substrate is observed for the extradiol catechol dioxygenases where the "prime" substrate ligates directly to the iron. The largest subfamily in the 2His-1Carboxylate group of non-heme FeII-dependent enzymes depends on a-KG as a co-substrate (6). The substrate binding is ordered. First a-KG ligates the FeII, then the prime substrate binds adjacent to the metal site, and finally O2 binds to the metal (Fig. 3b). The ternary complex reacts intramolecularly to yield the oxidizing species. This was characterized recently as an FeIV=O by Møssbauer, Raman, and EXAFS spectroscopy, which is the first enzymatic example of a non-heme FeIV (14). As with P450s, the FeIV=O is thought to be the oxidizing intermediate in a variety of reactions, including hydroxylation, desaturation, and oxidative ring closure (6). The multifaceted, highly controlled nature of this chemistry is exemplified brilliantly by the enzyme clavaminate acid synthase, which uses a single active site to catalyze a hydroxylation, desaturation, and a ring closure at different stages in a single biosynthetic pathway. The a-KG-dependent enzymes, like other oxygenases, can hydroxylate amino acid side chains in the vicinity of the active site under certain conditions (7). This illustrates the enzymes' need to control the oxidizing intermediate to minimize such deleterious side reactions. In fact, mammals are thought to require vitamin C for the reduction of FeIII, generated as a by-product of unproductive reaction cycles of a-KG enzymes.

The extradiol dioxygenases, similar to their intradiol counterparts, cleave aromatic compounds, but the position of ring opening differs (Fig. 2, 17). In the resting state, the extradiol dioxygenases contain a 2His-1Carboxylate, five-coordinate FeII. Bidentate binding of the catechol to the Fe activates the metal for O2 binding. This process is analogous to the a-KG enzymes.
Oxygen-Activating Enzymes, Chemistry of

*Figure 2* Some common reactions carried out by oxygenase enzymes. A * represents an oxygen atom derived from O₂. The reactions listed are those thought to be the enzyme’s biologically relevant reaction. Where appropriate the name of the enzyme catalyzing the example reaction is given.

except an additional substrate is not required to make the center five-coordinate. The Fe³⁺ superoxide formed on dioxygen binding can induce radical character in the catechol that leads eventually to ring opening.

Another subgroup of the 2His-1carboxylate family is dependent on a reduced pterin cofactor (5). They catalyze hydroxylations at the aromatic positions of amino acids in phenylalanine catabolism and hormone biosynthesis (*Fig. 2*). Unlike the α-KG-dependent enzymes, the pterin co-substrate does not ligate to the iron directly. In the reaction cycle, the pterin co-substrate supplies two electrons for the heterolysis of O₂ to give a yet to be characterized iron-oxygen hydroxylating species.

X-ray crystallography and magnetic circular dichroism spectroscopy have shown that, for the α-KG- and pterin-dependent mononuclear non-heme iron enzymes, the binding of the penultimate substrate promotes a change to a five-coordinate iron center. As with the P₄₅₀, this finding implies that dioxygen will not bind in the absence of an organic substrate, which prevents the build up of potentially damaging intermediates (23).

**Binuclear non-heme iron oxygenases**

Soluble methane monoxygenase (sMMO) is the best studied binuclear non-heme iron oxygenase enzyme, largely due to its remarkable ability to hydroxylate the stable C-H (440 kJ/mol) of methane (15). sMMO is a three-component enzyme system, which consists of the di-iron hydroxylating protein, a flavin-Fe⁵₇₃ protein, and a third regulatory protein that does not contain a cofactor. The role of the flavin-Fe⁵₇₃ protein is to provide two electrons from NADPH to form the active Fe⁴⁺-Fe²⁺ form of the di-iron hydroxylating protein. Although incompletely understood, the regulatory protein seems to coordinate the interaction of the other two components such that uncoupled reaction cycles do not occur. The intermediates in the sMMO reaction cycle accumulate in the absence of methane, which is a feature that has allowed their spectroscopic characterization (14). Studies on the Mössbauer suggest that O₂ binds initially to the binuclear Fe site forming a μ-1,2-peroxo-Fe⁴⁺ intermediate, which then decays to form the highly oxidizing intermediate termed Q (*Fig. 3c*). EXAFS and Mössbauer have indicated collectively that Q is a bis(μ-oxo)-Fe⁵⁺, with a short Fe-Fe separation of 2.46 Å. A strong consensus on how this intermediate compound hydroxylates methane has not yet been reached despite a wealth of studies using isotopically labeled and radical clock substrates (15). Currently, a radical rebound mechanism or a nonsynchronous concerted rearrangement remains possible.
Copper oxygenases

Copper-containing oxygenases are less common than their iron counterparts, perhaps because of the greater difficulty in obtaining more reactive, higher oxidation states. Examples include dopamine β-monooxygenase (DβM), tyramine β-monooxygenase (TβM), peptidylglycine α-hydroxylating enzyme (PHM), tyrosinase, and a membrane bound form of MMO (16, 17, 24). The membrane bound form of MMO contains both a Type 3 binuclear Cu site and a mononuclear Type 2 Cu site (25). At the moment, its mechanism has not been studied in detail and will not be discussed additionally here.

Figure 3 Illustration of possible partial reaction cycles of some oxygenase enzymes. Water molecules and protein ligands have sometimes been omitted for clarity. (a) P450 (18); (b) intradiol dioxygenase (7); (c) lipoxygenase (7); (d) α-KG-dependent non-heme iron enzymes (14); (e) soluble methane monooxygenase (15); (f) uncoupled binuclear copper (16); (g) coupled binuclear copper; (h) flavin monooxygenases (17).
Uncoupled binuclear copper oxygenases

DyM, TmM, and PHM all employ two uncoupled, Type 2 copper sites in the biosynthesis of neurotransmitters and of hormones (16). The reaction entails the hydroxylation of unactivated carbon centers, similar to the chemistry of the Fe(II) and u-FeG-dependent non-heme iron enzymes. The two metal sites of the uncoupled binuclear Cu monooxygenases are separated by more than 11 Å, which is a feature that distinguishes this family of enzymes from the binuclear Type 3 copper proteins. Each Cu site has a discrete coordination environment and serves a unique function in the hydroxylation reaction. The CuM site functions as the dioxygen activation site, whereas CuM serves as the electron transfer site, providing the additional electron required for substrate oxidation. Evidence suggests that the reaction of dioxygen with the reduced CuIIM generates a CuIIM-superoxide intermediate (Fig. 3f), which subsequently abstracts a hydrogen atom from the substrate (26, 27). A methionine ligand to CuM is believed to stabilize the reduced enzyme form, such that oxygen activation is coupled strongly to the ensuing C-H activation step. The steps following C-H activation are still unresolved. A CuII-oxyl radical (CuII\(\cdot\)) is a postulated, short-lived intermediate along the reaction pathway. This high energy intermediate is unprecedented in copper chemistry, but it could provide the driving force for the requisite electron transfer from CuIIM to CuII.

Coupled binuclear copper oxygenases

A characteristic Type 3, binuclear Cu site is found in tyrosinase (10). Tyrosinase catalyzes the conversion of monophenols to ortho-quinones, the formation of which is coupled to the two electron reduction of \(\text{O}_2\) and proceeds in two steps. The first step occurs via electrophilic attack on the phenol ring of the substrate and is followed by the oxidation of the di-hydroxybenzene to yield the quinone product. Thus, tyrosinase functions as both an oxygenase and an oxidase. Much of the structural information pertaining to the tyrosinase active site and various Cu-dioxygen intermediates was derived by comparison to several spectroscopic studies on model complexes (10). The resting form of tyrosinase assumes two forms: 15% of the enzyme exists in the met state, which is dehydrated to regenerate the resting state. Each Cu site has a discrete coordination environment and is dehydrated to the active conformation. The Baeyer-Villiger monooxygenases and mammalian flavin monooxygenases use an alternative strategy. The flavin is reduced readily by NADPH in the absence of substrate, but the (hydro)peroxyflavin intermediate is stabilized by the bound NADP until a suitable substrate binds.

Flavin oxygenases

Several types of flavoprotein monooxygenases exist. One group catalyzes electrophilic aromatic substitution or heteroatom oxidation reactions, whereas the other group catalyzes Baeyer-Villiger-type oxidations of ketones (Fig. 2) (13, 17). In the well-studied, single-component flavoprotein monooxygenases, the flavin is reduced by a hydride delivered from a single molecule of NADPH (Fig. 3h). The reduced flavin then reacts with \(\text{O}_2\) producing a peroxyflavin intermediate, which is in equilibrium with the hydroperoxyflavin form. For electrophilic substrates, as in the Baeyer-Villiger-type oxidation, it is the peroxyflavin intermediate that reacts with substrate. For hydroxylation of aromatic rings, the more electrophilic hydroperoxy intermediate is the reactive species. In the final step, the flavin is dehydrated to regenerate the resting state.

In the Baeyer-Villiger-type oxidations, the flavin is reduced by a hydride delivered from a single molecule of NADPH (Fig. 3h). The reduced flavin then reacts with \(\text{O}_2\) to form the peroxyflavin intermediate, which is in equilibrium with the hydroperoxyflavin form. For electrophilic substrates, as in the Baeyer-Villiger-type oxidation, it is the peroxyflavin intermediate that reacts with substrate. For hydroxylation of aromatic rings, the more electrophilic hydroperoxy intermediate is the reactive species. In the final step, the flavin is dehydrated to regenerate the resting state.

Oxidases

Oxidases couple the oxidation of an organic substrate to the two- or four-electron reduction of \(\text{O}_2\), producing \(\text{H}_2\text{O}_2\) or two molecules of \(\text{H}_2\text{O}\), respectively. Oxygen atoms from dioxygen are not incorporated into the product, unlike reactions catalyzed by oxygenases. Oxidase reactions may proceed via inner-sphere or outer-sphere mechanisms.

Iron oxidases

Non-heme mononuclear iron oxidases

As mentioned, many non-heme iron enzymes also catalyze oxidase-type reactions, such as desaturation, in biological systems (7). Similar to the non-heme iron oxygenases, the reactions are thought to proceed through an \(\text{Fe}^{3+}\text{O}^2-\) intermediate. Two examples of enzymes that catalyze biologically interesting oxidase reactions are isopenicillin N-synthase (IPNS) and 1-amino-2-cyclopropane-1-carboxylate oxidase (ACCO). IPNS catalyzes a double ring closure in the formation of isopenicillin concomitant with the four-electron reduction of dioxygen to two molecules of water, without the use of any cofactors or cosubstrates other than the \(\text{Fe}^3+\) (Fig. 4e). Ligation of the substrate thiolate activates the \(\text{Fe}^3+\) for reaction with dioxygen by converting the active site from a five-coordinate to a six-coordinate metal center. The first ring closure is believed to occur with heterolytic cleavage of \(\text{Fe}^{3+}\text{O}-\text{OH}\) forming \(\text{Fe}^{3+}\text{OH}\), which then closes the second ring.
ACCO breaks down 1-aminocyclopropane-1-carboxylate (ACC) in plants to form the growth hormone ethylene, hydrogen cyanide, and CO₂. In the initial stages of catalysis, ACC ligates the iron, priming the system for reaction with dioxygen. Electrons for the reduction of O₂ are derived from ascorbate, leading to formation of the reactive iron oxygen intermediate, possibly an FeIV=O. Although the later steps from ascorbate, leading to formation of the reactive iron oxyanion, ultimately generating a tyrosyl radical, which transfers an electron, are poorly characterized, it is thought that an FeIV=O may remove a hydrogen atom from ACC, generating a substrate radical, which breaks down to generate the gaseous product molecules.

Binuclear non-heme iron oxidases: Class I ribonucleotide reductases

Ribonucleotide reductases (RNRs) catalyze the conversion of nucleotides to deoxynucleotides and can be found in all organisms. RNRs are categorized into three classes and employ different cofactors for this process. The Class I RNRs contain a binuclear iron cofactor (29). The iron active site is located on the R2 subunit, one of two homodimeric subunits that compose the Class I RNRs, and is very similar structurally to the active site of sMMO (Fig. 1d). The resulting Cys radical abstracts a hydrogen atom from ACC, generating a substrate radical, which breaks down to generate the gaseous product molecules.

Oxygen-Activating Enzymes, Chemistry of

Copper Oxidases

Mononuclear copper oxidases

Most copper-containing oxygenases and oxidases use multiple metal centers to conduct their biotransformations. Copper amine oxidases (CAO) and galactose oxidase (GalO), instead, employ posttranslationally derived amino acid side chains as cofactors to supply additional electrons (24). CAOs use a quinone cofactor to catalyze the oxidation of primary amines to aldehydes (Fig. 5a) (30). The metal cofactor of CAO is a binuclear Cu(II) center ligated by three histidines and by two water molecules (Fig. 5a). A distance of ∼3 Å separates the copper site and the organic cofactor. Although its catalytic function is as an oxidase, CAO also functions as an oxygenase in the self-processing mechanism of quinone cofactor biogenesis (33). The biogenesis reaction requires two equivalents of molecular oxygen for the six-electron oxidation of an active site tyrosine to 2,4,5-trihydroxyphenylglycol one (TPQ). The catalytic oxidase reaction of the CAOs can be separated into two half-reactions. In the oxidative half-reaction, the amine substrate is oxidized to the corresponding aldehyde. Concurrently, TPQ is converted to the reduced aminoquinol form.
Dioxygen subsequently reacts with the reduced amino group, generating TPQ and one equivalent of hydrogen peroxide. Reduction of the copper center has not proven essential for oxidase activity. In selected CAOs dioxygen has been shown to be activated via direct electron transfer from the reduced TPQ (30). Kinetic studies suggest that the metal center in the oxidase reaction contributes primarily to charge stabilization of the activated dioxygen (Fig. 5a).

GalO contains a unique cofactor, which is composed of a cysteine-cysteine ligand, Tyr-cysteine, ligated to a copper center, derived from posttranslational crosslinking of the two amino acids (Fig. 5b, 32, 33). A second Tyr, two His residues, and water molecule comprise the remaining ligands of the square pyramidal CuII site. Cofactor formation is catalyzed by the enzyme itself, in a dioxygen and CuII-dependent reaction. The Tyr-Cys residue is oxidized during the posttranslational process to yield the active form of GalO, which is a CuII-Tyr-Cys cation radical. The reaction catalyzed by GalO is the oxidation of primary alcohols to their corresponding aldehydes. The oxidation of substrate is coupled to the two-electron reduction of dioxygen to H2O2. Kinetic studies support the oxidation of alcohols via a ping-pong mechanism. In the reductive half-reaction, a hydrogen atom is abstracted from the C-H position of the metal bound alcohol to produce CuI and the reduced cysteinyl-tyrosine. In the oxidative half-reaction (Fig. 5b), dioxygen is believed to bind directly to the reduced Cu site, displacing the water molecule. The Tyr-Cys ligand supplies the additional electron that is required for the two-electron reduction of dioxygen to peroxide, and for the regeneration of the cofactor radical.

Binuclear copper oxidases

The active site of the catechol oxidases is virtually identical to the Type 3 binuclear Cu site found in tyrosinasess (vide infra) (10, 34). In contrast to the tyrosinases, catechol oxidases do not exhibit monooxygenase activity and are capable only of the second reaction, the oxidation of diphenols to quinones. The resting form of catechol oxidases lies exclusively in the met state. According to the accepted mechanism, two molecules of catechol are oxidized on binding to either the reduced, deoxy form or the oxy form. Dioxygen reacts with the reduced form after product release to yield the CuII-oxy-educto adduct and allow binding of the next substrate molecule (Fig. 5c). The difference in reactivity toward O2 for the Type 3 copper centers in tyrosinasess, catechol oxidases, and hemocyanins has been attributed to the partial or the complete occlusion of the substrate binding site in the latter two enzyme families. The degree of flexibility around the copper active site has also been cited as a possible factor (35).

Multicopper oxidases

The multicopper oxidases couple the one-electron or two-electron oxidation of their substrates to the four-electron reduction of dioxygen to water (36). The reaction with substrate can proceed via an outer-sphere or an inner-sphere mechanism, and as a result, the substrate specificity varies substantially among the enzymes. The best-characterized enzymes are laccase, ascorbate oxidase, and ceruloplasmin. Radical phenol and amine species formed by laccase and ascorbate oxidase react further via polymerization reactions with other organic molecules, or disproportionate to generate the final biological products. The substrate for ceruloplasmin is FeII, which is oxidized to FeIII. Four substrate molecules transfer electrons sequentially to the Type 3 Cu site, which shuttles three electrons to the trinuclear cluster to generate the fully reduced enzyme form. Dioxygen is activated by the trinuclear copper cluster, which obtains two electrons from the Type 3 site (Fig. 5d). The Type 2 center is required for dioxygen activation but remains reduced at this stage in the reaction. The hydoperoxide is bound to the trinuclear cluster near the Type 3 binuclear Cu center, but both the Type 2 and the binuclear copper centers contribute significantly to the reduced oxygen molecule (37). The bound peroxide is reduced further by two electrons, one from the Type 2 Cu center and one from the distant Type 1 Cu site, forming the native intermediate. The native intermediate is believed to contain an oxo coordinated by the three Cu atoms of the trinuclear cluster. This species reacts with the substrate in the catalytic cycle.

Flavin Dependent Oxidases

Flavoprotein oxidases can conduct a variety of oxidation reactions, such as the conversion of alcohols and amines to aldehydes (12). Typically, the organic substrate provides an equivalent of hydride to reduce the flavin. As mentioned previously, the rate-limiting step in the reaction of the reduced flavin with dioxygen is the initial electron transfer to form a superoxide anion (30). Unlike the flavin oxygenases, a hydperoxylflavin intermediate has never been detected for the flavin oxidases during catalysis. Instead, the mechanism is thought to proceed by a sequential one-electron transfers forming H2O2 (Fig. 5e). The lack of a solvent deuterium isotope effect in kcat/Km(O2) for glucose oxidase has provided evidence that proton transfer is not rate-limiting for the reduction of dioxygen.

Cytocrome c Oxidase

Cytocrome c oxidase is a vital enzyme in the respiratory pathway of most aerobic organisms (38, 39). The enzyme couples the four-electron reduction of dioxygen to the generation of a proton gradient and the resulting synthesis of ATP, which is the primary source of energy for all cellular processes. Cytocrome c oxidase contains two iron-porphyrin units (heme a and heme a3) and two mononuclear Cu sites (CuB and CuA). Heme a and CuA serve as electron transfer sites, whereas heme a3 and CuB form a binuclear metal site that activates dioxygen. CuB is coordinated by three His residues and is located 5 Å from the Fe center of heme a3. A tyrosine residue is bound covalently to one CuB-His ligand and is believed to be critical to the four-electron reduction of O2. The first intermediate generated upon the reaction of dioxygen with the reduced CuA-FeIIa3 center has been identified, based on resonance Raman spectroscopic studies, as an FeIIa3-O2, although a peroxide bridged FeIIa3-O2–CuIIa3 species has not been ruled out entirely (Fig. 5f) (38). The heme-bound dioxygen molecule is reduced rapidly by four electrons, two of which are obtained from the iron center and one each obtained from CuB and the active site tyrosine. The resultant intermediate P.0 consists of
an $Fe^{II}=O$, Cu, and a tyrosyl radical, as deduced from Raman and EPR spectroscopy (40, 41). A series of proton and electron transfer events regenerates the resting, fully oxidized form of cytochrome $c$ oxidase. The electrons for this process are derived from cytochrome $c$ and shuttled through the CuA and heme $a$ sites, during which protons are pumped across the cell membrane. Thus, cytochrome $c$ oxidase functions as an electron-coupled proton pump.

Figure 5 Illustration of possible partial reaction cycles of some copper- and flavin-dependent oxidase enzymes. (a) Copper amine oxidase 30, 31; (b) galactose oxidase (32); (c) catechol oxidase (10); (d) multicopper oxidases (10); (e) flavin oxidases (30); (f) cytochrome $c$ oxidase (38).

Conclusions

The mechanisms of the oxygenases and oxidases detailed here represent some of the numerous strategies employed by enzymes to overcome the kinetic barrier for reaction of organic molecules with dioxygen. This list is far from exhaustive; new reactions are discovered continually, and many intermediates in more established systems have not yet been characterized. A
key feature of the dioxygen-activating enzymes is their ability to form highly ROS using carefully tuned cofactors. The en-
zymes prevent the release of ROS at a stage in the catalytic cycle when damage to the protein or the wider cellular environ-
ment could occur as a consequence. Efforts to mimic nature’s dioxygen chemistry with synthetic inorganic complexes often are only partially successful (7, 42). Many biomimetic com-
plexes will react with dioxygen, but the resultant M–O bonds are either too stable to catalyze the subsequent chemistry or the products are too unstable and lead to undesirable side reactions.

Although great strides have been made to understand dioxy-
gen activation by enzymes, many questions remain. The rela-
tion between protein structure and enzyme catalysis is not well understood. Changes in the active site, beyond the primary coordination sphere, lead to altered cofactor reactivity. An identical cofactor is often employed by different enzymes to carry out dissimilar chemistry. The protein fold tunes the reactivity of these sites through electrostatic effects, control of solvent and substrate access, and by carefully organizing the orientation of substrates around the active site. The role of pro-
tein dynamics in tuning enzyme reactivity has become the focal point of recent studies, as well. These aspects of enzyme catal-
ysis present the primary difficulties in the design of synthetic complexes that function as protein mimics.

References

4. Solomon EI, Lever ABP. Inorganic Electronic Structure and Spec-
8. Kuritz DM. Structural similarity and functional diversity in dimor-
12. Massry V. A activation of molecular oxygen by flavins and flavo-

Oxygen-Activating Enzymes, Chemistry of
Pyridoxal phosphate-dependent enzymes catalyze a wide variety of reaction types on amines and amino acids, generally by stabilizing carbanionic intermediates, which makes them very common in both primary and secondary metabolism. A fundamental mechanism by which reaction specificity is enforced is stereoelectronic effects; the bond to Cα that gets broken in the universal external amine intermediate is that which is aligned parallel to the p-orbitals of the conjugated, electron-deficient π system. When a bond to Cα has been broken, enzymes must also carefully control the fate of the resulting carbanionic intermediate. Structurally, all known pyridoxal S'-phosphate (PLP)-dependent enzymes are separated into five groups. The largest of these, Type 1, is the aminotransferase group, of which aspartate aminotransferase is the prototype. The prototypes for the other groups are as follows: Type 2, tryptophan synthetase; Type 3, alanine racemase; Type 4, D-amino acid aminotransferase; and Type 5, glycogen phosphorylase.

Biological Background

Pyridoxal S'-phosphate (PLP, the active form of vitamin B6) is an enzyme cofactor in amine and amino acid metabolism. Vitamin B6 was first identified as pyridoxine, which is a catalytically inactive form, in 1938, whereas the catalytically active aldehyde (pyridoxal) and amine (pyridoxamine) forms and their phosphorylated derivatives (pyridoxal S'-phosphate and pyridoxamine S'-phosphate) were discovered in the early 1940s (1). Snell discovered that heating pyridoxal and amino acids yielded the amino form of the coenzyme, and he proposed, based on these observations, that pyridoxal, or a derivative thereof, was the amino form of the coenzyme, and he proposed, based on these observations, that pyridoxal, or a derivative thereof, was the amino form of the coenzyme, and he proposed, based on these observations, that pyridoxal, or a derivative thereof, was the mediator of transamination reactions (1). These three decades observed the demonstration of the involvement of PLP and pyridoxamine S'-phosphate in a wide variety of both enzymatic and nonenzymatic reactions. The enzyme-catalyzed reactions that correspond to the nonenzymatic reactions were shown to follow the same basic chemical mechanisms, in which the enzymes enforce substrate and reaction specificity as well as enhance catalytic power.

PLP-dependent enzymes are ubiquitous; it has been estimated that 1.5% of genes encoded by free-living prokaryotes are PLP-dependent (2). PLP-dependent enzymes are central to nitrogen metabolism, commonly catalyzing racemization, transamination, decarboxylation, retro aldol cleavage, 3-elimination, as well as several other types of chemistry. Figure 1 outlines these mechanisms, with the “Others” group composed of 3-elimination, 3-decarboxylation, radical-based rearrangements, cyclopropyl ring formations and openings, and combinations of different individual activities into single enzymes. Indeed, 4% of all activities classified by the Enzyme Commission are catalyzed by PLP-dependent enzymes (2).

The biosynthesis and degradation of amino acids is the most prominent realm of PLP enzymes. Other metabolic arenas in which they are central include the biosynthesis and degradation of biogenic amines (e.g., histamine and dopamine), heme and chlorophyll, amino sugars, sphingolipids, vitamins (e.g., biotin and folate), antibiotics, polyamines, neurotransmitters (e.g., GABA), and ethylene. These examples include several pathways that are directly affected in or generate pathophysiological conditions. Understanding the fundamental mechanistic bases for the multifarious capabilities of this class of enzymes has both medical and basic science importance.

Despite the diversity of chemistry that PLP-dependent enzymes catalyze, the chemistry of the PLP is based nearly universally on the formation of stabilized carbanions. Without PLP, the Cα proton of amino acids is simply not reactive enough for reaction to take place under physiological conditions. For example, the pKₐ for loss of the Cα proton from an amino acid in the absence of PLP is ~30 (3), and the half-life for the uncatalyzed decarboxylation of an amino acid is estimated to be 1.1 billion years (4). However, by resonance stabilizing negative charge that develops at Cα in the transition state, as well as the...
fully formed carbanion (Fig. 2), PLP makes possible the wide variety of reaction types that PLP-dependent enzymes catalyze. Professor William Jencks once wrote... it must be concluded that pyridoxal phosphate was created to provide satisfaction and enlightenment to those enzymologists and chemists who enjoy pushing electrons, for no other coenzyme is involved in such a wide variety of reactions, in both enzyme and model systems, which can be reasonably interpreted in terms of the chemical properties of the coenzyme. (5)

PLP is used in a remarkable number of reaction types. The first and common step for most PLP-catalyzed reactions is the formation of the “external aldimine” intermediate (Schiff base between PLP and the substrate amino group) from the “internal aldimine” (Schiff base between an active site lysine and PLP) and the free-amine substrate. From this common external aldimine intermediate, the various reactions diverge mechanistically (Fig. 1).

The external aldimine intermediate is poised to undergo facile heterolytic bond cleavage to generate a carbanion next to the amino group, because it is strongly resonance stabilized by the extended conjugation of the Schiff base/pyridine ring $\pi$ system. The occurrence of distinct carbanionic intermediates in PLP enzymes is accepted commonly in the literature.

The chemistry of free PLP has been investigated in great detail and provides a strong foundation for work in enzyme systems. Free PLP is an effective catalyst of many reactions in solution. Early studies with PLP model systems found catalysis of racemization, transamination, and other types of chemistry in solution. In enzyme active sites, the intrinsic properties of PLP account for a large fraction of catalytic power. For example, it has been found that nearly half of the rate acceleration in alanine racemase is caused by PLP ($\sim 10^6$ fold), whereas the active site environment contributes another $10^6$-fold increase in the reaction rate (6). However, the protein environment does not just contribute to an increase in the rate of reaction. Model studies show that PLP can catalyze multiple types of chemistry with a single substrate in solution. For example, serine undergoes racemization, transamination, and retro aldol cleavage in the presence of PLP. However, PLP-dependent enzymes are specific for a single reaction type, which reduce potential side reactions to levels that are barely detectable.
Structural Relationships Among PLP-Dependent Enzymes

All PLP-dependent enzymes that have X-ray crystallographic structures solved to date are categorized into five-fold types (7–9). Each structural class contains representatives of multiple reaction types; the classes are unified not by the chemistry that is catalyzed but instead by three-dimensional structure.

The Type 1 or aspartate aminotransferase family includes many well-studied PLP-dependent enzymes, which include its namesake as well as GABA aminotransferase and diaminopimelate decarboxylase. Enzymes in this group function as homodimers or higher-order oligomers. Each monomer of the Type 1 enzymes has a large and a small domain. The active sites are at the dimer interface, with each monomer contributing residues. Four subgroups exist within the aminotransferase superfamily.

Enzymes in the Type 2 or tryptophan synthase family also generally function as homodimers or higher order oligomers. However, in Type 2 enzymes, the active site is composed of residues from a single monomer. In addition, these enzymes differ from Type 1 enzymes in that they often contain a regulatory domain.

The Type 3 or alanine racemase family contains its founding member, alanine racemase, as well as eukaryotic ornithine decarboxylase and diaminopimelate decarboxylase. Structurally, this family is significantly different from the other fold types. Enzymes in this family consist of an αβ barrel and a second β-strand domain. Despite these gross structural differences, PLP binding in this family is very similar to what is observed in the other fold types, in which the PLP phosphate acts as an anchor and the lysine Schiff base is present. The enzymes are dimers with each monomer contributing residues to the active site.

The Type 4 or D-amino acid aminotransferase family is similar to Types 1 and 2 because they also function as homodimers, and each monomer is composed of a small and a large domain. However, the cofactor is bound in a nearly perfect mirror image of the Fold Types 1 and 2 binding sites, such that the α rather than the β face is solvent exposed.

Finally, the Type 5 or glycogen phosphorylase family is completely different from all other fold types. These enzymes are mechanistically different as well; they use the phosphate group of the cofactor for catalysis.

Mechanisms of Common PLP-Dependent Enzymes

Figure 1 illustrates the breadth of reaction specificity enabled by PLP using serine as an example substrate. The first and common step for all PLP-dependent enzyme-catalyzed reactions is a Schiff base exchange reaction (transamination). All known PLP enzymes exist in their resting state as a Schiff base (internal aldimine) with an active site lysine residue. The incoming, amine-containing substrate displaces the lysine α-amino group from the internal aldimine, in the process forming a new aldimine with the substrate (external aldimine). This process involves many steps, which include several facile steps; it is frequently very rapid compared with the central steps in the reaction mechanism (e.g., deprotonation of Cu). The external aldimine is the common central intermediate for all PLP-catalyzed reactions. Divergence in reaction specificity occurs from this point. Most pyridoxal phosphate-catalyzed reactions depend on the formation of a carbanionic intermediate, whereas some 1,2-aminomutases employ PLP in radical-based reactions (10, 11).

From the external aldimine intermediate, carbanions formed by heterolytic cleavage of any one of the three potentially labile bonds to Cu can be stabilized. Dunathan proposed that enzymes determine which bond to Cu will be cleaved by orienting it parallel to the π orbitals of the conjugated x system (12). In this way, the developing negative charge in the transition state for heterolytic cleavage will be best stabilized, and the bond will be more labile.

Proton abstraction is the most common forward step that external aldimines undergo since racemization, transamination, and β-elimination, which are three common reaction types, all require it. Loss of CO2 gives a carbanion that is commonly protonated on Cu to give the corresponding amine as the product. Less commonly, for example with diaminoglycine decarboxylase (discussed below), the resulting carbanion is protonated on C4 of the coenzyme to give oxidized substrate and the reduced, amino form (PMP) of the coenzyme. Retro-aldol cleavage of serine, which is central to one-carbon metabolism, is initiated by abstraction of a proton from the β-hydroxyl group followed by Cu-Cδ bond cleavage. Other known reaction types include β-decarboxylation of aspartate, β-elimination and replacement, γ-elimination and replacement, ω,ω′-elimination, cyclopentyl ring opening, radical-based 1,2-amino migrations, and others. This wide range of reaction types makes PLP enzymes extraordinarily useful to cells. The enzyme commission (EC) has more than 140 EC numbers assigned to PLP enzymes (2).

The mechanism for stabilization of the resulting carbanion is resonance delocalization within the extended conjugated x system, which is illustrated in Fig 2 where the three most significant resonance forms are shown. The rightmost resonance structure is referred to as the “quinonoid” because its structure resembles that of a quinone. It has strong absorption at ~500 nm (ε500 ~ 40,000 M⁻¹cm⁻¹) and is sometimes spectroscopically observable in enzyme-catalyzed reactions. This quinonoid resonance structure commonly is considered the major species responsible for the catalytic power of PLP because the electrons from Cu are neutralized by the protonated pyridine nitrogen.

Aspartate aminotransferase is the prototype of the large aminotransferase family. The catalytic cycle of all transaminases is composed of two reversible half-reactions. These reactions comprise the same chemical steps but employ substrates with different carbon backbones. For example, the a-amino acid/α-keto acid pairs aspartate/oxalacetate and glutamate/β-ketoglutarate in the two half-reactions of aspartate aminotransferase. The mechanism of the transamination half-reaction with
aspartate/oxalacetate is shown in Fig. 3. The transamination cycle is completed by the reverse reaction of α-ketoglutarate with the amino form of the cofactor through the same series of chemical steps.

In decarboxylation, the external aldimine formed from the internal aldimine and amine acid substrate loses CO₂ and forms the quinonoid intermediate. The quinonoid is usually protonated at Cα, which results in nonoxidative decarboxylation, to generate the imine between the amine product and PLP. In the next step, the product imine undergoes transamination with the active site lysine residue to regenerate the internal aldimine and give free product. Most PLP-dependent decarboxylases catalyze nonoxidative decarboxylation (i.e., amine formation) as their primary reaction. For these enzymes, oxidative decarboxylation that results from protonation at C4′ is a side reaction that results in an inactive enzyme.

Many PLP-dependent decarboxylases studied have been shown to facilitate decarboxylation using ground-state destabilization achieved by placing the charged carboxylate in a hydrophobic environment. Examples include ornithine decarboxylase (13, 14), dopa decarboxylase (15), and arginine decarboxylase (16). DGD is an interesting exception that has a hydrophilic active site with residues stabilizing the carboxylate.

A common feature of PLP-dependent decarboxylases is the stereochemistry of the catalyzed reaction. All PLP-dependent decarboxylases studied that act on an L-stereocenter have been found to decarboxylate with retention of configuration, that is, the carboxylate is removed from one face of the molecule and the proton is placed on this same face. The only known exception is diaminopimelate decarboxylase, which decarboxylates a D-stereocenter with inversion of configuration (17, 18).

Dialkylglycine decarboxylase (DGD) is an interesting and unusual PLP-dependent decarboxylase that catalyzes decarboxylation-dependent transamination. In the first half-reaction catalyzed by DGD, a small dialkylglycine, such as aminoisobutyric acid, is decarboxylated oxidatively to give the ketimine between acetone and PMP. DGD is highly specific for oxidative decarboxylation (19). The ketimine formed is hydrolyzed to give free acetone and the PMP form of the enzyme. The second half-reaction is a classic transamination half-reaction, in which a ketimine forms between PMP and an α-keto acid, such as pyruvate, followed by a 1,3-prototropic shift from C4′ to Cα to give the external aldimine between PLP and L-alanine.

In the DGD active site (Fig. 4), PLP is bound covalently to the enzyme via a Schiff base linkage to Lys272, as observed in all PLP-dependent enzymes studied to date. Modeling of the L-isovaline and L-alanine external aldimines into the DGD active site allowed development of a functional active site model that is consistent with the previously observed reactivity and the idea of stereoelectronic control of reaction specificity proposed by Dunathan (20). Additional work with alternate substrates has validated this model (19).

In the functional active site model for DGD, three subsites exist (Fig. 4). The A subsite is near Q52 and K272 and is the position in which all bond breaking and making occurs, because it is here that the bond is aligned parallel to the π system of the PLP ring and is activated stereoelectronically. When decarboxylation occurs, the carbonyl is proposed to bind in the activated A subsite, which interacts with Q52 and K272. X-ray crystallographic studies with aminophosphonate...
inhibitors provide support for this proposition, which show the phosphonate near the A subsite (21). When transamination occurs, as with L-alanine, Dunathan’s hypothesis predicts that the Cα-phosphonate near the A subsite is lost, which produces the carboxylate near the B subsite. X-ray crystallographic structures show that the B subsite residues R406 and S215 interact well with a carboxylate in this position (21, 22). Thus, studies on DGD have allowed the experimental confirmation of Dunathan’s stereoelectronic hypothesis for control of reaction specificity.

Tryptophan synthase is representative of the Type 2 PLP-dependent enzymes. It catalyzes the replacement of the L-serine hydroxyl group with indole in the final step of tryptophan biosynthesis. In this reaction, the Cu ion is removed to give the carbinol, which eliminates hydroxide from Cu. This step is followed by addition of indole to the resulting aminoacyl-carbinol intermediate. This enzyme is an \( \alpha_2 \beta_2 \) heterodimer. The \( \alpha \) subunit catalyzes the release of indole from indole-3-glycerolphosphate, but this indole is not released into solution. Instead, the enzyme has a molecular tunnel that leads from the active site of the \( \alpha \) subunit to that of the \( \beta \) subunit (23, 24). The activities of the two active sites are regulated by allosteric interactions. This allosterism also regulates the channeling of indole between the active sites.

The Type 3 prototype is alanine racemase. This enzyme provides bacteria with D-alanine for cell wall biosynthesis. It catalyzes perhaps the simplest reaction of all PLP-dependent enzymes: the removal of the Cα proton from one stereochmical face and its replacement on the other. Lys39 and Tyr265 are the two bases that deprotonate/reprotonate Cα on opposite stereochmical faces (6). The Type 4 prototypes are D-amino acid aminotransferase, whose mechanism follows that outlined above for all aminotransferases (25). The Type 5 prototype is glycogen phosphorylase. This class of PLP-dependent enzymes is unique in that the phosphoglucone group and not the aldehyde group is involved directly in catalysis of polysaccharide phosphorylation.

References


Further Reading

http://en.wikipedia.org/wiki/Pyridoxal_phosphate
http://en.wikipedia.org/wiki/Vitamin_B_6
http://www.rcsb.org/pdb/home/home.do

©2006, John Wiley & Sons, Inc.
See Also

Amino Acids, Chemical Properties of
Cofactor Biosynthesis
Enzyme Catalysis, Chemical Strategies for
Enzyme Cofactors
Vitamins, Chemistry of
Proteins are biological macromolecules synthesized with genetically determined sequences of amino acids by the formation of peptide bonds through the removal of water molecules. The chemical nature of amino acids is important for the folding and stability of protein structures. In this article, we give an overview of the computational analysis of protein structure and function from the viewpoint of chemical biology. The biological insights obtained from the analysis of known protein structures and their applications to protein structure prediction and folding rates are outlined. The computational analysis needed for understanding the functions and stability of proteins and for predicting the functionally important residues is described. The methods developed for understanding the factors that influence the stability of proteins and predicting protein stability during mutation are explored. Furthermore, the current status of computational protein design and the approaches to understanding the recognition mechanism of protein complexes are discussed.

The formation of stable secondary structures and a unique tertiary structure of proteins are dictated by the interactions between constituent amino acid residues along the polypeptide chain and by their interactions with the surrounding medium. During the process of protein folding, the hydrophobic force drives the polypeptide chain to the folded state and overcomes the entropic factors while hydrogen bonds, ion pairs, disulfide bonds, and van der Waals interactions define the shape and keep it from falling apart. The structure of a protein mainly dictates its function, and the attainment of stable conformation is essential for proper function. Hence, many methods have been developed to determine the three-dimensional structures of proteins experimentally.

X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy are widely used to determine the structures of proteins and their complexes. These structures have been deposited in Protein Data Bank (PDB) and are available at http://www.rcsb.org/ (1). Currently, the PDB has nearly 50,000 structures, and the wealth of data provide valuable information on relating the structures of proteins with their functions, interactions, and evolution. For example, the structure of a protein explicitly reveals the presence of various interactions, cavities, clefts, active sites, binding regions, and so forth.

Figure 1 shows the crystal structure of X-linked inhibitor of apoptosis (XIAP) binding with small-molecule inhibitor, which helps to restore apoptosis and offset the cancerous state. The availability of many protein structures has enabled researchers...
to perform various computational analysis from different perspectives such as structure, function, and stability.

The applications of the chemistry of amino acids to the biological problem, protein structure and function, and folding and stability are the main focus of this article. The article is divided into five main sections that include the biological insights on protein structures, chemical applications including protein functions, thermodynamics of proteins, protein interactions, and computational protein design.

### Biological Insights

The computational studies on protein structures have been carried out from different perspectives in biology. These perspectives include 1) comparison of protein structures, 2) structural analysis, and 3) protein structure prediction. Table 1 lists the online resources that are providing such services.

#### Table 1: Online resources for protein structure analysis and prediction

<table>
<thead>
<tr>
<th>Protein Data Bank</th>
<th>Budget, <a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein classification</td>
<td>SCOP, <a href="http://scop.mrc-lmb.cam.ac.uk/scop/">http://scop.mrc-lmb.cam.ac.uk/scop/</a></td>
</tr>
<tr>
<td></td>
<td>CATH, <a href="http://cathwww.biochem.ucl.ac.uk/latest/index.html">http://cathwww.biochem.ucl.ac.uk/latest/index.html</a></td>
</tr>
<tr>
<td>Protein structure comparison</td>
<td>DALI, <a href="http://ekhidna.biocenter.helsinki.fi/dali/start">http://ekhidna.biocenter.helsinki.fi/dali/start</a></td>
</tr>
<tr>
<td></td>
<td>CE, <a href="http://ccl.ms.sdsc.edu/ce.html">http://ccl.ms.sdsc.edu/ce.html</a></td>
</tr>
<tr>
<td></td>
<td>PRIDE, <a href="http://hydra.icgeb.trieste.it/pride/">http://hydra.icgeb.trieste.it/pride/</a></td>
</tr>
<tr>
<td></td>
<td>MATRAS, <a href="http://biunilastnara.jp/mtras/">http://biunilastnara.jp/mtras/</a></td>
</tr>
<tr>
<td></td>
<td>TOPS, <a href="http://www.tops.leeds.ac.uk/">http://www.tops.leeds.ac.uk/</a></td>
</tr>
<tr>
<td>Protein three-dimensional structure prediction</td>
<td>FRankenstein, <a href="http://genesilico.pl/meta/">http://genesilico.pl/meta/</a></td>
</tr>
<tr>
<td></td>
<td>ROBETTA, <a href="http://robeta.bakerlab.org/">http://robeta.bakerlab.org/</a></td>
</tr>
<tr>
<td></td>
<td>FORTE, <a href="http://www.cbcj.jp/hbin/forte-cgi/forte_form.pl">http://www.cbcj.jp/hbin/forte-cgi/forte_form.pl</a></td>
</tr>
<tr>
<td>Computation of solvent accessibility</td>
<td>ACCESS, <a href="http://www.csb.yale.edu/">http://www.csb.yale.edu/</a></td>
</tr>
<tr>
<td></td>
<td>ASC, <a href="http://mendel.imp.univie.ac.at/mendel/ps/studies/asc.jsp">http://mendel.imp.univie.ac.at/mendel/ps/studies/asc.jsp</a></td>
</tr>
<tr>
<td></td>
<td>DSSP, <a href="http://www.cmbi.kun.nl/ps/dssp/">http://www.cmbi.kun.nl/ps/dssp/</a></td>
</tr>
<tr>
<td></td>
<td>NACCESS, <a href="http://wolf.bms.umist.ac.uk/naccess/">http://wolf.bms.umist.ac.uk/naccess/</a></td>
</tr>
<tr>
<td></td>
<td>POOPS, <a href="http://mathbio.nimr.mrc.ac.uk/~ffranca/POOPS">http://mathbio.nimr.mrc.ac.uk/~ffranca/POOPS</a></td>
</tr>
</tbody>
</table>

### Comparison of protein structures

The function of a protein mainly depends on its structure, and proteins of similar structures perform similar functions. Structure comparison is the process of analyzing two or more structures to evaluate the extent of similarity in their three-dimensional structures. The comparison of protein structures is very important in the design of algorithms that predict protein function and is expected to play a major role in structural genomics projects. The comparison also enables one to classify the protein structures into different structural classes, folds, families, and superfamilies. Typically, the structures being compared are superposed, and the measure, root mean square deviation (RMSD), is widely used to characterize structural relatedness among the proteins being compared. It has been reported that the method PRIDE (2), based on the distribution of amino acid Cα atoms, is fast and can handle large amounts of data and vast structural databases. However, it has many...
false positives and negatives. Yet, the method combinatorial extension (3), based on aligned fragment pairs, is an accurate one, but it is slow for comparison. In addition, comparison of protein structures is helpful to get the biological information in distinctly related protein structures to detect the similar and dissimilar regions between protein structures.

Analysis of protein structures

The computational analysis of unrelated protein structures (nonredundant set) enables one to identify the fundamental principles that govern their folding and stability. Different types of computational analyses such as solvent accessibility; surrounding hydrophobicity, flexibility; thermodynamic parameters; average medium- and long-range contacts; preference of amino acid residues in α-helical, β-strand, and coil regions; and so forth have been carried out. In these calculations, either the coordinates of α-carbon atoms or all atoms in protein structures have been used. Solvent accessibility is a measure to identify the location of amino acid residues (inside the protein core or on the surface) in protein structures (4). It is defined as the locus of the center of the solvent molecule (usually a sphere of water with a radius 1.4 Å) as it rolls over the van der Waals surface of the protein (Fig. 2). The concept of solvent accessibility is used to compute the hydrophobic or solvation free energy. Surrounding hydrophobicity of a residue in protein structure is the sum of the experimental hydrophobic indices of the residues that are located within a distance of 8 Å (5). The short-, medium-, and long-range contacts are computed with the information about the contacting residues in space (less than 8 Å) and their respective locations with the intervals of 1–2, 3–4 and > 4 residues, respectively in the sequence (Fig. 3). The conformational parameters have been derived by the ratio between the frequency of occurrence of a specific amino acid residue (say A→A) in a particular conformation (say α-helix) and the occurrence of the same residue in the whole protein (6). These parameters derived from the computational analysis of protein structures have been used extensively to predict the secondary structures, solvent accessibility, and residue contacts from amino acid sequence.

Furthermore, the parameters derived from protein three-dimensional structures characterize the topology of proteins, which play an important role in kinetics of protein folding. For example, contact order (7) and long-range order (8) are derived from the information about residue contacts in protein structures. Contact order reflects the relative importance of local and nonlocal contacts to the native structure of a protein, whereas long-range order accounts for the contacts between two residues that are close in space and far in the sequence in protein structure. These parameters have shown a strong correlation with protein folding rates, which is a measure of slow/fast folding of proteins from the unfolded state to their native three-dimensional structures. The amino acid properties and the topological parameters are successfully used to predict protein folding rates from amino acid sequence and structure, respectively.

Prediction of protein structures

Predicting the three-dimensional structure of a protein from its amino acid sequence, known as the protein folding problem, is a challenging task. Considering the rapid growth of amino acid sequences against known structures, prediction of protein structures from sequences is necessary in structural and functional genomics. As an intermediate step, several methods have been proposed for predicting protein secondary structures and solvent accessibility. These methods are based on statistical
Proteins: Computational Analysis of Structure, Function, and Stability

Valencia's group (14) proposed a method based on structural templates in the form of multiple alignments with structures and local conservation of residues for predicting the functionally important residues (e.g., catalytic sites and residues that are involved in substrate or cofactor binding). Generally, a database will set up the functional residues, and their characteristic features will be analyzed based on available sequence and structure analysis tools. These parameters are used for predicting the functionally important residues. Fisher et al. (15) analyzed the conservation of functional residues and their neighboring residues, the distribution of amino acid residues, and the predicted secondary structure and solvent accessibility and developed a method for predicting functionally important residues.

Chemistry

Protein function

The functions of proteins are dictated with the importance of specific amino acid residues that are acting as catalytic site residues and binding site residues, which are capable of binding small molecules (ligands) and interacting with macromolecules. The chemistry of amino acids plays a major role in developing computational models for characterizing the functionally important residues in proteins. Thornton's group (13) collected the catalytic site residues in protein structures and developed a database, Catalytic Site Atlas (CSA), which is available at http://www.ebi.ac.uk/thornton-srv/databases/CSA/. The catalytic site residues have been analyzed with the pattern of hydrogen bonding, location of amino acids in proteins, conservation of residues, ability to possess binding pockets, and so forth. Furthermore, the catalytic mechanisms of hydrolysis and transfer reactions have been compared with the aid of a database of enzyme catalytic mechanisms (http://mbs.cbrc.jp/Ecatal/).

Functional residues identified by site-directed mutagenesis

The site-directed mutagenesis experiments provided a large amount of data about the functional importance of amino acid residues in a protein. Systematic studies on p53 tumor suppressor protein showed that the mutation of each residue leads to different kinds of tumors. Furthermore, several residues are identified as important for drug resistance, protein translocation, substrate binding, increased affinity and activity, and so forth. The information about human genes, genetic disorders, and the effects of mutation has been collected in the form of databases, Online Mendelian Inheritance in Man (OMIM) and Human Gene Mutation Database (HGMD). The combined information about the mutational effects of protein stability and diseases along with sequence and structural details is available at http://pbk26.bse.kyutech.ac.jp/puhou/3dinsight/disease.html. These resources will be helpful in predicting the functionally important residues, which is eventually useful for protein engineering experiments.

Prediction of functionally important residues

The functionally important residues are dispersed in a protein, which makes it difficult to map them on structures. However, the comparison of protein structures provides useful information for predicting protein function. Valencia's group (14) proposed a method based on structural templates in the form of multiple alignments with structures and local conservation of residues for predicting the functionally important residues (e.g., catalytic sites and residues that are involved in substrate or cofactor binding). Generally, a database will set up the functional residues, and their characteristic features will be analyzed based on available sequence and structure analysis tools. These parameters are used for predicting the functionally important residues. Fisher et al. (15) analyzed the conservation of functional residues and their neighboring residues, the distribution of amino acid residues, and the predicted secondary structure and solvent accessibility and developed a method for predicting functionally important residues.

Protein Thermodynamics

Protein stability is the free energy difference (ΔG) between the folded and unfolded states at physiological conditions, and it is in the range of 5–25 kcal/mol. Site-directed mutagenesis experiments provided a wealth of data for understanding the importance of chemical interactions for the stability of proteins during amino acid substitutions. Protein stability is experimentally measured with differential scanning calorimetry, circular dichroism, fluorescence spectroscopy, and so forth. The availability of such data in an electronically accessible database would be a valuable resource for the analysis and prediction of protein mutant stability.

Thermodynamic database for proteins and mutants

ProTherm (16) is a large collection of thermodynamic data on protein stability, which has information on 1) protein sequence and structure (2) mutation details (wild-type and mutant amino acid: hydrophobic to polar, charged to hydrophobic, aliphatic to aromatic, etc.), 3) thermodynamic data obtained from thermal and chemical denaturation experiments (free energy change, transition temperature, enthalpy change, heat capacity change, etc.), 4) experimental methods and conditions (pH, temperature, buffer and ions, measurement and method, etc.), 5) functionality (enzyme activity, binding constants, etc.), and 6) literature.
Amino acid properties and protein stability

Analysis of the relationship between amino acid properties and protein stability showed that hydrophobicity is the major factor for the stability of proteins during substitution of amino acids in the interior of the protein. The stability of protein mutants is attributed to the number of carbon atoms, which shows the direct relationship between hydrophobicity and stability.

For the mutations on the surface of the protein, the classification based on the chemical nature of amino acids, such as hydrophobic amino acids (Ala, Cys, Phe, Gly, Ile, Leu, Met, Val, Trp, and Tyr), amino acid side chains that can form hydrogen bonds (Asp, Cys, Glu, His, Lys, Met, Asn, Gin, Arg, Ser, Thr, Trp, and Tyr), and so forth, improved the correlation between amino acid properties and protein mutant stability. Furthermore, the inclusion of neighboring and surrounding residues remarkably improved the correlation in all the subgroups of mutations. This result indicates that the information from nearby polar-charged amino acid residues and the aliphatic and aromatic residues that are close in space is important for the stability of exposed mutations.

Empirical potentials and protein stability

The information about the preference of amino acid residues for contact with each other in protein three-dimensional structures has been used to derive empirical potentials for understanding protein folding and stability. The most widely used approach is the computation of the frequencies of sequence and structure features and the conversion of the frequencies into free energies (17). The potentials based on torsion angles and conformation features and the conversion of the frequencies into free energies (17) have also been used to estimate the stability change between wild-type and mutant proteins by Serrano’s group (19). In this method, all contact residues have been used for predicting protein mutant stability (18). Furthermore, free energy approach has also been used to derive empirical potentials for understanding protein stability (18). Furthermore, free energy approaches have been systematically considered to compute the stability, especially the van der Waals contributions of all atoms, solvation energies (17), amongst others.

The potentials based on torsion angles and conformation features and the conversion of the frequencies into free energies (17) have also been used to estimate the stability change between wild-type and mutant proteins by Serrano’s group (19). In this method, all contact residues have been used for predicting protein mutant stability (18). Furthermore, free energy approach has also been used to derive empirical potentials for understanding protein stability (18). Furthermore, free energy approaches have been systematically considered to compute the stability, especially the van der Waals contributions of all atoms, solvation energies (17), amongst others.

Stabilizing residues in protein structures

The key residues that are important for the stability of protein structures have been delineated from a consensus approach (20). This method includes the concept of long-range interactions, hydrophobicity, and conservation of amino acid residues in homologous sequences. It shows good agreement with experimental thermodynamic data of globular proteins, and the consensus method could be used to identify potential candidates for protein engineering.

Protein Interactions

Protein–Protein Interactions

Protein interactions play a key role in many cellular processes. Despite the fact that protein interactions are remarkably diverse, all protein interfaces share certain common properties (21). The structural analysis of protein–protein complexes showed that a specific preference for amino acid residues is found at the interface, and most interfaces consist of completely buried cores surrounded by partially accessible rims. The binding sites consist of a few highly packed regions, which contribute significantly to the free energy of binding. In protein–protein complexes, contacts between hydrophobic residues were favorable, pairs of hydrophilic and polar residues were unfavorable, and the charged residues tended to pair subject to charge complementarity. The buried interface is mainly created by nonpolar (carbon-containing) groups followed by polar (N-, S-, and O-containing) groups. Furthermore, hydrogen bonds and salt bridges are found to be important for protein–protein interactions.

Protein–protein interactions represent a highly populated class of targets for drug discovery (22). Integrins are cell surface receptors and are the first example of drug-like protein–protein inhibitors. They bind various protein ligands, such as fibronectin and vitronectin, which possess an Arg-Gly-Asp motif. Furthermore, inhibitors of integrin receptors from various chemical
Regarding the interacting pairs of protein–protein interactions. Protein interaction databases provide a wealth of information used to predict interaction networks in other species, which is networks. The interaction maps obtained for one species can be has enabled the construction and analysis of protein interaction networks. However, the fast development of mass spectroscopy and other experimental techniques for studying protein interactions has enabled the construction and analysis of protein interaction networks. The interaction maps obtained for one species can be used to predict interaction networks in other species, which is useful for identifying the functions of unknown proteins. The protein interaction databases provide a wealth of information regarding the interacting pairs of protein–protein interactions.

Protein–nucleic acid interactions

Protein-nucleic acid interactions play a key role in many vital processes, including regulation of gene expression, DNA replication and repair, and packaging. The remarkable specificity with which proteins recognize target DNA sequences is of considerable theoretical and practical importance, and its basis has been demonstrated through structural analysis of many protein–DNA complexes. The structural analysis of protein–DNA interactions showed that hydrogen bonds are the most frequent interactions followed by van der Waals, hydrophobic, and electrostatic interactions. The electrostatic interactions between the atoms in amino acids and the bases of DNA make many hydrogen bonds in the major groove of DNA, whereas hydrophobic interactions are dominant in the minor groove. The protein–DNA recognition is characterized by both nonspecific and specific interactions via the contacts between the atoms in amino acids and the phosphate/sugar chains of DNA and between the side chains of amino acids and the bases of DNA. In specific interactions, the positive-charged amino acids, Arg and Lys, are dominant in the binding sites of proteins. The main contributor from DNA is thymine with Lys and guanine with Arg for specific interactions; nonspecific contacts are mainly coordinated with phosphate groups of DNA. In addition, the weak interactions including CH...π contacts that involve the thymine methyl group and position C5 of cytosine are comparable to the number of protein–DNA hydrogen bonds. Furthermore, the chemical and physical properties, such as polarity, size, shape, and packing; conformational changes in DNA (bending and flexibility); and water molecules bridging amino acids and bases, are important for the specificity of protein–DNA complexes.

The free energy of binding is a measure for understanding the importance of specific residues/bases for binding, and the experimentally measured specificities of protein–nucleic acid interactions have been compiled in PDBbind database (http://pdb2bind.bse.kyutech.ac.jp/pdb2pocket/pronit.html). The binding of protein with DNA is also characterized by the preference of secondary structures in proteins as well as specific motifs, such as helix-turn-helix, leucine zipper, zinc finger, helix-loop-helix, and so forth. Recently, for the first time, the thermodynamic characterization of the stability and specific DNA binding of a full-length gene product of the MyoMadMax family, namely, Max protein isoform p21 (p21) with helix-loop-helix-leucine zipper motif, have been reported (23). They showed that the association is driven by a large exothermic effect, which is partly compensated for by entropic factors.

In protein–RNA interactions, the recognition frequency occurs in noncanonical and single-strand-like structures that allow interactions to occur from a much wider set of geometries and make use of unique base shapes and hydrogen-bonding ability. Although it forms a glove-like tight binding pocket around RNA bases, the size, shape, and nonpolar binding patterns differ between specific RNA bases. A dinucleotide is distinguishable from guanine based on the size and shape of the binding pocket and steric exclusion of the guanine N2 exocyclic amino group. The unique shape and hydrogen-bonding pattern for each RNA base allows proteins to make specific interactions through a very small number of contacts. Furthermore, knowledge-based potential functions are shown to be useful for predicting the specificity and binding free energy of protein–RNA complexes.

Protein–ligand interactions

The biological functions of proteins depend on their direct physical interactions with other molecules. Specificity of such interactions is crucial so that each protein must interact only with the appropriate molecule and not with any others present in the cell. A molecule of any type that interacts with a protein is designated as a ligand. The affinity between a protein (P) and a ligand (L) is measured by the association constant (K_a), and for the binding reaction at equilibrium, K_a = [P•L]/[P][L]. The inverse of K_a is dissociation constant (K_d), and the free energy of binding, ΔG_bind = −RT ln K_a = RT ln K_d.

Protein–ligand interactions are important for finding new pharmaceuticals, industrial compounds, and functional molecules for food products. It has been shown that a ligand’s partial charges are critically important for estimating the electrostatic interactions in ligand–receptor interactions. Furthermore, the van der Waals energy between proteins and ligands and solvation energy based on the interactions of chemical groups with aqueous medium are used to identify the ligand binding sites.

Prediction of ligand binding sites is an essential part of the drug discovery process. Generally, the binding sites are predicted with bioinformatics approaches, whereas the binding free energy of binding can be determined experimentally using isothermal titration calorimetry or fluorescence polarization.
Computational Protein Design

Computational protein design is a practical option for solving problems in protein engineering. Several advances have been made in this area that include the development of new potential functions, efficient ways of computing free energies, flexible treatments of solvent and solvent-mediated effects (pairwise approximation to continuum electrostatics implemented with the finite-difference Poisson–Boltzmann model), energy function approximations, ensemble-based approaches for inclusion of entropic effects, improvements to stochastic search techniques, and methods to design combinatorial libraries for screening and selection. These new approaches have several applications such as the successes in the design of specificity for protein folding, binding, and catalysis, in the redesign of proteins for enhanced entropic effects, improvements to stochastic search techniques, and methods to design combinatorial libraries for screening and selection. These new approaches have several applications such as the successes in the design of specificity for protein folding, binding, and catalysis, in the redesign of proteins for enhanced entropic effects, improvements to stochastic search techniques, and methods to design combinatorial libraries for screening and selection.

Computational protein design is also applied for redesigning a whole protein. Mayo’s group (25) redesigned a protein, Drosophila melanogaster engrailed homeodomain, which has a sequence of 51 amino acid residues. They used dead-end elimination theorem to overcome the complexity of calculations including the energy functions for simultaneously modeling the residues in solvent-exposed protein surface and in hydrophobic core. Two of the designed sequences were evaluated experimentally, and those two sequences differ by 11 mutations and share 22% and 24% sequence identity with the wild-type protein. Interestingly, computation effectively to design stable proteins and protein engineering.

Conclusions

The chemical nature of amino acids mainly dictates the contacts between them in protein structures in terms of hydrophobic, electrostatic, hydrogen bonding, disulfide bonds, and van der Waals interactions, which are key determinants for the structure, folding, stability, and functions of proteins. The computational analyses of protein structures, such as comparison, prediction, and development of structure-based parameters, can identify the structurally similar regions, functionally important residues, and modeling and folding rates of proteins. The functionally important residues have been identified with catalytic sites and substrate binding sites and are important to protein transport, signaling, channels, enhance activity, affinity, and so forth. These residues are predicted with structurally similar regions, conservation of residues, and physico-chemical properties.

The relationship between amino acid properties and protein stability revealed that the number of carbon atoms (methyl and methylene groups) that reflects the property hydrophobicity has a strong relationship with protein stability for the mutations in the interior of the protein. Yet, hydrophobic, hydrogen bond, electrostatic, and other polar interactions are important for the stability of mutation at the surface of the protein. The atom pair potentials set up on the basis of chemical nature and connectivity successfully could predict protein stability during amino acid substitution.

The structural analysis of protein complexes showed that the chemical behavior of the interacting partners is important for understanding the recognition mechanism. The binding of small molecules with proteins provides deep insights for drug discovery. In addition, computational approaches have been used effectively to design stable proteins and protein engineering.

References

Proteins, Chemistry and Chemical Reactivity of

S. David Tilley, Neel S. Joshi, and Matthew B. Francis,
Department of Chemistry, University of California, Berkeley, Berkeley, California
doi: 10.1002/9780470048672.wecb493

Protein bioconjugation is a critically important tool for the elucidation of enzyme mechanisms, the tracking of biomolecules in living systems, the improvement of pharmacokinetic properties, and the construction of new materials. All these applications rely on a continually expanding set of chemical reactions that can modify native protein functionality in aqueous solution under mild pH and temperature conditions. To survey these techniques, this article provides an introduction to the chemical reactivity of the amino acid side chains, with an emphasis on the selectivity that can be achieved using a particular reactive strategy. Site-selective techniques that target the unique reactive properties of N-terminal residues are also reviewed, as are native chemical ligation methods for the modification of the C-terminus. Whenever possible, the mechanistic aspects of the reactions are discussed, as these considerations provide the foundation for future reaction development. The article concludes with a brief description of labeling reactions that selectively target unnatural functional groups in the presence of native protein functionality.

Our understanding of the chemical reactivity of proteins lies at the very heart of chemical biology. In concert, the chemical behavior of the amino acid functional groups provides the basis for enzyme catalysis, and the differences between their reactive properties allow chemoselective protein labeling techniques to be developed. The ability to attach new functionality to a protein of interest allows the rich diversity of biomolecular function to be expanded to the limits of one’s imagination, which affords hybrid structures that can enhance our understanding of and interactions with living systems. For example, the biological role of many proteins has been clarified by tracking their cellular localization with attached fluorescent probes (1, 2). In many cases, chemical cross-linking strategies have then facilitated the identification of their binding partners. Radiolabels are commonly introduced to assess protein biodistribution in vivo, and polymers have been attached to improve circulation behavior (3, 4). Surface immobilization offers exceptional promise for the generation of microarrays for proteomic studies, and chemically modified proteins have been used to build new materials that use biological structure to template inorganic crystal growth (5, 6).

At the core of these studies is the expanding set of chemical reactions that can modify biomolecules in buffered aqueous solution. These techniques push organic reactivity to the limits of efficiency and chemoselectivity, often with the ability to modify a single functional site among hundreds of spectator groups. This article focuses on the reactive principles that have led to the development of these strategies, with a particular emphasis on the site selectivity that can be achieved for each. It is hoped that this survey can serve as a useful guide for those individuals wishing to modify a protein of interest and that it can facilitate the design of new strategies for modifying proteins with ever-increasing levels of precision and yield.

General Reactivity and Stability Considerations

Proteins have evolved to catalyze a remarkably diverse set of chemical transformations, largely by employing general acid/base chemistry and radical pathways (7). Although this range of reactivity provides the very reason for our interest in the chemical behavior of proteins, it also limits our ability to make general predictions about how specific reagents and conditions will affect a biomolecular target. As long as we remember that each protein has a unique reactivity “personality,” however, we can formulate a few guidelines that can serve as logical starting points for protein analysis, purification, and modification protocols.
The amide linkages of proteins are generally stable under both acidic and basic conditions, as are the native amino acid functional groups. However, changes in pH will alter the conformation state of many side chains (Fig. 1), which will result in changes in their reactivity and the overall folding state of the protein. To maintain higher-order structure, most modification techniques are conducted at a pH between 6 and 8.5, although many proteins will remain folded outside this range. The three-dimensional structure of most proteins is stable in aqueous solution between room temperature and 37 °C, and many proteins (such as those isolated from thermophilic organisms) can be heated to temperatures that are significantly higher before unfolding occurs. Large fractions of organic cosolvents (such as methanol, DMSO, and DME) are generally disruptive to protein structure, whereas small amounts are often used to solubilize the organic reagents that are used for protein modification. Although it is sometimes possible to refold proteins into their native conformation after denaturation occurs, this refolding can be difficult to achieve in practice because of unwanted aggregation or the absence of chaperones that facilitate the correct folding pathway. Thus, unless a refolding protocol is available or anticipated, it is recommended that proteins be processed under the mildest conditions possible if more biochemical studies are to be pursued.

The side chain functional groups contain a wide range of nucleophiles, including amines, thiolates, carbonylates, phenolates, and electron-rich aromatic rings, but disulfide bonds represent the only appreciably electrophilic functional groups. Therefore, the bulk of protein modification reagents are themselves electrophilic in nature. Amino acid side chains in the active sites of enzymes often show distinct nucleophilicity because of the alteration of their pKₐ values or other forms of activation. This nucleophilicity allows these positions to be targeted by less-reactive electrophiles that are ineffective for the modification of “ordinary” residues. An excellent review of this technique for the activity profiling of enzymes has recently appeared (8).

Many side chain groups are prone to oxidation, including thiolates, thioethers, and indole rings, but proteins are relatively nonreactive toward reducing agents (disulfide bonds again being the one exception). Although several proteins are known to catalyze reactions through radical pathways, proteins themselves are often unreactive toward radical species in solution (such as polymerization initiators (9)). The lack of alkene functional groups also precludes the direct participation of proteins in most electrocyclization reactions.

A critical consideration for any protein labeling experiment is the level of site specificity that can be achieved. Largely, the inherent selectivity is dictated by the abundance of the targeted amino acid in the protein sequence. As listed in Fig. 1, the commonly modified amino acids occur with average frequencies ranging from 1.4% (tryptophan) to 6.3% (glutamic acid) of the amino acids overall (10). However, significant variations are commonly observed, as tabulated for the soluble cellulase domain in Fig. 1 (11). The hydrophobicity of the amino acids also influences their relative surface accessibility and, thus, their ability to participate in a particular reaction. The best strategy for the modification of a particular protein, therefore, varies on a case-by-case basis, which makes it important to have a range of modification techniques that can target several different amino acids.

These considerations form the basis for the numerous techniques that are now available for the chemical modification of proteins. The sections that follow will examine these techniques and the reactive principles by which they function. A section describing reactions that display orthogonal reactivity to native protein functional groups has also been included because of the growing importance of these reactions as tools to label proteins in complex mixtures. Because it is not practical to summarize all protein bioconjugation methods here, this information instead is intended to serve as an introduction to the concepts that drive the development of these reactions. Several additional reviews and books on protein modification have been listed in the Further Reading section.

### Chemical Reactivity of Amino Acid Residues in Proteins

It should be stressed that proteins are more than just the sum of the amino acids that comprise them. Surprises abound when predicting the chemical modification behavior of a particular protein, largely because of the complex interplay between the individual functional groups and the local environments in which they reside. Nevertheless, the first step to understanding the chemical behavior of proteins is to characterize the reactive properties of the individual side chain groups. The following sections describe these aspects, with an emphasis on the chemical techniques that are commonly used to modify these residues for applications in chemical biology.

#### Chemical reactions of lysine residues

The amino groups of lysine residues are highly nucleophilic in aqueous solution, which makes them the most commonly targeted sites for covalent protein modification. As they must be deprotonated to react, these reactions are typically run at slightly elevated pH (often 8–9). However, many lysine modification reagents undergo competing hydrolysis under alkaline conditions, and thus a balance must be struck to obtain optimal modification yield. Although the average pKₐ of lysine amino groups is 10.5, this value can be altered significantly by interactions with other charged groups on the protein surface. A recent review has described an analysis technique to evaluate these effects (12).

The most common reagents for lysine modification are NHS-esters, which react with amines at rates that are significantly greater than background hydrolysis, Fig. 2a (13). These reagents are most commonly prepared from carboxylic acids and isolated before exposure to the protein substrate. Because of the popularity of this method, dozens of premade NHS-esters are now commercially available. In cases where aqueous solubility is problematic, sulfonated NHS-esters can be used. Because virtually all proteins have many lysine residues on their surface, this reaction affords the most reliable and general method...
for protein modification, although initial screens are still recommended to determine the required amount of reagent and optimal buffer conditions. It should be noted that the abundance of lysine residues also renders this technique inherently non-site selective, which leads to mixtures of products that vary in both the number and location of the modifications. Although lysine acylation is the principal reaction that occurs with these reagents, the N-terminal can also be acylated, as can a variety of other nucleophilic side chains. In the latter case, the transient species that result are often cleaved by exposing the protein to hydrazineamine before isolating the product by gel filtration.

Several related reagents are also available for lysine modification, including isocyanates, which are often prepared from acyl azides using the Curtius rearrangement (Fig. 2b), and isothiocyanates (Fig. 2c), which are more stable for long-term storage. Amine arylation can also be accomplished using 2,4-dinitrofluorobenzene (Sanger’s reagent, Fig. 2d). The products of this reaction are fluorescent, which assists their detection in sequencing applications. All of these reactions proceed under similar conditions and are often selected based on the convenience of reagent preparation from the compounds on hand.

As an alternative strategy, lysine residues can be modified through reductive alkylation, Fig. 2e. This method is most frequently carried out by exposing the protein to aldehydes in the presence of hydride-containing agents that reduce the transiently formed imines. NaB(CN)H3 and NaB(OAc)3H are commonly used for this purpose. As an alternative, transfer hydrogenation can be carried out in the presence of an Ir(III)[Cp*]2(bipyridyl) catalyst, which allows imine reduction to occur under mild conditions using buffered formate as the hydride source (14). Relative to acylation strategies, modification via reductive alkylation preserves the overall charge state (and thus the solubility) of the protein. In general, this technique also suffers from poor site selectivity.

**Chemical reactions of cysteine residues and disulfides**

By far the most widely used methods for site-selective protein modification target cysteine residues. In contrast to lysine, cysteine is one of the rarest amino acids (10), and it is unusual to find it in the reduced form on protein surfaces. Thus, it is often possible to introduce a uniquely reactive cysteine residue using genetic methods. Similar to lysine-modifying reagents, many cysteine-reactive small molecules are commercially available because of the success of this overall strategy.

In the reduced state, the sulfhydryl group of cysteine can be deprotonated (pKa ∼8) to generate a potent thiolate nucleophile. This species then can be intercepted by "soft" electrophiles, such as iodoacetamides (Fig. 3a), maleimides (Fig. 3b), acrylamides, and vinyl sulfones (Fig. 3c), to result in the formation of a new carbon-sulfur bond. Although lysine residues can also be modified using many of these reagents, cross-reactivity can often be minimized by running the reactions at relatively low pH to encourage protonation of the amine groups. Varying amounts of cysteine-alkylating reagents (from 1 to 1000 equivalents) are used to reach the desired level of modification. Although possible, the acylation of cysteine residues is not commonly used as a modification strategy because the resulting thioesters hydrolyze over time in solution.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbreviation</th>
<th>Side Chain pH5</th>
<th>Average Abundance10</th>
<th>Number in Cellulase Sequence11</th>
<th>Modify With</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lysine</td>
<td>K</td>
<td>10.5</td>
<td>5.9%</td>
<td>25 (5.8%)</td>
<td>NHS Esters</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>8</td>
<td>1.9%</td>
<td>6 (1.4%)</td>
<td>Maleimides</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>D</td>
<td>3.5</td>
<td>5.3%</td>
<td>33 (7.7%)</td>
<td>EDC + Amines</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>E</td>
<td>4</td>
<td>6.3%</td>
<td>13 (3.0%)</td>
<td>EDC + Amines</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>10</td>
<td>3.2%</td>
<td>30 (7.0%)</td>
<td>Diazonium Salts</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>n/a</td>
<td>1.4%</td>
<td>11 (2.6%)</td>
<td>Rhodium Carbenoids</td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>6</td>
<td>2.3%</td>
<td>10 (2.3%)</td>
<td>Bis(2-dimethylaminopyridyl)</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>n/a</td>
<td>2.3%</td>
<td>10 (2.3%)</td>
<td>Cyanogen Bromide</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>12.5</td>
<td>5.1%</td>
<td>5 (1.2%)</td>
<td>Glyoxal Compounds</td>
</tr>
</tbody>
</table>

Figure 1: Tabulated data for amino acid side chains commonly targeted in chemical modification reactions. As a specific example, a space-filling model of a soluble cellulase domain from C. cellulolyticum (PDB ID 1IA7) shows the relative abundance and surface accessibility of these residues. Examples of commonly used modification reagents are also listed.
Cysteine-alkylating groups have been coupled to small molecules that bind to specific sites on a protein to gain an additional level of labeling selectivity. As one example, a weakly reactive fluoromethyl ketone was fused to pyrrolopyrimidines known to inhibit a series of kinases. When binding to two kinases (RSK1 and RSK2), selective alkylation was observed for a cysteine residue adjacent to the binding site (15). Ligand-receptor interactions have also been implicated in the modification of ion channels. In these studies, maleimides bearing channel-blocking groups were used to alkylate genetically introduced cysteine residues. After modification, the activity of the ion channel could be controlled by the photoisomerization of an attached azo moiety (16).

A second method for the modification of cysteine residues involves the formation of disulfide bonds (Fig. 3d). In this method, reduced cysteine residues exchange with exogenous disulfides, which leads to the formation of new disulfides on the protein surface. This reaction can be driven to completion by mass action using a large excess of disulfide, or its efficiency can be improved by using asymmetric disulfides that possess good leaving groups. Disulfides of 2-thiopyridone (17) and 3-carboxy-4-nitrothiophenol (Ellman’s reagent) are commonly used for this purpose. To assist with the thermodynamic considerations of these reactions, a helpful table of disulfide reduction potentials has been published (18). A unique aspect of this class of cysteine modification reactions is that they are reversible, which is a feature that could be exploited for drug delivery applications (19).

Disulfides themselves are often modified during protein analysis. Most commonly, this modification is accomplished by reducing them with dithiothreitol, mercaptoethanol, or tris (2-carboxyethyl)phosphine (20) and following this reduction by a subsequent alkylation step using any of the reagents listed above, Fig. 3e (21). An interesting version of this procedure has been developed to link both of the cysteines that result from the reduction step through two consecutive alkylation reactions (22). The ability of this technique to replace a disulfide bond with a more robust link could be used to improve the stability of proteins, hormones, or antibodies for in vivo applications. To distinguish reduced cysteines from disulfides in proteolytic digests, a biotin-based affinity capture strategy has been developed (23).

Chemical reactions of aspartic and glutamic acid residues

The surface of most proteins displays many aspartic and glutamic acid residues, which provides an additional set of locations for protein modification. The carbohydrate groups are predominately deprotonated at neutral pH ($pK_a = 3.5–4$), which makes it generally difficult to distinguish between the many copies that are present. In cases where site selectivity is not a concern, the carbohydrate functional groups can be activated through the use of water-soluble carbodiimides, such as N-$\text{N'}$-$\text{N'}$-dimethylaminopropyl carbodiimide (EDC), Fig. 4. This reagent forms an $\text{O}$-acylisourea intermediate that can react with amines to form amides. This intermediate can also hydrolyze to regenerate the carboxylate anion, and in some cases it will undergo an acyl shift to yield an $\text{N}$-acyl urea. This latter pathway results in a permanent modification of the protein and in the deactivation of the carboxylate group toward more modification. Nucleophilic catalysts (such as NHS and HOBT) make it generally difficult to distinguish between the many copies that are present. In cases where site selectivity is not a concern, the carbohydrate functional groups can be activated through the use of water-soluble carbodiimides, such as N-$\text{N'}$-$\text{N'}$-dimethylaminopropyl carbodiimide (EDC), Fig. 4. This reagent forms an $\text{O}$-acylisourea intermediate that can react with amines to form amides. This intermediate can also hydrolyze to regenerate the carboxylate anion, and in some cases it will undergo an acyl shift to yield an $\text{N}$-acyl urea. This latter pathway results in a permanent modification of the protein and in the deactivation of the carboxylate group toward more modification. Nucleophilic catalysts (such as NHS and HOBT)
have been shown to suppress this pathway (24). When lysine side chains supply the amino groups, this reaction results in protein cross-links that can be used for topology mapping purposes. As one example, this technique has been used to determine the number of individual protein subunits in an ion channel, which affords information about its multimeric state (25).

**Chemical reactions of tyrosine residues**

Although the chemical modification of tyrosine residues has enjoyed a long history, this residue remains an underused target for bioconjugation reactions. It is typically modified through electrophilic aromatic substitutions (EAS), which makes its reactivity distinct from other amino acid side chains. This reaction complementarity is particularly useful in cases when cysteine chemistry cannot be used or in combination with cysteine chemistry when multiple labels need to be introduced in specific locations. Although tyrosine is a relatively common amino acid, it is often buried in the interior of the protein. This placement effectively reduces the number of phenolic side chains that can participate in reactions, which often leads to higher site selectivity than can be expected for lysine modification strategies. In general, EAS reactions are selective for the positions adjacent to the phenolic hydroxyl group, and in some cases (such as iodination) two additions can be observed for a single ring. When increased reactivity is needed, reactions are often run at elevated pH (pH 8–10) to access an appreciable population of the more nucleophilic phenolate anion.

(a) Alkylation with Iodoacetamides

(b) Alkylation with Maleimides

(c) Alkylation with Vinylsulfones

(d) Formation of Disulfides

(e) Reduction of Disulfides

![Figure 3](image)

**Figure 3** Chemical modification strategies for cysteine residues.

![Figure 4](image)

**Figure 4** Chemical modification of aspartic and glutamic acid residues with EDC.
The oldest method for the modification of tyrosine residues occurs through the use of diazonium salts, as reported in 1904 (26). This reaction results in the formation of azo compounds that are brightly colored. Both tyrosine and histidine participate in these reactions, and it has been reported that aliphatic amino groups can react to form unstable triazenes (27, 28). Thus, it should not be assumed that tyrosine is the only modified amino acid without additional confirmation.

In the case of tyrosine, these reactions typically require the participation of the phenolate anion and thus are carried out at pH 8-10 at 4°C (26). In most cases, the diazonium salts are generated through the reaction of anilines with sodium nitrite under acidic conditions and are used immediately. This reaction has been used to produce antigens (30) and to elucidate the requirements of tyrosine and histidine in enzyme active sites (31). The azo products have also served as probes of amino acid surface accessibility (29), thus providing an indirect method to determine protein conformation. One recent example has used this reaction to immobilize pig liver esterase on silica to prepare a reusable catalyst for the hydrolytic kinetic resolution of racemic esters (32).

Diazonium coupling reactions have also been applied to the modification of protein assemblies, reaching nearly quantitative conversion for the modification of tyrosine residues that are displayed on M S2 (33) and TMV (34) viral capsids. These proteins are particularly good targets for this reaction, as each capsid monomer possesses a single solvent-exposed tyrosine residue and contains no histidines. In these studies, it was found that electron-withdrawing substituents in the 4 position of the diazonium salt significantly enhanced the efficiency of the reaction. These modification reactions have been used to introduce hundreds to thousands of ketone and aldehyde groups for the subsequent attachment of MRI contrast enhancement agents (35) and water-solubilized carbon nanotubes (36).

Phenol nitration with tetranitromethane (37) (TNM) is typically run at pH 8 at room temperature and is selective for tyrosine residues under these conditions (although some oxidation of cysteine residues has been reported) (38). The product of this reaction could be thought to develop through an electrophilic aromatic substitution reaction, but the mechanism instead has been shown to proceed through a radical coupling pathway (39). The degree of nitration is easily quantified by measuring the absorbance of nitrotyrosine at 428 nm. If the reaction is performed under more rigorous conditions, such as with increased equivalents of TNM, then histidine, tryptophan, and methionine can also react.

The in vivo nitration of tyrosine by peroxynitrite and other reactive nitrogen species is implicated in many disease states and is an area of active research (40). It has recently been proposed that tyrosine nitration, mediated by nitric oxide and superoxide, is a regulated cell signaling pathway that provides quick response to the microenvironment of the cell (41–43).

The iodination of proteins has proven to be invaluable in the area of biomedical research. These techniques allow the biodistribution of biomolecules to be tracked after they have been labeled with radioisotopes, such as iodine-131. Although early studies used molecular iodine and bromine for this purpose (37), more efficient reagents have since been developed to make better use of small quantities of radioisotopes. An improved iodination method that uses iodine monochloride (ICl) has been exceptionally useful for protein radiolabeling, with efficiencies as high as 60–70% being reported, Fig. 5b (44). Since its original disclosure, several modifications of this reaction have been reported to improve efficiency and reduce handling hazards (45–47). An alternative Chloramine-T based method has also been developed to achieve both low and high levels of specific radioactivity, with iodine use as high as 60–75% (48).

Formaldehyde-induced cross-linking is commonly used for many biochemical applications, from studying protein-protein interactions to immobilizing whole cells on surfaces. Although a detailed early review of the effect of formaldehyde on proteins (49) implicated tyrosine as one residue responsible for cross-linking, a few reports have been written regarding the rigorous chemical characterization of this mechanism with modern analytical techniques. A recent investigation into the nature of these cross-links used extensive enzymatic degradation and MS/MS analysis of insulin that had been exposed to formaldehyde to determine the relative level of participation of different residues and the yield of specific modifications (50).
In follow-up studies, it was demonstrated that nickel-catalyzed irradiation times of less than one second. In this reaction, the acceptor, effectively cross-links tyrosine-containing proteins with persulfate, which generates Ru (III) as the one-electron oxidant.

In a more recent advance in tyrosine modification chemistry has been the development of a Mannich-type reaction that involves the condensation of an amine derivative with aldehydes, followed by the nucleophilic addition of tyrosine to generate a new carbon–carbon bond, resulting in peptide bond cleavage. This procedure thus provides a useful tool for the study of pro-enzyme activation.

Chemical reactions of tryptophan residues

Several features of tryptophan residues render them strategic targets for chemical modification. Tryptophan is the most fluorescent of the native amino acids, and the emission wavelength and intensity of this residue are often used to assess changes in the folding state of the protein or to detect ligand binding. Through the chemical alteration of this residue, these properties could be enhanced, shifted, or abrogated to determine the environment of a single tryptophan residue in the presence of others. As tryptophan residues are often hydrophobic contributors in binding sites, the modification of this residue could offer a way to modulate substrate binding.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

A more recent advance in tyrosine modification chemistry has been the development of a Mannich-type reaction that involves the condensation of an amine derivative with aldehydes, followed by the nucleophilic addition of tyrosine to generate a new carbon–carbon bond, resulting in peptide bond cleavage. This procedure thus provides a useful tool for the study of pro-enzyme activation.

Chemical reactions of tryptophan residues

Several features of tryptophan residues render them strategic targets for chemical modification. Tryptophan is the most fluorescent of the native amino acids, and the emission wavelength and intensity of this residue are often used to assess changes in the folding state of the protein or to detect ligand binding. Through the chemical alteration of this residue, these properties could be enhanced, shifted, or abrogated to determine the environment of a single tryptophan residue in the presence of others. As tryptophan residues are often hydrophobic contributors in binding sites, the modification of this residue could offer a way to modulate substrate binding. Studies have also revealed the importance of tryptophan side chains as mediators of electron transfer (62), which suggests opportunities to tune the electrical potential of tryptophan side chains through derivatization.

As tryptophan residues are often hydrophobic contributors in binding sites, the modification of this residue could offer a way to modulate substrate binding. Studies have also revealed the importance of tryptophan side chains as mediators of electron transfer (62), which suggests opportunities to tune the electrical potential of tryptophan side chains through derivatization. It is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.

The electron-rich indole ring of tryptophan is susceptible to attack by electrophilic reagents. Although the 3-position is typically the most nucleophilic site of the aromatic ring, the alkyl substituent present in tryptophan residues can direct reactions to occur at the 2-position instead. Electrophilic halogen species, such as N-bromosuccinimide (NBS) (63) and in situ-generated dimethylchlorosulfonium ions (64) have been used for some time to effect the oxidation of tryptophan residues through the intermediary of halohydrins. As it is the rarest of the amino acids (particularly on protein surfaces), its selective modification could provide a strategic handle for bioconjugation when cysteine chemistry cannot be used. Because of its importance, a growing number of techniques are becoming available for the modification of this residue.
3-bromo-2-(2-nitrophenylsulfonyl)-3-methylindole (BNPS-skatole) is often used (66). It was recently reported that the tryptophan residues of proteins could be nitrated by the action of peroxynitrite (67). This reactive nitrogen species (RNS) is generated from the reaction of nitric oxide with superoxide at a rate that is ten times greater than the destruction of superoxide by dismutases. The authors propose that the nitration of tryptophan, although less common than tyrosine nitration, could serve to modulate the function of some proteins. However, at this time there is no in vivo evidence for tryptophan nitration by RNS, and this has yet to be reported. One common way to modify tryptophan residues occurs through the addition of 2-hydroxy-5-nitrobenzyl bromide (HNB). Also known as Koshland’s reagent (68), this compound first alkylates the 3-position of the ring and eventually leads to the formation of a product mixture through subsequent intramolecular cyclization pathways, Fig. 6c. This reagent exhibits good tryptophan selectivity, although some levels of cysteine cross-reactivity (about fivefold less) are sometimes observed. Multiple additions to a single tryptophan residue have also been reported (69). Nonetheless, this compound has long served as the reagent of choice to estimate the tryptophan content of proteins (70, 71), and it has been used extensively to probe the role of tryptophan in the active sites of enzymes (72–74).

As an alternative reaction pathway, a recent report has shown that the nitrogen atom of the indole ring can be modified using malonyl dialdehyde derivatives with quantitative conversion, Fig. 6d (75). When carried out under strongly acidic conditions, such as 50% aqueous trifluoroacetic acid, competing Schiff base formation and arginine modification can be avoided. Although the strongly acidic conditions required for this reaction will undoubtedly denature protein substrates, this reaction offers a convenient method for the identification of tryptophan-containing peptides for proteomics applications. The aldehyde group remaining in the product can be derivatized through hydrazine...
formation, which allows enrichment of tryptophan-containing peptides through solid-phase capture (76).

A transition metal catalyzed reaction has been developed to modify tryptophan residues on proteins with high chemoselectivity, Fig. 6c (77). This reaction involves metallocarbenoid intermediates generated in situ through the degradation of vinyl diazo compounds by sodium carbonate salts. The resulting species react with indole rings to form a mixture of N- and 2-substituted products, likely proceeding through direct N-H insertion and cyclopropanation/ring opening pathways, respectively. Although no organic cosolvent is required for the reaction to occur, the addition of small amounts of ethylene glycol can be used to achieve solubilization of the diazo compound. Hydroxyamine hydrochloride was found to promote this reaction significantly, presumably by binding the catalyst and attenuating the reactivity of the metallocarbenoid species. However, the use of this additive requires that the reaction be carried out under low pH conditions (pH 3.5) that would be expected to denature most proteins. A subsequent study has reported that diazido-1,6-heptadiene-3,5-dione derivatives can be used at neutral pH without the use of this additive (78). This report demonstrated the use of this approach for the attachment of fluorescent probes to a tryptophan residue on beta-lactoglobulin.

Chemical reactions of other amino acid residues

Several other amino acids are known to react selectively with chemical reagents. Although they are less commonly targeted for bioconjugation, the modification of these residues is often used for the purposes of analysis or to annihilate catalytic activity. As one example, histidine residues can be acylated using diethylpyrocarbonate (Fig. 7a), a commonly used strategy to deactivate RNase when intact RNA molecules are required. Methionine residues react selectively with cyanogen bromide (CNBr) (79), which results in the formation of cyclic imidates that can be cleaved during acidification (Fig. 7b). This method has long been used to degrade full-sized proteins into smaller peptides for sequencing analysis. Methionine residues are reactive toward oxidants, such as periodate, yielding sulfoxides. Arginine residues can be modified selectively using dicarbonyl compounds (Fig. 7c) (80). Both single- and double-addition products are obtained in these reactions. Serine and threonine residues are generally difficult to modify in aqueous solution unless they are present in the active sites of proteases (8) or at the protein N-terminus, as described below.

Modification of the N- and C-Terminal

As multiple copies of even the rarest amino acids are often present, cases exist in which none of the chemical modification strategies described above can be used to functionalize a single site. To address this challenge, many groups have developed reactions that take advantage of the unique chemical reactivities of the polypeptide termini. In addition to providing improvements in site selectivity, these methods have been used to form native-like peptide links in some instances.

Historically, the most important technique to target the N-terminus has been the Edman degradation, Fig. 8a (81). In this sequencing method the N-terminal amino group is first reacted with phenylisothiocyanate to form a thiourea. During subsequent acidification, the sulfur atom attacks the proximal amide bond, which results in the removal of the first amino acid from the polypeptide chain. The initially formed thiazolidine quickly rearranges under the acidic conditions to form the more stable thiohydantoin. HPLC analysis then is used to identify the liberated residue. The remaining polypeptide chain bears a new N-terminal amino acid that can be analyzed in a subsequent sequencing round. Peptides up to 50 amino acids in length can be analyzed in this fashion.

The N-terminal amino group is less basic (pKₐ = 6–8) for the protonated form (82) than lysine amino groups, which makes it possible to achieve a degree of selective acylation by using NHS-esters at relatively low pH. However, this strategy usually does not yield absolute site selectivity because of the large number of competing lysines that are present. More selective N-terminal modification reactions also involve the side chain functionality of the first amino acid. A common example is the oxidation of beta-aminoo alcohols of N-terminal serine and threonine residues to yield aldehyde groups for subsequent modification (Fig. 8b) (83, 84). N-terminal serine and cysteine residues can also be condensed with aldehydes to form oxazolines and thiazolines, respectively (Fig. 8c) (85). Both of these links are stable at high pH, but display varying hydrolysis rates in acidic media. Aldehyde groups are known to react readily and irreversibly with N-terminal tryptophan residues, yielding high yields of Pictet-Spengler products (Fig. 8d) (86). Finally, N-terminal cysteine residues can participate in native...
Proteins, Chemistry and Chemical Reactivity of

(a) Edman Degradation with Phenylisothiocyanate

(b) Oxidation with Periodate

(c) Oxazolidine and Thiazolidine Formation

(d) Pictet-Spengler Reaction

(e) Transamination

Figure 8 Reactive strategies targeting the N-terminus.

chemical ligation with thioesters, as detailed in the next section (see Fig. 9). As an alternative to these techniques, reactive functionality can also be introduced at the N-terminus using biomimetic strategies. Early reports by Dixon indicated that reactive functionality can be accomplished by exposing proteins to aldehydes (such as glyoxylic acid) and Ni$^{2+}$ or Cu$^{2+}$ ions (87, 88). More recently, a metal-free version of this reaction has been reported using pyridoxal phosphate (PLP) to effect the transamination reaction (89). This method proceeds in buffered aqueous solution at pH 6.5 at 22–50°C. Several N-terminal amino acids, such as glycine and aspartic acid, have been shown to be compatible with this method, whereas tryptophan can react with the aldehyde of PLP through other pathways, as described in the previous paragraph. This technique provides a convenient method for the site-specific modification of several protein targets and has been demonstrated for the attachment of polymerization initiators (9) and surfaces to proteins through well-defined links (90). This method has also been applied to the modification of monoclonal antibodies in sites that are adjacent to the antigen binding regions (91).

Modification of the C-terminus

The native chemical ligation (NCL) has enjoyed tremendous success as one of the few methods for C-terminal protein modification. In its most general form, this approach is a ligation between a peptide that contains a C-terminal thioester and a peptide that contains an N-terminal cysteine, which results in a “native” amide bond (Fig. 9) (92). The full power of this protein synthesis technique is realized in the expressed protein ligation (EPL), a variant of the NCL in which the thioester-containing peptide is produced recombinantly (93). In EPL, the N-terminal portion of the desired peptide is expressed contiguous with an intein domain. The intein catalyzes an N-to-S acyl shift to form a thioester between the peptide target and a cysteine thiol at the end of the intein sequence. During the addition of a functionalized N-terminal cysteine, a transthioesterification occurs to join the two coupling partners and release the intein. A final S-to-N acyl shift forms the native peptide bond to yield the desired product. A addition small-molecule thiols are often added to accelerate this reaction.

A wide variety of biologically relevant molecules have been linked to the C-terminus of proteins using the EPL approach, ranging from naturally occurring biomolecules, such as lipids and polysaccharides, to non-natural probes like fluorophores, affinity tags, and metal chelating agents (94). In combination with solid-phase peptide synthesis (SPPS), EPL has made valuable contributions to proteomics by enabling the synthesis of proteins containing posttranslational modifications that are difficult to introduce (95). This approach has been used to attach fluorophores to proteins for the detection of protein–protein interactions by emission wavelength shifts (96) or through FRET (97). The site-specific incorporation of stable isotopes has been used also to facilitate NMR structural studies. In one example, two intein-splicing events were used to 15N-labell an internal
Bioorthogonal Methods for Protein Labeling

Although this review has focused on the reactivity modes that can be used to modify the amino acids directly, it is equally important to be able to predict what will not react with the native functional groups of proteins. Reactions that can proceed in aqueous solution while ignoring the natural amino acids are exceptionally useful for the development of secondary labeling strategies. In these approaches, one reaction described above is used to attach a chemically distinct functional group to a particular amino acid. A second reaction is then used to install a desired label that might have interfered with the chemistry of the first step. The targeting of unique chemical functional groups is also important for the labeling of a single biomolecule that is present among many others, as is the case in cell lysates, living cells, or entire organisms (101).

In particular, the development of “bioorthogonal” labeling reactions has been propelled by the availability of new tools for the direct incorporation of artificial functional groups into biomolecules. For proteins, powerful techniques have been developed for the introduction of new amino acids on the translational level, both through stop codon suppression (102) and amino acid codon reassignment (103). Metabolic engineering techniques have also been developed for the incorporation of artificial groups into carbohydrates (104, 105), and sequence-selective enzymatic labeling techniques have been used to introduce lipids (106) and cofactors (107, 108) that are substituted with new functionality.

The purpose of this section is to provide a brief list of the new chemical strategies that have been developed to target these artificial functional groups because the importance of these techniques will undoubtedly grow in the coming years.

Methods targeting carbonyl groups

The first methods that were used to label non-native functional groups on proteins targeted aldehydes and ketones. These groups can be introduced directly through the periodate oxidation of N-terminal serine residues (83), transamination (89), or carbohydrate oxidation with periodate (100). They can also be installed using metabolic engineering (104) or translational techniques (110). Aldehydes and ketones condense selectively with hydrazine and alkoxyamine derivatives to form hydrazine- and oxime-type derivatives, respectively (Fig. 10a) (110, 104).

Both reactions are carried out in aqueous solution using an excess of the ketone-reactive reagents. Early mechanistic studies determined that the rate-limiting step in the reaction is the dehydratation of the tetrahedral intermediate that is formed after nucleophilic attack on the carbonyl group (111). Therefore, mild acidic conditions (e.g., pH 6.5) are typically used to accelerate the dehydration step without fully protonating the nucleophilic reagents. This reaction exhibits excellent chemoselectivity and works well for protein labeling with a diverse array of functional groups. However, the presence of competing ketone and aldehyde metabolites limits its use for protein labeling in crude cell lysates.

Methods targeting azides

More recently, azide groups have emerged as popular targets for chemoselective protein modification. They seem to be ignored completely in a variety of biological settings (112), and yet they possess a favorable thermodynamic reaction potential through the loss of nitrogen gas (113). They are also small in size and thus can be incorporated using a variety of enzymatic, metabolic, or translational techniques. The azide group also provides a unique IR chromophore for spectroscopic characterization convenience.

The first technique that was used to modify azide groups in a biological setting was a modified Staudinger ligation (105),
Proteins, Chemistry and Chemical Reactivity of

(a) Modification of Ketones and Aldehydes

(b) Staudinger Ligation with Azides

(c) Copper-Catalyzed Huisgen Cyclization of Azides

(d) Strain-Promoted Huisgen Cyclization of Azides

(e) Oxidative Coupling of Anilines

Figure 10  Chemoselective modification strategies targeting non-native functional groups.

which used a pendant ester group to capture the iminophosphorane intermediate. The resulting amide bond served to link additional functionality on the phosphine to the protein permanently (Fig. 10b). Although the original version of this reaction incorporated the triarylphosphine oxide into the coupling product, more recent "traceless" versions have appeared (114, 115). When used in conjunction with native chemical ligation techniques (see above), this reaction can serve as a means to synthesize full-sized proteins (99). A version of this technique has also been developed for the fluorescent detection of azido groups (116). Perhaps the ultimate testament to the bioorthogonality of this method is its reported use in living animals (112), which renders it a uniquely useful approach for the study of glycans using in vivo imaging techniques (117). A detailed mechanistic investigation of this reaction has appeared (118).

A second mode of reactivity for azide groups occurs through a [3 + 2] cyclization reaction with alkynes, as originally described by Huisgen (119). Although this reaction occurs thermally in the absence of additional reagents, a critical advance for the targeting of azide groups on biomolecules was the observation that Cu(I) salts catalyzed the formation of the triazole products in aqueous solution at ambient temperature (120, 121) (Fig. 10c). This technique has been used for the labeling of viral capsids (122, 123) and the surface of bacterial cells (124), and it has been used to attach proteins to surfaces (125). Both azide and alkyne groups have been introduced on the translational level to provide the appropriate coupling partners (124, 126). This reaction has been used with particular success as a detection method for the identification of new proteases, which were labeled with azides at active sites (127). The monumental success of this reaction for the facile construction of diverse molecules in organic synthesis suggests that it will have a very bright future ahead for biomolecule labeling. The mechanism of the copper-catalyzed reaction has been described (128).

For situations in which copper ions are observed to interfere with protein function or to exhibit toxicity, a metal-free version of this reaction has been developed. This strategy uses a strained cyclooctyne as the alkyne component, which readily undergoes the Huisgen cyclization with azides at room temperature (Fig. 10d) (129). Although unsubstituted cyclic alkynes participate in the reaction with relatively slow rates, a subsequent report has shown that the presence of electron-withdrawing fluorine substituents can accelerate the cycloaddition significantly (130). Taken together, this set of reactive pathways renders the azide group one of the most promising targets for bioconjugation reactions in which complete functional group tolerance is required.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Modification through oxidative coupling reactions

An alternative method has been developed for the modification of proteins through the oxidative coupling of aniline groups (131) (Fig. 30c). This method targets aminophenylalanine, which is one of the most successful amino acids to be introduced into protein sequences using the stop codon suppression technique (132). In the presence of oxidants, such as periodate or cerium (IV), this group reacts rapidly with phenylamine derivatives to afford adducts that are highly stable toward hydrolysis, reduction, or oxidation. The reaction displays very high chemoselectivity, although the reliance on periodate leads to varying levels of methionine oxidation in addition to the desired products.

Conclusion

Although protein bioconjugation has enjoyed a long and successful history, recent years have witnessed a dramatic increase in the number of available techniques for chemical modification. This increase has developed largely because of the availability of high-resolution mass spectrometry and NMR characterization techniques, which have accelerated the rate with which detailed chemical information can be obtained for the modification products. With the addition of each new reaction, previously inaccessible protein bioconjugates of ever-increasing complexity can be realized. The chemical underpinnings of these reactions also form the basis for future modification strategies and expand our understanding of protein reactivity in living systems.

References

Proteins, Chemistry and Chemical Reactivity of


Further Reading


Further Reading

Chemical Modification of Proteins

See Also

Chemical Modification of Proteins

Amino Acids, Chemical Properties of

Enzyme Catalysis, Chemical Strategies for

Proteins, In Vivo Chemical Modifications of

Peptides, Chemistry of
The Chemistry of Metallo-Enzymes and Metallo-Proteins

Robert J.P. Williams, Oxford University, Oxford, United Kingdom
doi: 10.1002/9780470048672.wecb327

Metallo-enzymes and metallo-proteins are a combination of a metal ion or several metal ions in which the metal ion is strongly held in a protein. Hence, before describing their properties and functions, it is necessary to appreciate which metal ions bind strongly to proteins and exchange slowly. The biological significance of metallo-enzymes is that they are essential catalysts, especially for the metabolism of small molecules such as H₂, CH₄, CO, N₂, and O₂ as well as for long-range electron transfer. The control of uptake and the synthesis of their proteins rest with feedback in a series of metallo-proteins.

Before describing these enzymes (see the examples in Table 1), it is necessary to make a broad division of metal interactions in enzymes. In one case, the metal ion is firmly attached to the protein so that like a nonmetal of an amino acid it does not exchange within days. In such a situation, the isolation of the metallo-protein, correctly named, can be followed through all steps of purification until further purification procedures fail to alter the stoichiometry of the metal/protein, and this is a whole number. At the same time, if the metallo-protein is an enzyme, activity will become optimal. This approach to metallo-enzymes was first developed by Professor B. L. Vallee of Harvard. A different extreme is one in which the metal ion is loosely attached to the protein when isolation procedures may well result in an apoprotein, and if it was an enzyme, it would then be devoid of activity. The problem in such a case is to know which metal ion was intrinsically involved in activity in vivo. Now a great number of metal ion/protein interactions are intermediate in character between extremes so that purification results in the discovery of 'fractional stoichiometry' and can be beset by contamination. Confusion is increased because many enzymes have several metal ion centers when it may require astute experimentation to reveal the nature of the original metal ion complement. We must be aware of both binding constants of metal ions to proteins of different kinds and of their rates of exchange so as to establish their nature correctly as it is related to function (1–4).

Binding Constants

Over many years orders of binding strengths of metal ions have been established. Some general features are that the ions of sodium and potassium bind very weakly if at all, those of magnesium and calcium bind more strongly, and certain transition metal ions bind with very considerable strength. In more detail, the binding of divalent ions to virtually all ligands follows the Irving-Williams order and the metal ions toward the end of the series are those that form stable metallo-enzymes.

\[
\text{Mg}^{2+} < \text{Mn}^{2+} < \text{Fe}^{2+} < \text{Co}^{2+} < \text{Ni}^{2+} < \text{Cu}^{2+} > \text{Zn}^{2+}
\]

Binding strength of the monovalent Cu⁺ is large like that of Cu²⁺, but binding of trivalent ions may also be very strong, especially Fe³⁺ and Mn³⁺. (Changes of oxidation state can occur during isolation when the properties of interest may be hidden.) A further specific difficulty is that the binding of Ca²⁺ is extremely variable in magnitude relative to that of Mg²⁺. Although the above sequence is roughly that of increasing electron affinity and hence of binding constant strength, as the ions do not differ greatly in size, the size of Ca²⁺ is very different so ligands can be devised for it that are of greater binding strength than Mg²⁺ by using structural constraints on ligand conformation, creating a disadvantage for Mg²⁺ binding.

The fact that so many metal ions are available and that they vary in availability in environmental waters means that specifically separating each one with a given ligand, assuming at first that equilibrium is attained (i.e., that is, binding constant orders are obeyed), make for a problem of selection. It is largely solved in cells as follows; see Fig. 1.

A cell can limit the free concentration of a metal ion in a given compartment by using energized pumps either inwardly or outwardly directed in the containing membrane of the compartment (Fig. 1). It can also use selective reagents outside the cell, siderophores, or proteins to scavenge for metal ion,
The Chemistry of Metallo-Enzymes and Metallo-Proteins

Table 1 Examples of metallo-enzymes

<table>
<thead>
<tr>
<th>Class of catalysis</th>
<th>Example of active site Mn⁺</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acid/Base Hydrolysis</td>
<td>Zn (Mg, Co, Ni, Ca)</td>
<td>Carbonic Anhydrase</td>
</tr>
<tr>
<td>Electron Transfer</td>
<td>Fe, heme (Fe, Cu)</td>
<td>Cytochromes</td>
</tr>
<tr>
<td>Oxidation (H₂O₂, RO₂⁻)</td>
<td>Fe, heme (Fe, Cu)</td>
<td>Catalase, Peroxidase</td>
</tr>
<tr>
<td>Hydroxylation</td>
<td>Ni, Fe</td>
<td>Hydrogenase</td>
</tr>
<tr>
<td>Group Transfer (I-CH₃)</td>
<td>Co (B₁₂)</td>
<td>Methylmalonyl Isomerase</td>
</tr>
<tr>
<td>Group Transfer (OPO₃⁻)</td>
<td>Mg (weakly bound)</td>
<td>Kinase</td>
</tr>
<tr>
<td>Group Transfer (CO)</td>
<td>Ni (F₄₃₀₀)</td>
<td>Acetyl Synthetase</td>
</tr>
</tbody>
</table>

NOTES: Organic side chains and metal ions can be substituted. The organic side-chain substitution is usually done by gene mutation, but metal ion substitution is done by direct exchange and, in fact, can be done for S and Se.  

In these cases, the metal ion is in a metal complex.

particularly iron, bringing them into the cell where they may be reduced (3). In this manner, the concentration of ions in a cell can be fixed; for example, Mg²⁺ at about 10⁻³ M in the cytoplasm greatly exceeds that of all other ions. The cell has certain binding agents, potential ligands (L) that, although they bind in a given order of strength, Mg²⁺ < Mn²⁺ < Fe²⁺ < Cu²⁺ < Ni²⁺ < Co²⁺ < Zn²⁺, show only a modest change from Mg²⁺ to Cu²⁺ (e.g., the pyrophosphate of ATP). Given that concentration of the ion bound depends on the product K[M²⁺][L] = [ML], the high value of [M⁺] for Mg²⁺ makes it the only ion able to bind to a very weak-binding ligand L, which has a general binding constant between 10³ M⁻¹ and 10⁴ M⁻¹. In fact, only Mg²⁺ is found in many enzymes using ATP, where Mg²⁺ ATP is the substrate. No other metal ion is observed in these metal/ligand/protein complexes (e.g., the kinases).

Now consider the opposite extreme of binding strength Cu²⁺(Cu⁺). Its concentration in a cell is limited to around 10⁻¹⁵ M by pumps, so a very strong binding ligand is required to retain copper. A Cu²⁺(Cu⁺) has such a strong binding constant, it alone can still form its Ml complex so long as the free concentrations of the other metal ions in the product K[M][L] are weaker than for Cu²⁺, which then links Cu to a given L, but if excess [L] over [CuL] were allowed, then the excess of [L] above that of [CuL] would let this ligand, L, bind other metal ions. This possibility is prevented as the free [Cu] binds to a transcription factor for the synthesis of its L (Fig. 1), with equal strength such that when free [Cu] is at a specific

![Figure 1](image-url)  
Figure 1: An example of the way metallo-enzymes are under controlled formation through both controlled uptake (rejection) of a metal ion and controlled synthesis of all the proteins connected to its metabolism and functions. The example is that of iron. Iron is taken up via a molecular carrier by bacteria but by a carrier protein, transferrin, in higher organisms. Pumps transfer either free iron or transferrin into the cell where Fe²⁺ ions are reduced to Fe⁰ ions. The Fe⁰ ions form heme, aided by cobalamin (cobalt B₁₂ control) and a zinc enzyme for α-laevulinic acid (ALA) synthesis. Heme or free iron then goes into several metallo-enzymes. free Fe⁰ also forms a metallo-protein transcription factor, which sees to it that synthesis of all iron carriers, storage systems, metallo-proteins, and metallo-enzymes are in fixed amounts (homeostasis). There are also iron metallo-enzymes for protection including Fe SOD (superoxide dismutase). Adenosine triphosphate (ATP) and H⁺ gradients supply energy for all processes. See References 1–3.
value, say $10^{-15}$ M, the production of L ceases; hence, $[\text{CuL}]$ alone is formed. No other metal ion has a sufficient $K[M][L]$ to form such complexes. The transcription factors bind to promoter regions of DNA. By a subtle combination over such controls of both free $[M]$ and free $[L]$, specific combinations can be obtained for all divalent ions. In fact, very little cross contamination of $M$ and $L$ is observed in vivo (see the cases in Table 1). (In all these considerations we must remember that $K$ is an effective binding constant in cells dependent on pH and the presence of other interacting ions or molecules. It is the size of the effective constant $K$ that makes certain that binding is selective.) We estimate the free ion concentration, molarity, in the presence of other interacting ions or molecules. It is the size of the effective constant $K$ that makes certain that binding is selective. We estimate the free ion concentration, molarity, in cells to be in the inverse order to binding strength $K$.

In contrast, the effective constant $K$ can be looked upon as a simple ratio of the rate of ligand, $L$, binding, $k_{on}$, relative to the rate of leaving $K_{eq}$ (i.e., $K = k_{on}/k_{off}$). As $k_{off}$ is limited by loss of water from around the ion, which cannot exceed $10^9$ s$^{-1}$, $k_{off}$ is always less than 1 second for $K > 10^6$ M$^{-1}$, and for $K > 10^{12}$ M$^{-1}$ the off-rate is close to or more than 1 hour. For slower on-rates, the off-rate for a given $K$ is correspondingly slower. It is these rates, often related to equilibrium binding constants, that decide whether an enzyme containing a metal can be separated in pure bound form, a true metallo-enzyme, that is in slow exchange, (8), or will need a metal ion to be added to an isolated apo-enzyme to form a metal ion protein complex that is a fast exchange (9-12). Now certain ions always have relatively fast exchange rates. Examples are Na$^+$, K$^+$, Mg$^{2+}$, Mn$^{2+}$, and sometimes Ca$^{2+}$. Examples of metal ions that have slower exchange rates and readily form metallo-enzymes are Cu$^{2+}$, Co$^{2+}$, Zn$^{2+}$, Ni$^{2+}$, Mo$^{6+}$, Fe$^{3+}$, Mn$^{3+}$, and possibly Co$^{3+}$. In an intermediate group is Fe$^{2+}$, the enzymes of which are usually included in the class of metallo-enzymes. We must also remember that some metal ions are trapped not bound directly to protein side chains but to complex ligands such as porphyrins, for example, Fe, Mg, Co, and Ni (see Fig. 2) or special dithiols such as M or see Fig. 2 and W from which exchange is slow and the metal ions are frequently in low-spin states. The bindings of these complexes is often very tight and in slow exchange, and we include their enzymes in the class of metallo-enzymes. It is generally found that these ions are inserted into the complex...
ligand by selective processes. Slow exchange is also observed for Fe₅S₄ centers, but certainly some Fe₄S₄ centers exchange at intermediate rates.

To transfer or insert the cations that are in slow exchange with aqueous cations or are very dilute in solution demands special transporting proteins (Fig. 1). The first of these to be discovered was the protein for calcium transport across cells, calbindin, a member of the S-1 class of proteins. More recently, transporters have been discovered for Ni, Cu, Mo, and for molecules such as vitamin B₁₂ (Co). In essence, a transport metallo-protein, T, with only very slow loss of the metal ion, carries the ion and delivers it from a pump to an apoprotein, A, which becomes an enzyme and implies that the transporter, sometimes called a chaperone, and the apoprotein form a complex TMA, which by exchange becomes TMA. TM must also exchange with a pump, P, in the same way PM:T becomes P:M:T. It is interesting that the genes for T, P, and A often lie close together under one promoter and transcription factor F (Fig. 1), which senses either the CM or that of the free M concentration giving F:CM, and all these metallo-proteins have rather similar binding constants and selective functional groups.

We have now described the intracellular movement leading to the synthesis of a metallo-enzyme and we see that the control of synthesis rests with the supply of M to the cell from the environment, with its genetic system, which controls synthesis of the apoprotein, with its carrier and its pumps. All must be in a tight feedback circuit so that each metal has its selected proteins and its free concentration in homeostasis, which implies that all their binding constants are closely related to the Irving-Williams series and that to obtain selectivity the binding groups of the proteins are such that thiols are common for binding Cu²⁺, Zn²⁺, Mo⁶⁺, W⁶⁺, and Fe³⁺; N-donors are more common for Fe⁴⁺ and Mn²⁺; whereas O-donors are observed to bind Mg²⁺, Ca²⁺, and K⁺. We need to go forward to the properties of these proteins as isolated and then return to their cellular functions. We shall take it that clear-cut metallo-proteins are those containing simple Cu, Zn, Ni, and Mo (W) ions including the cases of heme (Fe), some Fe₅S₄, B₁₂ (Co) and F-430 (Ni), as well as some mixed metal ion proteins of Fe²⁺/Fe³⁺, of Mn²⁺, or of NiFe and MnFe.

**Properties Of Isolated Metallo-Proteins**

Metallo-proteins have been studied by a great range of physical methods, which have revealed that frequently the metal ion is held in a constrained (sometimes called entatic) state (Figs. 3 and 4 (B-28)). The simplest explanation of this state is that the binding energy of the metal ion in the resting state of the enzyme is such as to lower the activation energy required for it to function in a catalytic act relative to that energy required for free ion activity, which is most easily understood if we refer to a free energy diagram for a reaction pathway. In the course of the reaction the substrate is transformed while the catalyst cycles. Referring to the discussion of substrate activation in a catalyst, Pauling proposed that its binding energy was used in part to distort its structure to match as far as possible the transition state for the reaction presuming the structure of the catalyst was fixed. Turning now to the metal ion, or any other part of a catalyst that acts in the catalytic act, we can examine its
bonding relative to that normally observed in the ground state of its simple compounds where its bonds are relaxed to optimal free energy binding conditions. As examples, the bond angles and bond lengths of amino acids and of metal ions in simple aqueous complexes (no steric hindrance or other constraint) are well known and the corresponding physical-chemical properties are understood. It is now known that in a protein complex the binding of any group can be strained by the need for the overall structure to have as great an activity as possible. It is usually found that a metal ion in a metallo-protein has distinctly unusual bond properties. The observed structural consequence is that the bonding geometry of any single unit, here a metal ion, can be seen to match the catalytic requirement. Some ways in which catalytic properties are achieved are for the metal ion 1) to have a small number of ligating atoms giving the metal ion an enhanced electron affinity for acid attack (Fig. 4) or 2) to have a ligand geometry matching its redox properties with those required for redox catalysis (Fig. 3). In addition, if during the catalytic cycle the metal ion has to change coordination number or oxidation state, then the constrained condition of the metal ion ensures that the relaxation energies between states in the cycle remain small. Note immediately that strong binding and non-exchanging ions are most likely to show these properties. In addition, it may be helpful if the metal ion has an open-sided structure and has a disposition of ligands to repel product while accepting substrate. Examples illustrate these generalities (see Table 2).

Table 2 Examples ofl constraints in metallo-enzymes

<table>
<thead>
<tr>
<th>Metal</th>
<th>Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu</td>
<td>Cu^{+}/Cu^{2+} held in fixed coordination allowing fast electron transfer (Fig. 3)</td>
</tr>
<tr>
<td>Zn</td>
<td>Zn^{2+} held in 5-coordination (one water molecule) which is readily adjustable (Fig. 4)</td>
</tr>
<tr>
<td>Fe</td>
<td>Heme enzymes have open-sided Fe^{2+} to give easy access to O_{2}</td>
</tr>
<tr>
<td>Co</td>
<td>Cobalt in coenzyme B_{12} (Fig. 2) has a Co-carbon bond that breaks on substrate binding</td>
</tr>
<tr>
<td>Ni</td>
<td>Ni^{2+} in close to tetrahedral geometry in hydrogenase</td>
</tr>
<tr>
<td>Mn</td>
<td>Mn^{2+} held in a peculiar cluster with calcium in O_{2}-production (Fig. 2)</td>
</tr>
</tbody>
</table>

NOTES: The constraints have been shown to be related to function in many cases.

The case of Lewis acid function of metallo-enzymes is given by many a zinc enzyme (Fig. 4). Taking the case of zinc in carbonic anhydrase, the ground state structure shows it to be open sided with easy access for H_{2}O and CO_{2} but not larger substrates. Moreover, these neutral molecules bind more readily than the product HCO_{3}-. The zinc ion is known to be open sided and to go readily between 4- and 5-coordination, which is extremely useful in the required catalytic cycle. Moreover, the pK_{a} of attached water is close to 7.0, an exceedingly low value for Zn^{2+}(H_{2}O) complexes. The reason for the selection of zinc as a biological Lewis acid catalyst is now clear. Functionally, it has optimum properties by being readily constrained in the ground state and it easily passes through different coordination...
states all the time in fast exchange internally with substrate intermediates in the site. If we look at other available Lewis acid cations, we observe that of the metal ions with possible similar advantages are 1) Mg2+, but it binds weakly and is a weak acid; 2) Ni2+, but it does not change geometry easily; and 3) Cu2+, but it would readily react with oxygen that can access the site. Only Co2+ is an ideal substitute, but it is a somewhat weaker acid than Zn2+. However, in confirmation that Co2+ is the only obvious substitute, it is found that the catalytic power of the M2+ ions in carbonic anhydrase follows the order

\[
\text{Zn}^{2+} > \text{Co}^{2+} > \text{Mn}^{2+} > \text{all other cations} \quad (0.0)
\]

Notice that this order is not the relative Lewis acid strength or binding strength to the site as demonstrated in simple aqueous complexes that follow the order

\[
\text{Cu}^{2+} > \text{Zn}^{2+} > \text{Ni}^{2+} > \text{Co}^{2+} = \text{Fe}^{2+} = \text{Mn}^{2+} > \text{Mg}^{2+}
\]

but is a result of the similar coordination geometries of Zn2+ and Co2+. In addition, Fe2+, Mn2+, and Mg2+ are used in enzymes for certain acid catalyses; however, these enzymes are not true metallo-enzymes but are metal/enzyme complexes from which the metal ion dissociates relatively easily, and hence these ions function both as a catalyst and as a free unbound regulatory ion (Fig. 1). Probably the tightly bound ions Zn2+, Cu2+, and Ni2+ only act as regulators while bound to proteins.

We turn next to a simple example of a redox metallo-enzyme—azurin, which is a copper electron-transfer protein and, as the electron has no volume, it is best if the metal ion is enclosed in a chelate from which it can escape (Fig. 2). Very well-known examples of such metallo-enzymes are provided by porphyrin and its derivatives as in Fe2+/heme, Co2+/B12, Ni/F-430, and even Mg2+ chlorin, although the last is only found in electron-transfer enzymes of photo-centers. For the three transition metal ions, the placing of them in the strong ligands of the porphyrin series generates low spin states in contrast with all the examples we have given so far where metal ions are high spin. Low spin states are more common among second and third rows of transition metal ions such as of elements Rh, Pd, and Pt, and Mn has made great use of these ions as catalysts, but they are not environmentally available to organisms, hence the resort to porphyrins. A big advantage of porphyrin is that it induces low spin states in its Fe, Co, and Ni complexes where redox change is usually of small relaxation energy so that relaxation is usually easy as various oxidation states can be held in or closely in the porphyrin plane. Of course the binding of additional ligands above and below the plane or just below it tunes the function and once again the protein can create constrained advantageous conditions. With two ligands bound, the heme iron is an excellent electron-transfer metallo-enzyme known in the series of cytochromes a, b, and c. The metal ions Cu2+/B12 and Ni2+/F-430 by contrast are naturally open sided. They, together with open-sided Fe-heme, readily bind ligands but of very different kinds. Open-sided Fe2+/heme is

\[
\text{O}_2 \quad \text{H}_2\text{O}
\]

for certain acid catalyses; however, these enzymes are not true metallo-enzymes but are metal/enzyme complexes from which the metal ion dissociates relatively easily, and hence these ions function both as a catalyst and as a free unbound regulatory ion (Fig. 1). Probably the tightly bound ions Zn2+, Cu2+, and Ni2+ only act as regulators while bound to proteins.

We turn next to a simple example of a redox metallo-enzyme—azurin, which is a copper electron-transfer protein and, as the electron has no volume, it is best if the metal ion is enclosed in a chelate from which it can escape (Fig. 2). Very well-known examples of such metallo-enzymes are provided by porphyrin and its derivatives as in Fe2+/heme, Co2+/B12, Ni/F-430, and even Mg2+ chlorin, although the last is only found in electron-transfer enzymes of photo-centers. For the three transition metal ions, the placing of them in the strong ligands of the porphyrin series generates low spin states in contrast with all the examples we have given so far where metal ions are high spin. Low spin states are more common among second and third rows of transition metal ions such as of elements Rh, Pd, and Pt, and Mn has made great use of these ions as catalysts, but they are not environmentally available to organisms, hence the resort to porphyrins. A big advantage of porphyrin is that it induces low spin states in its Fe, Co, and Ni complexes where redox change is usually of small relaxation energy so that relaxation is usually easy as various oxidation states can be held in or closely in the porphyrin plane. Of course the binding of additional ligands above and below the plane or just below it tunes the function and once again the protein can create constrained advantageous conditions. With two ligands bound, the heme iron is an excellent electron-transfer metallo-enzyme known in the series of cytochromes a, b, and c. The metal ions Cu2+/B12 and Ni2+/F-430 by contrast are naturally open sided. They, together with open-sided Fe-heme, readily bind ligands but of very different kinds. Open-sided Fe2+/heme is
Some metallo-enzymes are formed by a set of subunits. Shown here is the structure of nitrogenase that has an energized supply of electrons (energy from ATP) via two iron/sulfur clusters and an active site FeMoco shown in the inset. There are two distinct protein units linked to one supply line of a separate unit. Note there is a light non-metal in a hole in the structure of Fe7MoS9 but, as it is of uncertain nature, it is not shown. 

Figure 5

Some metallo-enzymes are formed by a set of subunits. Shown here is the structure of nitrogenase that has an energized supply of electrons (energy from ATP) via two iron/sulfur clusters and an active site FeMoco shown in the inset. There are two distinct protein units linked to one supply line of a separate unit. Note there is a light non-metal in a hole in the structure of Fe7MoS9 but, as it is of uncertain nature, it is not shown.

Further Reading

The following volumes contain detailed structures and properties of virtually all metallo-enzymes and metallo-proteins.

Transient State Enzyme Kinetics

Kenneth A. Johnson, Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas
doi: 10.1002/9780470048672.wecb609

Transient-state kinetic analysis defines the elementary steps along the pathway of an enzyme-catalyzed reaction. The methods rely on rapidly mixing a substrate with a sufficient concentration of enzyme to allow a direct observation of intermediates and products formed at the active site during a single enzyme cycle. Stopped-flow methods afford observation of reactions by monitoring changes in optical signals (i.e., absorbance, fluorescence, and light scattering) as a function of time after mixing. Analysis of the substrate concentration dependence of the rates and the amplitudes of observable species can define the sequence of events that occur after substrate binding. Rapid-chemical quench-flow methods require two sequential mixing events: one to start the reaction and a second to stop the reaction by the addition of a quenching agent. One then quantifies the amount of product formed, often by chromatography, to resolve radiolabeled substrate from the product. Because one can observe the conversion of substrate to product directly on a known concentration scale, rapid quench experiments can be interpreted directly.

In the final analysis, the most rigorous interpretation relies on global analysis fitting both stopped-flow and rapid quench-flow data simultaneously to establish a reaction pathway and the rates of interconversion of kinetically significant intermediates. Accordingly, transient-state kinetic methods allow a definition of a reaction sequence by direct measurement of each step to establish the kinetic and thermodynamic basis for enzyme specificity and efficiency.

Structural and kinetic methods provide a powerful combination of related data to explore the remarkable efficiency and specificity of enzymes. Each method fills in details left out from the other. Structural studies provide a static picture, whereas kinetic studies bring the structure to life by providing the data to define the dynamics that underlie catalysis. To relate kinetic studies directly to structure, it is crucial to examine the reactions that occur at the active site of the enzyme, namely, after substrate binding and before product release. Unfortunately, steady-state kinetic methods cannot reveal details regarding the conversion of substrate to product at the active site of an enzyme. Therefore, we turn to rapid-transient kinetic methods in which we examine the enzyme as a stoichiometric reactant rather than as a trace catalyst. We can thereby observe enzyme-bound species formed during catalysis by examining the time scale of a single enzyme turnover. The reactions that govern enzyme specificity and efficiency can then be defined by quantifying the rates of substrate binding, enzyme isomerization, chemical conversion of substrate to product, and product release.

Measurable Kinetic Parameters

Traditional steady-state kinetic studies rely on indirect observation of catalysis by monitoring the accumulation of product or consumption of substrate as a consequence of many reaction cycles with a trace of catalyst. Conclusions are limited to inference of the possible pathways for the order of addition of multiple substrates and release of products and quantification of two bulk kinetic parameters, $k_{cat}$ and $k_{cat}/K_m$. The parameter $k_{cat}$ defines the maximum rate of conversion of enzyme-bound substrate to product released into solution, but it cannot be used to establish whether the maximum rate of reaction is limited by enzyme conformational changes, rates of chemical reaction, or rates of product release per se. It does, however, set a lower
limit on the magnitude of any rate constant in that sequence. The term \( k_{cat}/K_m \) is known as the specificity constant because it quantifies which enzyme or substrate wins in a competition. Moreover, \( k_{cat}/K_m \) defines the rate of substrate binding multiplied by the probability that once bound, the substrate is converted to product and released. As such, \( k_{cat}/K_m \) sets a lower limit on the second-order rate constant for substrate binding, and it also provides a measure of enzyme efficiency relative to the diffusion-limited substrate-binding rate. The term \( K_m \) is best understood as the ratio between the two primary parameters, \( k_{cat} \) and \( k_{cat}/K_m \), and it is generally not equal to the dissociation constant for substrate binding. These steady-state kinetic parameters form the minimal background to design experiments than can define the sequence of individual steps along the reaction pathway.

Transient-state kinetic analysis allows definition of a minimal reaction sequence of kinetically significant intermediates responsible for the conversion of substrate to product at the enzyme active site. Resolution of how individual rate constants lead to a given value for \( k_{cat}/K_m \) provides an understanding of the reactions that govern enzyme specificity and efficiency. Direct measurement of events that occur at the active site of the enzyme fills the void of information left by steady-state kinetic studies, namely, the steps that occur after substrate binding through to the point of product release. Using transient kinetic methods, the rates of a single reaction cycle are measured using an enzyme as an observable reactant often at concentrations approximately equal to the substrate. For example, given the pathway shown in Scheme 1, all eight rate constants can be measured by a combination of equilibrium, steady-state, and transient-state kinetic experiments. Moreover, these kinetic studies lay the foundation for the identification of enzyme-bound intermediates (EI).

Comprehensive kinetic analysis defines the mechanistic basis for enzyme specificity and efficiency in ways that can be directly related to enzyme structure. In this article, the rationale will be described for design and interpretation of experiments to define the pathway of enzyme-catalyzed reactions using transient kinetic methods. These principles will be illustrated with three examples of biologically important reactions, none of which could have been solved with steady-state kinetic methods alone. This article is by no means a comprehensive survey of this extensive field, but rather, selected examples from the author’s laboratory will be used to illustrate the methods to provide a flavor for what is possible.

### Transient Kinetics of Biological Reactions

Enzyme catalysis is the basis for life. Enzymes accelerate desired reactions so they occur on a time scale requisite for living organisms. Moreover, enzymes can synthesize complex biological macromolecules that exist long enough to sustain life by coupling desired synthetic reactions to thermodynamically favorable catabolic reactions. Kinetic analysis allows us to decipher the pathway by which enzymes carry out these biologically important reactions with incredible speed, efficiency, and specificity. Traditional steady-state kinetic methods provide estimates for the kinetic constants, \( k_{cat} \) and \( k_{cat}/K_m \), which govern the flux through metabolic pathways. Transient-state kinetic methods allow definition of the kinetically significant reaction sequence and can provide estimates for each intrinsic rate constant along the pathway, including rates of formation and breakdown of enzyme-bound intermediates. Thus, transient-state kinetic analysis is essential to understanding how enzymes bring about catalysis. Here, three examples of biologically important reactions are provided.

#### DNA polymerase fidelity

As one example, we have long been fascinated by the incredible speed and accuracy of DNA polymerases that can replicate DNA at a rate of 300 base pairs per second while introducing errors only 1 out of a billion times. Kinetic analysis has revealed how these enzymes use information from both hydrogen bonding and shape of the base pair to discriminate correct base pairs from the three possible competing mismatches at any DNA template position. The most important step in which specificity is determined is during a change in enzyme structure after nucleotide binding, in which substrate binding energy is used to alter the enzyme conformation. After the conformational change, a correct substrate (deoxynucleotide) becomes tightly bound in a state that slows release of the bound substrate and organizes the active site residues to promote catalysis. In contrast, an incorrect substrate (a mismatched deoxynucleotide) disorganizes the active site residues leading to a state that promotes release of the substrate and slows the chemical reaction. Thus, it is the kinetic partitioning of the conformational isomer formed after substrate binding that establishes substrate specificity. Transient-state kinetic analysis has revealed how these structural dynamics endow the polymerase with the ability to replicate DNA with sufficient accuracy and speed to sustain a living organism.

#### Detection of enzyme intermediates

Enzymes often catalyze conversion of a substrate to product through a sequence of intermediate states. In many instances, the intermediates must be inferred because they are not sufficiently stable kinetically or thermodynamically to be observed directly. However, in several instances, transient-state kinetic analysis has allowed identification of enzyme-bound intermediates. Rigorous kinetic analysis is essential to distinguish intermediates from the side products of the reaction. As one example, excitatory postsynaptic potential (EPSP) synthase, which is the target of the herbicide glyphosate, catalyzes the reaction of shikimate-3-phosphate (S3P) with phosphoenolpyruvate (PEP) through a tetrahedral intermediate to form the product, EPSP.

**Scheme 1**

\[
E + S \rightarrow ES \rightarrow\begin{array}{c}
E + P \rightarrow EP \rightarrow EI \rightarrow E + P
\end{array}
\]

**WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.**
for more effective herbicides. Transient-state kinetic studies revealed the rapid formation and breakdown of the intermediate in less than 100 milliseconds. In contrast, structural studies by nuclear magnetic resonance suggested another species (EPSP ketal) as an intermediate, but subsequent kinetic analysis showed that its rate of formation at the active site was a million-fold too slow to account for catalysis (9, 10). Thus, kinetic competence is a critical criterion for the identification of intermediates; the intermediate must be formed and broken down at rates sufficient to account for the net conversion of substrate to product at the enzyme active site (11). Single-turnover transient-state kinetic studies provide the crucial data to identify enzyme intermediates.

**Force production by motor proteins**

Motor proteins such as myosin, dynein, and kinesin, produce a force for movement by coupling the hydrolysis of ATP to conformational changes in protein structure that are linked to the cyclical interaction of the motor with a filament (actin or microtubule). Initial characterization of actin-stimulated myosin ATPase kinetics using steady-state methods led to the puzzling observation that actin fully stimulated ATP turnover while associating with only small fraction of the myosin (12). In a set of simple and direct experiments using transient kinetics, Lymn and Taylor (13) showed that actin associates with myosin for a very brief time and then rapidly dissociates as part of the myosin crossbridge cycle. ATP binding to the actomyosin complex stimulates rapid release of the myosin from the actin, whereas the binding of actin to the myosin-ADP-P complex stimulates the release of products (ADP and Pi). Conformational changes that occur with nucleotide binding and release are then coupled to force production. A motor protein, which is called kinesin, walks along microtubules in a hand-over-hand fashion because of an alternating-site ATPase pathway (14, 15). Strain in the linkage between the two motor domains of this dimeric protein distinguishes leading from lagging ATPase sites when they are both bound to the microtubule surface. This strain leads to rapid dissociation of ADP from the leading motor domain while stimulating ATP hydrolysis on the lagging motor domain. Moreover, ATP binding to the lagging head stimulates the release of ADP from the leading head. Transient kinetic analysis established this alternating site ATPase cycle by showing directly that ATP binding to one site stimulated the rate of ADP release from the other site (14).

**Kinetic and Structural Basis for Catalysis**

Structural analysis is often the basis for discussions of enzyme specificity and mechanism. However, structural studies alone cannot establish mechanism or define the origins of enzyme specificity. In the three examples cited above, structural studies were critical to illustrate the active site residuum, but could not by themselves address the most pressing questions. In particular, enzyme specificity is a kinetic phenomenon which is quantified by $k_{cat}/K_m$ but the underlying origins of specificity are not revealed by pondering the magnitude of $k_{cat}/K_m$ or by inspecting the three-dimensional arrangements of amino acids. Rather, in combination with transient-state kinetic studies, the structures and steady-state kinetic parameters take on new meaning and can be interpreted to reveal fascinating details of enzyme dynamics and specificity.

**The chemistry of DNA polymerization**

The structure of the active site of a DNA polymerase with bound DNA and an incoming substrate, dCTP, is shown in Fig. 1a (16). In this structure, the DNA is terminated with a deoxyribonucleotide that lacks the 3'-OH so that catalysis is prevented, and so this structure is thought to represent a close approximation to the closed enzyme state immediately preceding catalysis. The chemistry of the reaction is well established in that the reaction is catalyzed by a two metal ion mechanisms, in which one Mg$^{2+}$ increases the nucleophilic reactivity of the 3'-OH while the other Mg$^{2+}$ stabilizes the developing negative charge on the $\beta$-phosphate during the transition state (17). However, this model does not explain the extraordinary efficiency and specificity of the reaction. This polymerase copies DNA at a rate of 300 base pairs per second and makes a mistake approximately one out of a million times. After making a mistake, the polymerase stalls, which gives time for the DNA primer strand to melt away from the template strand and migrate to an exonuclease site 25 Å away to remove the mismatched base. The primer then rapidly renatures and polymerization continues so that the overall error frequency is only one in a billion (15). Efficiency and specificity are kinetic properties of the enzyme that can be addressed directly by transient kinetic analysis to define the kinetically significant intermediate and thereby establish the kinetic and thermodynamic basis for nucleotide selectivity. Detailed kinetic studies have revealed that although the metal ions mechanism accounts for a base level of catalysis, the alignment of positively charged amino acids on the O-helix (Fig. 1a) plays a critical role in specificity by their involvement in recognition of a correct base pair.

**The chemistry of EPSP synthase**

EPSP synthase catalyzes the synthesis of EPSP by an addition-elimination reaction through the tetrahedral intermediate shown in Fig. 2a. This enzyme is on the shikimate pathway for synthesis of aromatic amino acids and is the target for the important herbicide, glyphosate, which is the active ingredient in Roundup (The Scotts Company LLC, Marysville, OH). Transient-state kinetic studies led to proof of this reaction mechanism by the observation and isolation of the tetrahedral intermediate. Moreover, quantification of the rates of formation and decay of the tetrahedral intermediate established that it was truly an intermediate species on the pathway between the substrates (S3P and PEP) and products (EPSP and Pi) of the reaction. The chemistry of this reaction is interesting in that the enzyme must first catalyze the formation of the intermediate and then catalyze its breakdown, apparently with different requirements for catalysis. Quantification of the rates of each step of this reaction in the forward and reverse directions has afforded a complete description of the free-energy profile for the reaction and allows
has been the topic of research for more than a half century. Studies in the light microscope established a sliding filament model for muscle contraction, which was subsequently shown to be caused by the cyclical interaction of the myosin ATP-Pase with actin filaments. A major revelation came with the realization that the largest changes in free energy are associated with the binding of ATP and the release of ADP and Pi, whereas the actual chemical step involving hydrolysis had an equilibrium constant of only 4 at the active site of the enzyme (13). The very tight binding affinity \( K_a \sim 10^{-11} \text{M} \) of ATP to myosin is used in the cycle to induce a change in enzyme structure that breaks the otherwise tight interaction of myosin with actin in the so-called rigor complex. After dissociation of the myosin-ATP complex from the actin, the hydrolysis of ATP converts the myosin to a myosin-ADP-Pi state that can reassociate weakly with actin. Subsequent conformational changes coupled to the sequential release of Pi and then ADP drive the swinging of a lever arm during force production (18, 19). The key to understanding coupling is to recognize that the myosin motor cycles between tight- and weak-nucleotide-binding states in reactions that are opposed and therefore coupled to cycling between tight- and weak-filament-binding states. The details of this energy transduction cycle are derived from a combination of structural, spectroscopic, equilibrium, and kinetic measurements; transient-state kinetic analysis has played a central role in defining the pathway.

The coupling of ATP hydrolysis to force production by kinesin is different, but it relies on similar principles in that the binding of substrate and release of products are coupled to conformational switching between tight- and weak-filament-binding states (14, 15, 20). Like myosin, cycling between two nucleotide-binding states is coupled to cyclical changes in the affinity of the kinesin motor domain for binding to the microtubule surface. However, the pathway is complicated by the negative cooperativity between the two ATPase domains of the dimeric motor. It is believed that strain induced when simultaneously binding both motor domains to the microtubule surface causes the leading head into a weak-nucleotide-binding state stimulating the release of product ADP. In contrast, the trailing head binds ATP tightly, and the tight association with the microtubule surface also stimulates ATP hydrolysis. The subsequent release of Pi is then coupled to the dissociation of this trailing head from the microtubule that allows it to leapfrog forward, thereby being converted from the trailing to a leading motor. As this motor domain then binds to the microtubule surface in its new role as leading motor, it reaches a weak-nucleotide-binding state and rapidly releases the bound ADP. Unlike myosin in which one step of the cycle can be attributable to the force-producing step, net displacement of kinesin along the microtubule occurs with a complete cycle of the ATPase shown in Fig. 3a, which is rate limited by the release of Pi from the trailing head. This complex pathway was derived by a relatively small set of simple and direct experiments using transient kinetic methods to measure the rate of each reaction and to provide definitive evidence for this alternating site ATPase pathway. For example, Fig. 3b shows the effect of increasing ATP concentrations in stimulating the release of ADP from kinesin bound to a microtubule.

**Figure 1** Mechanism of DNA polymerization. (a) The structure of T7 DNA polymerase in a complex with DNA and an incoming nucleotide is shown with a fluorescent label attached to C314. Changes in the fluorescence allow quantification of the nucleotide-induced change in structure and its role in selectivity. Residues 235-411 and 436-454 have been removed to reveal the active site. Shown also are the 52-Nbs and key catalytic residues from PDB 1T7P (17). (b) The time dependence of the fluorescence change induced by nucleotide binding is shown at three concentrations of dCTP. The inset shows the measurement of the rate of dCTP dissociation from the E.DNAdd. d N T Pc o m p l e x. A n a l y s i s o f t h e s e d a t a d e f i n e d t h e r o l e o f t h e n z y m e estmodation of the extent to which the enzyme catalyzes each step (8).

**The chemistry of energy transduction by motor proteins**

The free energy available from the hydrolysis of ATP to form ADP and Pi in solution is generally in the neighborhood of 15 kcal/mol, depending of course, on the intracellular concentrations of ATP, ADP, and Pi. The means by which an enzyme can couple this free energy to produce a force for movement...
Transient State Enzyme Kinetics

Transient-Kinetic Tools and Techniques

Transient-kinetic techniques most often rely on the rapid mixing of reactants with enzyme to initiate the reaction. This mixing is essential so that all enzyme molecules start reaction in synchrony with one another; therefore, the time dependence of the observable reactions defines the kinetics of interconversion of enzyme intermediate states. Because mixing requires a finite amount of time, conventional methods are limited in their ability to measure very fast reactions. For example, a typical value for the "dead time" of a stopped-flow instrument is approximately 1 ms, which is because of the time it takes to fill the observation cell. Thus, reactions with a half-life of less than 1 ms (rate > 700 s⁻¹) are difficult to observe depending on the signal to noise ratio. With a favorable signal to noise ratio, rates on the order of 2000 s⁻¹ can be measured, but it is rare. Other transient kinetic techniques have been devised to break this time barrier. Both temperature-jump and pressure-jump experiments can approach the 1-µs limit, but both only allow observation of relaxation of a system of reactants from one equilibrium position to another and therefore have less widespread use. Modern microfluidic methods have also achieved mixing times in the microsecond range and offer the promise of more widespread use. In addition, recent efforts have succeeded in combining rapid mixing followed by mass spectrometry analysis to detect enzyme intermediates.

Stopped-flow methods

The two prominent transient-kinetic mixing methods are stopped flow and rapid chemical quench flow. In the stopped flow, the

Figure 2 Intermediate in the EPSP synthase pathway. (a) The mechanism of the reaction catalyzed by EPSP synthase is shown. The reaction proceeds by an addition-elimination mechanism via a stable tetrahedral intermediate. (b) A single turnover reaction is shown in which 100-µM enzyme was mixed with 100-µM S3P and 3.5-µM radiolabeled PEP. Analysis by rapid-quench kinetic methods showed the reaction of PEP to form the intermediate, which then decayed to form EPSP in a single turnover. The smooth lines were computed from a complete model by numerical integration of the equations based on a global fit to all available data. Reproduced with permission from Reference 7.
instrument mixes and then forces reactants through a flow cell. After an abrupt stop of the flow, the observation cell contains reactants that were just mixed within the previous millisecond. One can then trigger a computer to collect data to measure changes in optical signals such as absorbance, fluorescence, or light scattering as a function of time after mixing. Other spectroscopic methods such as electron paramagnetic resonance have been employed, but many traces must be collected and averaged to improve the signal-to-noise ratio. Stopped-flow methods have a distinct advantage in that one can often obtain data of high quality (high signal-to-noise ratio) over a series of concentrations of substrate with a modest investment of time and materials. These data can then accurately define the concentration dependence of observable rates of reaction; this information is critical in developing a model for the enzyme pathway. However, stopped-flow signals are sometimes difficult to interpret uniquely, especially when fluorescence data scheme multiple exponentials, as described below. An important solution to the problem can be achieved by correlating stopped-flow data with measurements of the chemical conversion of substrate to product using rapid quench-flow methods.

Figure 1b illustrates the analysis of substrate binding and dissociation using stopped-flow fluorescence methods. The concentration dependence of the binding rate defines a two-step sequence that involves a weak rapid-equilibrium binding followed by isomerization to tighter binding. The inset to Fig. 1b shows an experiment to measure the substrate dissociation rate. Combined, the two experiments accurately define the kinetic parameters that govern the two-step binding sequence.

Rapid quench-flow methods

Rapid chemical-quench-flow methods are employed to measure the chemical conversion of substrate to product without necessarily relying on an optical signal. Often, radiolabeled substrates are used, and the substrate and product are resolved chromatographically to allow quantification of the conversion of substrate to product. The reaction is started by mixing the enzyme with substrate and then stopped by mixing with a quenching agent, such as 2 N acid or base. In a quench-flow apparatus, reactants are driven through loops of tubing at defined speeds to control the timing and to achieve reaction times as short as 2 ms. Longer reaction times, up to approximately 100 ms, are achieved by using longer loops of tubing. Even longer reaction times, greater than 100 ms, require a push-pause-push mode in which the first push forces the reactants into a reaction loop and after a pause of known duration, a second push forces the reactants to mix with a quenching agent and then out into a collection tube. The method can be time consuming and tedious compared with stopped-flow, but rapid quench-flow methods have two distinct advantages. First, they provide a direct measurement for the rate of conversion of substrate to product at the active site; second, they allow one to assess the extent to which dead ends are present. This can be interpreted more easily in evaluating a given model. In addition, they allow one to assess the extent to which dead enzyme or microheterogeneity in the enzyme and/or substrate might influence interpretation of the data.

Figure 2b shows the results of a rapid-quench single-turnover experiment performed with EPSP synthase with enzyme in excess over the radiolabeled substrate, PEP. The data show the transient formation and decay of the tetrahedral intermediate, which led to its subsequent isolation and structure determination.

When transient kinetic methods fail

Conventional transient-kinetic methods are not always applicable to every enzyme system. In particular, enzymes that have
very fast turnover rates such as carbonic anhydrase are not amenable to single-turnover kinetic methods and may require the use of stopped-flow methods just to observe reactions in the steady state. In addition, enzymes with very high $K_w$ values are sometimes difficult to study because the high substrate concentrations needed to saturate the rate of reaction increase the background in attempts to observe product formed at a concentration equal to only one per enzyme site. However, these limitations are being challenged constantly by the development of better instrumentation, new methods, and better signals. If one is lucky or wise enough to choose an enzyme with $K_w$ values in the micromolar range and $k_{cat}$ less than several hundred per second, then a large range of experiments is possible.

Experimental Design and Interpretation of Data

The overall goal of a comprehensive kinetic study is to provide estimates for each step in the pathway and to identify enzyme-bound intermediates. The time dependence of an enzyme-catalyzed reaction can be rather complex for even a simple reaction sequence that contains only one substrate, one product, and one enzyme-bound intermediate such as the reaction illustrated in Scheme 1. Nonetheless, careful design and rigorous interpretation of a combination of equilibrium and kinetic data can define all eight rate constants under favorable circumstances; in particular, definition of the reverse rate constants usually requires that the reaction be measurable in the reverse direction. In those cases in which rates cannot be defined with certainty, the degree of uncertainty must be stated. For example, in some cases the data may only indicate that $k_2$ and $k_4$, for example, must be greater than $1000 \text{s}^{-1}$, but no data define an upper limit. In either case, the data could show that the formation of the intermediate is the kinetically significant step and that the intermediate breaks down at a fast rate. Given the very nature of the reaction, this method may be the best possible and perhaps all that is important mechanistically.

Equilibrium measurements

Steady-state and equilibrium measurements are extremely important to place constraints on the interpretation of more complex kinetic experiments. For example, measurement of the overall equilibrium constant for the reaction defines the product of equilibrium constants for all four steps, $K_{eq} = K_1K_2K_3K_4$. In addition, it is often possible to measure the two internal equilibria, $K_2$ and $K_3$, if concentrations of enzyme can be attained that are higher than the dissociation constants for substrate and product. Finally, dissociation constants for substrates and products can be measured in some instances, particularly with multiple substrate reactions when one of the reactants can be left out to allow the binding but not reaction of the other substrate. Whatever information can be obtained will facilitate the design of subsequent experiments (i.e., what concentrations of substrate are needed to saturate the enzyme) and in their interpretation (i.e., knowing that $k_{cat}/K_w = K_2$ provides an important constraint in globally fitting data).

Binding kinetics

It is useful to approach the solution of an enzyme pathway from both ends and work inward. Measurement of the substrate-binding kinetics is a good first start, but it will depend on a useful signal. Often, changes in protein fluorescence occur that reflect changes in protein structure induced by substrate binding; these signals can be monitored to measure the kinetics of binding and isomerization. In other cases, a fluorescent analog of a substrate or a fluorescently labeled protein can provide a useful signal. In either case, definition of rate constants that govern binding assist in the design and interpretation of subsequent experiments to measure rates of substrate to product conversion at the active site. Figure 1b illustrates the measurement of substrate binding kinetics.

Pre-steady-state burst kinetics

Important results are obtained in an experiment where enzyme is mixed with an excess of substrate and then the formation of product is monitored over the time period of a few enzyme turnovers and with an enzyme concentration high enough to allow quantification of one product per enzyme site. Under conditions in which substrate binding and chemical conversion to product are faster than the release of product, one observes a “burst” of product formation equal to one product per enzyme site, followed by steady-state turnover. The mere observation of a burst implies that a step after chemistry is at least partially rate limiting for steady-state turnover. In addition, if no observable burst occurs, then the data imply that chemistry is largely rate limiting. When a burst can be observed, quantitative fitting of the amplitude and rate of the burst phase relative to the steady-state phase affords estimates for three rate constants: $k_2$, $k_{-2}$ and $k_4$ in the pathway shown above (Scheme 2).

Single turnover kinetics

If enzyme can be obtained at concentrations approaching the $K_w$ or $K_v$ for substrate, then experiments can be performed with enzyme in excess of substrate. In these experiments, one can observe the net conversion of substrate to intermediate and form product in a single enzyme cycle. The real advantage of the method is in the signal-to-noise (or background) ratio of the data. While in a pre-steady state burst experiment, one might be looking for the appearance of an intermediate in the background of 100-fold higher concentration of substrate; in a single turnover experiment, that same intermediate might appear as 30% of the total substrate. This finding was the case in studies on EPSP synthase, in which single-turnover experiments revealed the presence of the tetrahedral intermediate, which led to its isolation and structure determination (Fig. 2b).
Interpretation of transient kinetic data

Unlike the standard protocols for steady-state kinetic analysis, transient kinetic analysis is dependent on the availability of signals to measure individual steps of the reaction. Moreover, the observable kinetics can change and be complex or deceivingly simple depending on the relative magnitudes of sequential steps in a pathway. The rule of thumb is that one exponential phase exists in the time dependence of a reaction for each step in the pathway. For example, the time dependence of signals observable according to Scheme 2 will follow a triple exponential function: $Y = A_1 e^{-\lambda_1 t} + A_2 e^{-\lambda_2 t} + A_3 e^{-\lambda_3 t} + c$. Moreover, it is important to note that the observed rates ($i$ values) do not translate directly to rate constants for individual steps except in the rare case where each step is irreversible. Rather, the observable rates are a function of all rate constants that are reversibly linked (1, 21, 22). In general, the observed rate approaches the sum of the rate constants for all steps that contribute to the observable transient. A complete discussion of this is beyond the scope of this review, but some simplifying concepts are important to note. For example, if the binding of substrate is a rapid equilibrium or the experiment is conducted at sufficiently high substrate concentration, then the exponential term that governs the formation of ES is so fast that it is dropped from the equation. In addition, in a pre-steady-state burst experiment, the last exponential term is lost because the release of products renders the formation of ES so fast that it is dropped from the equation. For example, in a pre-steady-state burst experiment, the last exponential term is lost because the release of products generates free enzyme, which blunts the time dependence of the reaction into the steady-state phase. Thus, it is common that for this three-step pathway, the observed rate of formation of product during a pre-steady-state burst experiment follows a single exponential with a rate constant provided by the sum, $k_{observed} = k_2 + k_2 + k_3$ (followed by linear steady-state turnover).

Under some conditions, the kinetics of the reactions simplify, but the validity of such simplifying assumptions is limited. Moreover, as a reaction becomes more complex, the mathematical modeling of the data becomes intractable. For these reasons, methods have been developed for modeling and fitting kinetic data by computer simulation based on numerical integration of the full rate equations without simplification. The data shown in Fig. 3b, for example, can only be analyzed by simulation to obtain meaningful information because of complexities in the experiment. However, in fitting the data by simulation, the rate constant for ATP binding is defined accurately by the observed increase in rate of the slow reaction phase as a function of ATP concentration.

Computer simulation and global data fitting

Kinetic data can be fitted directly to a given model without invoking simplifying assumptions needed to solve mathematical equations. Rather, by numerical integration of the rate equations, complex models can be examined in fitting data over a wide range of concentrations and differing starting conditions. When fitting data by computer simulation, one bypasses mathematical modeling entirely and fits data directly to the model. Moreover, many experiments can be fitted simultaneously to a single unifying model. Several computer programs are currently available including Dynafit (BIoK, Ltd., Watertown, MA), Copasi (EML Research, Heidelberg, Germany), and KinTek Global Kinetic Explorer (KinTek Corporation, Austin, TX), and all are available on the internet. The major difficulty in using these programs occurs because of the ease with which one can propose overly complex models in attempting to explain kinetic data. Understanding which kinetic constants are actually constrained by the data becomes extremely difficult when fitting multiple parameters to several data sets. KinTek Explorer addresses this problem by allowing the user to scroll through each rate constant, starting concentration, or output factor and watch the curves change in shape. This dynamic simulation immediately reveals whether a given constant has a significant effect on the fit to the data and shows when a model is overly complex.

The future of kinetic analysis lies in the use of these powerful computer simulation methods to design experiments and to fit data rigorously to extract mechanistic information. Ongoing work is directed toward the development of more robust algorithms to address the confidence intervals with which individual parameters can be obtained from modeling a given set of data. With these tools, one can accurately establish a given reaction mechanism and provide estimates for the rates and free-energy changes for each kinetically significant step in the pathway.

References


Further Reading


See Also

Enzyme Kinetics
Enzyme Catalysis: Roles of Structural Dynamics in Techniques to Study Enzyme Kinetics
DNA Replication: An Overview
Electron Transfer, Chemical Roles of Water in Biology

Ilya A. Balabin and David N. Beratan, Duke University, Durham, North Carolina
Spiros S. Skourtis University of Cyprus, Nicosia, Cyprus

doi: 10.1002/9780470048672.wecb148

Water plays multiple roles in biological electron transfer (ET): energy bath, polarizable medium that defines the reaction coordinate, electronic coupling bridge, and intimate participant in molecular recognition. This article explores these many faces of water in ET. Links are drawn to reactions in photosynthesis, oxidative phosphorylation, proton-coupled ET, and DNA damage and repair.

Biological Background: The Many Roles of Water in Biology

Biological electron transfer (ET) reactions are ubiquitous in nature. They define the fundamental reactions of bioenergetics, biocatalysis, and nucleic acid synthesis and repair. These reactions occur both in membranes and in aqueous environments. In both settings, water plays multiple roles in the ET reaction kinetics and dynamics (1–5). The roles of water are diverse. Its roles vary from establishing a high dielectric screening medium to supporting the solvation of counterions to establishing hydrogen-bonded relays to providing electron tunneling pathways and proton wires for key transport reactions. Our aim is to review the influence of water on biological ET and, as such, examine some of the many roles of water in these reactions.

Chemical Roles of Water in Biomolecular ET

The last 50 years have witnessed the establishment of a truly molecular-level description of electron transfer chemistry. From the Marcus description of how solvent polarization defines the ET reaction coordinate, to fully quantum treatments that describe electron and nuclear tunneling contributions to the kinetics, to atomistic simulations of reaction coordinate motion, a comprehensive view of biological ET is emerging (1–5).

Inner-and outer-sphere ET

Electron transfer reactions are categorized as outer or inner sphere. In outer-sphere processes, structural changes during ET largely involve solvent reorientation (polarization). Inner-sphere reactions involve changes in the bonding to the redox unit and are more difficult to describe using the ET theory as it is explained here.
Outer-sphere biological redox reactions occur between cofactors beyond van der Waals contact distances, and they occur mostly via electron tunneling. The ET rate is proportional to the product of an electron tunneling probability and an Arrhenius-like activation factor (1–5). The electron tunneling probability is proportional to the square of the donor-acceptor electronic coupling, and the Arrhenius-like factor is given by $\exp[-\Delta G^*/k_B T]$, where $\Delta G^*$ is the activation free energy, $k_B$ is Boltzmann’s constant, and $T$ is the temperature. The activation factor arises from changes in nuclear polarization coupled to the electron transfer event.

The activation free energy $\Delta G^*$ is defined by two terms. One term is the reaction free energy $[\Delta G(0)]$, which is derived from the redox potential difference of the donor and acceptor. The second term is the medium reorganization energy ($\lambda$), which is the energy stored in the solvent internal degrees of freedom when the electron is shifted suddenly from donor to acceptor (1–5). Marcus theory predicts the activation free energy:

$$\Delta G^* = [\Delta G(0) + \lambda^2/4\lambda]$$

(1)

Because $\Delta G^*$ enters the rate expression in the argument of an exponential, changes in driving forces and reorganization energies sweep ET rates over many orders of magnitude. In photoinduced electron-transfer reactions (as in the photosynthetic reaction center and photosystem II (2, 7)), the thermodynamic driving force for charge separation and charge recombination reactions are very different. Indeed, this difference is believed to enhance the efficiency of photosynthetic charge separation (1).

The reorganization energy $\lambda$ in Equation 1 has contributions from “inner-sphere” and from “outer-sphere” (i.e., solvent) motion. The outer-sphere reorganization energy ($\lambda_{\text{out}}$) often dominates the activation free energy. The activation free energy for ET depends on the match between $-\Delta G(0)$ and $\lambda$. That is, the reaction rate is maximized (for a fixed donor-acceptor distance) when $-\Delta G(0) \approx \lambda$. In the regime where $-\Delta G(0) < \lambda$, the “normal” regime, the reaction rate increases with increasing thermodynamic driving force. When $-\Delta G(0) > \lambda$, the reaction is “inverted” and its rate slows with additional increases in driving force. A great achievement of the 1980s was to access both of these regimes (1, 3), which validated the most dramatic prediction of the Marcus theory. Polar solvents like water have larger $\lambda_{\text{out}}$ than nonpolar solvents ($\lambda_{\text{out}}$ for water is typically a few electron volts and for nonpolar solvents is typically tens of electron volts). Therefore, the inverted regime of ET in aqueous solvent is accessed at larger driving forces than ET in low-polarity solvents, and aqueous ET reactions are more likely to be in the “normal” regime. As such, docking or association processes that exclude water would decrease $\lambda_{\text{out}}$ and accelerate the kinetics. This simple observation explains a possible advantage for carrying out the charge separation reactions of biomembranes of nonpolar transmembrane proteins rather than in aqueous media. Membranes also play a second and important role by defining a simple directorality for electron and proton transport.

Inner-sphere ET reactions are of great interest in chemistry and biology (8). However, the kinetics of these reactions are rather complex as they involve bond breaking and formation with the redox group. For inner-sphere processes, electron transfer need not be rate limiting and the reaction coordinate may not be determined as described above for pure ET events. ET kinetics may also be gated by conformational charges or ion motion (9, 10).

When inner-sphere motion is coupled to ET but does not control the reaction mechanism, it adds to the outer-sphere reorganization energy to define a total $\lambda$ value for Equation 1. Changes in bond lengths and vibrational frequencies contribute to the reorganization energy and are known as inner-sphere components. Inner-sphere reorganization energies routinely are computed with quantum chemical methods and are typically on the scale of tenths of electron volts; outer-sphere reorganization energies can be larger than this size and frequently are computed using continuum dielectric methods (20).

In addition to bulk water, which contributes to $\lambda_{\text{out}}$, buried water molecules can make significant contributions to the overall reorganization energy of biological ET reactions. The magnitude of the buried-water contribution depends on the location of the waters relative to the D and A moieties and on their number. Finite-difference Poisson-Boltzmann calculations for the R. viridis photosynthetic reaction center, which contains more than 100 buried waters, show that the buried water contribution to the total reorganization energy varies from 0.05 to 0.27 eV for different PRC ET reactions (Reference 11 and references therein). The largest contribution (0.27 eV) is for the cytochrome to the special pair ET reaction and constitutes 25% of the total computed reorganization energy (total meaning bulk and buried water, protein, and membrane). The effect of water on $\Delta G^*$ strongly depends on the ET system considered, for example, on how deeply the D and A moieties are buried inside the protein (which screens the bulk solvent) and on the positions of internal waters relative to the D and A molecules.

### Water-mediated molecular interactions

Water has $C_2$ symmetry. In the gas phase, the measured O–H bonds are 0.957 Å, and the H–O–H angle is 104.5° (12). Liquid water and ice have structures controlled by the formation of hydrogen bonds. These bonds make it possible for hydrogen ions to exchange among water molecules on the millisecond to picosecond time scale (13), depending on pH. The extensive and dynamic hydrogen bond networks account for many unusual properties of water and hydrated biomolecules (12).

Water-mediated molecular interactions, including hydrogen bonding, hydrophobic effects, ion pairing, and cation–π interactions, play a central role in biomolecular structure and function (5). In soluble proteins, most nonpolar residues that cannot form hydrogen bonds with water are located in the protein core, whereas polar and charged residues form most of the protein surface, which increases the protein-solvent stability (14). Cation–π interactions, which involve electrostatic attraction between a cation and the negative electrostatic potential associated with the face of a π system, influence both the protein structure and the ion distribution in solvents (5). These factors play a central role in protein-protein and protein-ligand binding, particularly influencing the structure of the thin water layers between biomolecules in an encounter complex and thereby influencing intermolecular ET reactions (15).
Water and nonequilibrium solvation

Water also influences the nature of the motion along the ET reaction coordinate (donor, acceptor, and solvent degrees of freedom that are coupled to ET). A measure of the short timescale nonequilibrium solvation dynamics is the Franck-Condon time \( \tau_{FC} \). This time is the time it takes for the ET system to move away from the crossing point of the electron donor and acceptor potential surfaces where the activation complex is formed. For temperature \( T \) and reorganization energy \( \lambda \), \( \tau_{FC} \) is \( \tau_{FC} \approx \hbar/\sigma \Delta U \), where \( \sigma \Delta U = \sqrt{\lambda T} \) is the root mean square fluctuation in the donor to acceptor energy gap (16) and references therein). For biological ET, \( \tau_{FC} \) was computed for azurin ET (Ru\(^2+\)–Ru\(^3+\)) (17) using a semiclassical molecular dynamics (MD) methodology (18). The computed \( \tau_{FC} \) was 2.47 fs with significant protein and solvent (water) contributions. For protein alone, it was \( \tau_{FC} \approx 4.3 \) fs, and from solvent alone, it was \( \tau_{FC} \approx 3.0 \) fs. Water solvation dynamics are equally important for excited-state ET, in which the formation of the D excited state by photoexcitation may induce a large dipole moment change of the D moiety that is followed by rapid protein matrix and solvent response. Excited-state ET from (FADH\(^2\)) to DNA in DNA photolyase is an example of an ET reaction in which the change in dipole moment of the D moiety induced by photoexcitation [(FADH\(^2\)) \( \rightarrow \) DNA] has shown very fast and significant solvation dynamics induced by photoexcitation with relaxation timescales ranging from 2 to 60 picoseconds (20). This solvation dynamics, which is much faster than the ET rate, is partially attributed to water molecules in the DNA photolyase active site (20). Solvent relaxation can limit rates that are fast compared with the solvent relaxation timescale (21). This water is characterized by a broad distribution of relaxation timescales (subpicosecond and slower) and may therefore act as a rate-limiting solvent for ET reactions (21).

Tunneling mediation

The tunneling characteristics of biological ET have placed considerable recent focus on protein and solvent (3, 22). The energetic proximity of solvent and protein orbitals to the redox active states of the donor and acceptor cofactors provides supereexchange pathways for tunneling (23). That is, the energy cost of tunneling via these protein states is much lower than the cost of tunneling through “pure” empty space. As such, the intervening medium facilitates weak tunneling propagation of the electron from donor to acceptor via coupling pathways (24). The tunneling probability through this medium multiplies the Marcus factor described above. Early models for protein-electron tunneling predicted that rates would drop by about a factor of 10 for a 1 Å decrease in the 1980s and 1990s, approaches were introduced to explore the influence of the specific medium structure on the tunneling. In recent years, the influence of medium dynamics was added to complete the models (14, 25–29, and references therein).

The role of disordered water as a tunneling medium has been the subject of several electronic-structure computational studies. Early calculations focused on the Fe(H\(_2\)O)\(_{6}\)\(^{3+}/2+\) self exchange reaction (30–32). These calculations concluded that disordered water is not a very effective electron-tunneling medium. More recently, one-electron pseudopotential analysis of electron tunneling through water layers placed between electrodes indicated that water is a modest tunneling mediator (33). A further computational study of aqueous ET between ion pairs used MD simulations with INDO/S and CAS-SCF calculations of the electronic coupling (34). This study showed that water can be an essential tunneling medium. The computed INDO/S tunneling matrix element decay parameter was 2.0 Å\(^{-1}\) and the anisotropy values ranged from 1.5 to 1.8 Å\(^{-1}\). Experiments in aqueous glasses give average decay of 1.8 Å\(^{-1}\) (22).

Of particular interest is how oriented water in confined spaces contributes to tunneling. Several experiments indicate a special role for bridging waters (22, 35–37). Recent studies of interprotein ET reveal details of how the ensemble of docked structures contributes to the observed rate. The “dynamic docking” paradigm for ET between weakly associated proteins suggests that minority population configurations dominate the electron tunneling mediation (38). It seems likely that water can contribute to dynamic docking by influencing donor-acceptor coupling pathways or by modifying the reorganization energies as the proteins approach a docked structure. Newer developed experimental methods enable the study of electron transfer kinetics between co-crystallized proteins. These experiments are likely to provide direct answers to some of the questions of how specific waters at protein-protein interfaces may influence ET kinetics (22, 39).

The fleeting nature of water structures in proteins makes it extremely challenging to study their influence on ET by direct means. Indications that a few key bridging waters seem to accelerate protein ET reactions motivated a theoretical simulation of the presence of explicit bridging water molecules (35–37). Pathway studies indicate that waters at cytochrome c\(_o\) and the photosynthetic reaction center serve as key mediators (Fig. 1) (40) Quantum studies of water mediation in protein-protein systems indicate that spatially confined waters between proteins may enhance electronic coupling in a regime just beyond protein-protein contact (Fig. 2) (41). In this regime, the distance permits entry of a few intervening waters and the coupling is enhanced by constructive interference among multiple coupling pathways. At larger distances, the effect rapidly dissipates. These studies are motivating small-molecule studies of water-mediated electron transfer in systems with well-defined cavities that may bind water molecules (D. Waldeck, private communication).

Proton-coupled ET reactions

In bioenergetics, electron transfer drives proton transfer, proton gradients in turn power ATP synthesis. In addition to coupling to medium polarization, biological ET reactions may also couple strongly to proton transfer. As such, rates may be limited by proton transfer or electron transfer or may fall into an intermediate regime. Rapid advances are being made in both the theoretical and experimental investigation of these coupled rate processes (42, 43, and references therein).
Electron Transfer, Chemical Roles of Water in

Figure 1  Water-mediated ET pathways between cytochrome c2 and the photosynthetic reaction center in an encounter complex (left panel) and the more fully docked state (right panel) (40). The pathways are shown for 50 conformations obtained by MD sampling. Cofactor-mediated pathway fragments are shown in green, protein-mediated fragments are shown in purple, and water-mediated fragments are shown in orange.

Figure 2  Dependence of the mean square electronic coupling on distance between two porphyrin rings in the cytochrome b5 self-exchange ET reaction (41). For each distance, system conformations were sampled using MD and the coupling was computed for each conformation at the extended Hückel level. The black line marked XH(P, W) shows the water-mediated coupling; for comparison, the red line marked XH(P) shows the coupling computed for the same protein conformation in vacuum. Conformational snapshots typical for the three coupling regimes are shown.

Isotope effects on ET reactions

Substitution of H2O solvent by D2O generally leads to a change in ET kinetics (44, 45, and references therein). In the simplest model of ET coupled to a single solvent vibration involving H, substitution of H by D lowers the frequency of the vibration and increases the localization of the vibrational wave functions. These changes influence the Franck–Condon factor for the ET reaction in a way that depends on the temperature and on reaction energetics. For example, at low temperatures, when the vibrations are in their ground states, the increased localization (for deuteration) of the vibrational wave functions is expected to speed up an activationless reaction by enhancing the probability of being at the crossing point between D and A potential energy surfaces. In contrast, for an activated ET reaction at low temperature, the increased localization will reduce the probability of reaching the crossing point by nuclear tunneling, which, thus, slows down the ET reaction.

Protein ET systems have large and somewhat flexible macromolecule assemblies with folded structure and motion strongly coupled to the surrounding solvent. In these systems, isotopic substitution of the solvent may affect multiple factors that influence the ET rate, and, generally, the overall isotope effect on the ET rate is system-dependent. For example, D2O changes protein solvation because the solvent–solvent hydrogen bonds are stronger in D2O compared with H2O, which leads, therefore, to an increased hydrophobic effect. Changes in the fold and volume of a protein may alter the donor–acceptor distance, the redox potentials, and the tunneling-pathway networks, which all influence the ET rate.

The strengthened hydrogen bond networks in D2O, which lower molecular and solvent flexibility, produce larger reorganization energies and affect nuclear fluctuation timescales, which are especially important in gated and solvent-controlled ET. For proton-coupled ET reactions, D2O solvent substitution changes the proton transfer–electron transfer equilibria and kinetics. The literature on the isotope effect on ET is large and growing (44, 45, and references therein).

ET reactions controlled by conformational transitions and solvent viscosity

Because of the rapid decay of the electronic coupling with distance, very long-distance (>40 Å) ET reactions rarely occur in a single step. Instead, extremely long range ET involves an array of multiple redox centers, mobile electron carriers, or large-scale motion of redox-active domains (46). All intermolecular ET reactions require either protein–protein docking or formation of an encounter complex in which the two protein cofactors are

\[
\begin{align*}
\text{direct contact} & \quad \text{structured water} & \quad \text{bulk water} \\
10^{-14} & \quad 10^{-12} & \quad 10^{-10} & \quad 10^{-8} & \quad 10^{-6} & \quad 10^{-4} & \quad 10^{-2} & \quad 10^0 & \quad 10^2 \\
\text{XH(P)} & \quad \text{XH(P, W)} & \quad \text{XH(P)} & \quad \text{XH(P, W)} & \quad \text{XH(P)} & \quad \text{XH(P, W)} & \quad \text{XH(P)} & \quad \text{XH(P, W)} & \quad \text{XH(P)} & \quad \text{XH(P, W)}
\end{align*}
\]
within a few angstroms (40, 41). In these cases, protein diffusion in solvent, influenced by solvent viscosity, pH, and other factors, can be rate limiting if the formation of an encounter complex is slower than the ET reaction itself.

In unimolecular ET, the rate can be controlled by large-scale cofactor motion, such as the quinone motion in the photosynthetic reaction centers, the Rieske subunit motion in the cytochrome bc₃ complex (47), or the cytochrome bc₁ domain in sulfite oxidase. Theoretical models for conformationally controlled ET reactions have been suggested by Hoffman and Ratner (48) and Brunschwig and Stuiv (49). Large-scale protein or domain motions are themselves linked to the movement of water molecules.

**Experimental and Theoretical Methods**

**Experimental methods**

Ultrafast ET reaction kinetics usually is measured using transient absorption or luminescence spectroscopy. ET reactions are initiated by exciting the native or modified ET cofactor or by injecting an electron or hole by a flash-quench technique (22). Slower ET reactions can be studied by using a variety of methods, including calorimetry, stopped flow, dipole relaxation, and other techniques (4). Recent experiments on the structural relaxation of water provide intriguing observations of its structure and dynamics. For example, the coexistence of ordered surface water and crystallite-like ice structures following exposure to a femtosecond-range infrared laser pulse was observed with ultrafast electron diffraction methods (50). Studies of the first hydration shell of a water molecule in liquid water using X-ray absorption spectroscopy and X-ray Raman scattering reveal that water molecules typically form only two hydrogen bonds (one donor and one acceptor bond), contrary to predictions of MD simulations that suggest three or four hydrogen bonds (51). Theoretical calculations find the optimal combination of bonded and non-bonded links from the donor to the acceptor cofactors using an X-ray-determined protein structure (24). Intriguingly, with essentially one parameter (the ratio of through bond to through space coupling decay), the Pathways model qualitatively described the nature of ET rates in several redox proteins, facilitating an early phase of mapping structural effects on protein ET reactions (24).

**Theoretical methods**

Theoretical methods for estimating the reorganization energy have to account for its large variations in different environments. The Marcus model for ϵ_loxp assumes spherical donors and acceptor moieties embedded in a dielectric continuum that represents the solvent. ϵ_loxp is proportional to (1/ϵ_loxp - 1/ϵ_in), where ϵ_loxp is the optical dielectric constant and ϵ_in is the static dielectric constant of the medium. The dependence of ϵ_loxp on ϵ_loxp explains why ϵ_loxp is higher in water and highly polar solvents compared with nonpolar solvents. In low dielectric media like membranes, ϵ_loxp and ϵ_in are nearly matched, thus lowering the reorganization energy. The simple continuum model cannot be used to compute accurate reorganization energies in complex molecular environments. More sophisticated numerical approaches have been developed for the computation of ϵ (e.g., References 11, 56, 57, and references therein). Numerical approaches permit the computation of reorganization energies for cofactors of arbitrary shape (e.g., Reference 24). They can be combined with MD simulations to compute the reorganization energy for different MD conformations of the ET system to assess the influence of fluctuations. Different methods have also been developed for computing λ (e.g., References 11, 56, and references therein).

A variety of theoretical descriptions of hydrogen-bonding water networks has been proposed, including flickering clusters, percolation networks, fluctuating charges, random networks, continuum models, and so forth. (58), yet many structural aspects of liquid water remain open to debate (59, 60). Since the first explicit (atomic) model for water was proposed by Bernal and Fowler in 1933, a variety of water potential functions have been developed (61, 62). These water potential functions typically include noncovalent interactions, for example, Coulombic interactions among partial charges assigned to each atom (and to each electron lone pair in some models) and van der Waals interactions among atoms. Some force fields include elastic terms, namely O–H bond stretch and H–O–H angle deformations. Two force fields common in biomolecular simulations are the SPC model and the TIP3P model (61, 62). Both force fields are rigid, three-site descriptions that were parameterized to reproduce the bulk liquid water phase structure and thermodynamics. However, none of these (and similar) force fields accurately predicts the temperature dependence of liquid water’s density (60). Later force fields incorporated intramolecular flexibility (e.g., the TIP4P model (62)).
Electron Transfer. Chemical Roles of Water in

Other improvements include adding fourth and fifth sites, polarizability, and quantum effects (62). Although models like TIP4P and TIP5P provide a more accurate description of homogeneous bulk water, they perform poorly in describing gas-phase clusters and nonpolar solutes in polar solvents (62). In biomolecular simulations, these models cannot correctly account for situations when a nonpolarizable moiety is exposed to different electrostatic environments, either within a static structure or during simulations. Another problem with nonpolarizable models is that of including the average bulk polarization in the potential. Whereas polarizable force fields, such as the fluctuating charge and related models, attempt to address these problems with varying success, SPC and TIP3P remain the most appropriate force fields for simulations of large biomolecular systems (61). Sampling of the biomolecular conformations is usually performed using MD simulations or Monte Carlo methods (61, 62). The protonation state of titratable amino acids can be treated with constant pH dynamics, QM/MM calculations, or continuum electrostatics methods (61, 62). Formation of a protein–protein encounter complex is often studied using Brownian dynamics (63). Studies of protein–protein docking involve electrostatic potential analysis, and more recently, protein flexibility models, for example, normal mode analysis (64).

Frontiers in Biomolecular ET Reactions

Protein-mediated ET reactions

The photosynthetic cytochrome b6f complex is an integral part of the oxygenic photosynthetic electron–proton transport chain. It transfers electrons from photosystem II to photosystem I and moves two protons per electron transferred across the photosynthetic membrane to build a transmembrane proton gradient. The solution of the crystal structure of the cytochrome b6f complex (65) revealed the relative positions of its redox-active subunits, cytochrome f, cytochrome b, and the Rieske 2Fe–2S protein, as well as the structural similarities with the respiratory cytochrome bc1 complex. In the cytochrome b6f complex, a lipid-soluble quinol transfers an electron to the Fe2S2 center of the Rieske protein, an electron to the heme b6 of cytochrome b6 and two protons to the aqueous side of the thylakoid membrane. The Rieske 2Fe–2S protein subsequently reduces the heme c center of cytochrome f that extends into the aqueous side of the membrane. Cytochrome f reduces the Cu center of water-soluble plastocyanin that in turn transfers the electron to photosystem I. Cytochrome f contains a buried chain of five water molecules inside its heme-binding large domain. The water chains are hydrogen bonded to surrounding residues that are highly conserved in cytochrome f sequences (66). It was shown that mutations that affect this water chain impair electron transfer and photosynthetic function (67). The mutations also affected the redox potentials of cytochrome f (Fig. 3). It is possible that the water chain interacts with the Rieske 2Fe–2S protein when the latter approaches the hydrophilic edge of cytochrome f to transfer an electron to the heme (67). Water chains (as long as 10–12 water molecules) and smaller water clusters are also present in bacterial photosynthetic reaction centers. They connect the secondary quinone to the cytoplasmic surface and are thought to participate in proton transfer from the cytoplasm to the secondary quinone (6, and references therein; Fig. 4). Water molecules seem to play a more complex role in cytochrome c oxidase, the terminal component of the mitochondrial and to other bacterial respiratory chains. The protein performs the four-electron reduction of O2 to water and couples it to transmembrane proton transport. The mechanism of function for this enzyme is not fully understood and is the subject of ongoing studies (6, and references therein; Fig. 1).

Apart from the role of the water molecules as proton wires, it has also been suggested (although not demonstrated experimentally) that internal waters are involved in redox-linked switching of the proton wires. Electron transfer from the heme a of cytochrome c oxidase to the binuclear Cu–heme a3 site may invert the local electric field between the redox sites and thus cause restructuring of internal waters and the switching of the proton pathways (6, and references therein).
DNA-mediated ET reactions

DNA is a polyelectrolyte with a tightly bound spine of water. Both DNA damage and repair mechanisms involve ET (10). Water is an important determinant of DNA structural stability, both in the form of organized water that occupies the major and minor grooves of the DNA double strand and as bulk solvent (10).

The mechanism of DNA-mediated ET remains poorly understood. Theoretical and experimental studies are consistent with a mechanism of superexchange (tunneling) mediation at short distances and multistep hopping at longer distances. Intensive investigations are exploring the degree of delocalization of “holes” in the DNA stack and the polaronic nature of the transport (10). Water may influence DNA charge transport in a variety of ways. Bulk and structured water around DNA and specific water molecules in the minor groove (70) and ordered water chains surrounding and internal water molecules (71) mediate hole transfer by acting as a superexchange bridge (Fig. 4). DNA decamer duplexes determined the hydrogen and deuterium atomic positions of the water molecules that hydrate the DNA (69). These experiments show that the water network in the DNA minor groove can form interstrand and intrastrand bridges that connect different parts of the duplex by hydrogen bonds. Density functional electronic structure calculations using the hydrated duplex structure identified in these experiments showed that the structure has three nearly degenerate highest occupied molecular orbitals, each spatially distributed over the G bases, and specific water molecules in the minor groove (70). These results suggest that hydrating water molecules in DNA not only influence the electronic structure of the polaron hole (10) but also mediate hole transfer by acting as a superexchange bridge (Fig. 4).

Conclusions

A unified view of how water influences biological ET is yet to emerge. Many qualitative effects are apparent: Water produces large reorganization energies, and structured water may produce favorable electron and proton transport pathways. Water at protein–protein interfaces can perturb docking energetics and influence both the coupling mechanisms and ET reaction coordinate. Structured water chains provoke intriguing questions regarding electron and ion communication, coupling, and switching. In nucleic acids, water and DNA counter ions influence both the structure of “holes” and the kinetics of their transport.

References

Electron Transfer, Chemical Roles of Water in


Electron Transfer, Chemical Roles of Water in


Further Reading


See Also

Chemistry of Electron Transfer Chain
Chemical Basis of Electron Transport
Electron Transfer Chemistry in Photosynthesis
Bioelectronics
Hydration Forces
Brian A. Todd, Nina Y. Sidorova and Donald C. Rau, NICHD, NIH, Bethesda, Maryland
Christopher Stanley, NIST, Gaithersburg, Maryland
doi: 10.1002/9780470048672.wecb230

Force measurements on a wide variety of biologic macromolecules indicate that the perturbation of water structuring as hydrated surfaces closely approach can dominate interaction energetics. Forces with remarkably similar characteristics are observed for the repulsive interactions of charged polymer surfaces, of uncharged polymer surfaces, of charged solutes with uncharged surfaces, and of uncharged solutes with charged surfaces stressing the universality of the hydration force in aqueous solution. Repulsive interactions are caused by a disruption of the hydrogen bonding network of the intervening water between surfaces. The available measurements of forces between macromolecules that spontaneously assemble are consistent with an attractive hydration interaction. A water-order parameter theory has been developed that is successful for predicting force curve behaviors. The direct involvement of water structuring in the interaction between macromolecules has important consequences for recognition, assembly, and conformational transitions that underlie biologic function.

In the simplest representation, hydration forces reflect the energetic consequences of perturbing the water structure as two hydrated surfaces approach. The concept developed as a result of direct measurements of interactions between macromolecules or surfaces in water that were unlike the forces commonly considered important (1-14). As is amply described elsewhere in previous articles, polar or charged surfaces bind and structure water strongly. Hydration contributes substantially to structure and stability. Water is a highly associative liquid because of its strong hydrogen-bonding ability; the structuring of water in direct contact with the macromolecular surface will alter the structure of subsequent water layers in solution. As two hydrated surfaces approach, the intervening water must accommodate both surfaces. Depending on the nature of the opposing surfaces, the hydrogen-bonding network of water can be disrupted, resulting in repulsion, or reinforced, causing attraction. Ample evidence from infrared or Raman spectroscopy, neutron scattering, and dielectric relaxation shows that the structuring of water in confined spaces, as in zeolites, inverse micelles, or very small droplets, can be different from bulk water (15-22) and references therein). In practice, hydration forces are observed between macromolecules only over the last 1–1.5-nm separation between surfaces, which corresponds to about the two last water layers on each surface. Although the energy perturbation per water molecule can be relatively small, the force can be large when summed over all the water molecules that must be reorganized. An order parameter theory has been developed to formalize the treatment of hydration forces and has been surprisingly insightful.

Force Measurements
We present the force measurements that led to the hypothesis of hydration forces, focusing on the osmotic stress measurements of several very different biologically important assemblies: lipid bilayers, DNA, collagen, and several different polysaccharides. The measurement of repulsive hydration forces between macromolecules at close approach has now been supplemented by similar measurements for the exclusion of small solutes from macromolecular surfaces. Atractive hydration interactions are more difficult to probe, but the available evidence indicates that the polycation-mediated assembly of DNA and the temperature-favored assemblies of collagen and hydrophobically modified cellulose are examples of attractive hydration forces. Lastly, the magnitude of the observed hydration forces means that polar surfaces will avidly cling to the hydration waters in the absence of a stringently complementary surface that can replace the favorable water interactions. This has profound consequences for recognition and assembly reactions in the cell that are typically characterized by both binding strength and specificity. Water structuring forces or energetics have
be sensitive to salt, solutes, pH, temperature, and so on. Along the macromolecule, the partial derivatives in the above expression represent the forces acting to push \( N(CH_3)_4^+ \)-DNA helices to touching, for example, in a DNA helix-DNA helix interaction. The same kind of force is observed to dominate the interactions of small solutes with macromolecular surfaces. The osmotic stress approach has also been used to map the distribution function for the exclusion of small nonpolar solutes from DNA and of polar solutes from hydrophobically modified cellulose. The partial derivatives in the above expression remind us that forces between macromolecules can be sensitive to salt, solutes, pH, temperature, and so on.

Hydration Repulsion

Figure 1 shows osmotic stress force curves, \( \log(\Pi_{\text{osm}}) \) versus the surface separation between macromolecules, for several different macromolecular assemblies that are charged, zwitterionic, or completely uncharged. Unlike hard materials such as mica or clays that show oscillating forces as surfaces approach \( 3–4 \AA \) exponentially, biologic macromolecules have forces that decay as \( 3.5–4 \AA \) exponentials with a 3.5–4-Å decay length \( \lambda \).

Order Parameter Theory for Hydration Forces

A phenomenologic theory has been developed to account for hydration forces (42–45). Water organized by surfaces is characterized by an order parameter. The linearized theory is mathematically equivalent to the Debye-Hückel formulism for electrostatic double-layer interactions. Homogenous surfaces that contain salt and other small solutes that equilibrate between the backbone phosphate neutralized by \( N(CH_3)_4^+ \), for egg phosphatidyl choline that has a \( CH_3 \) \( H_2 \) \( CH_2 \) group covalently linked to the phosphate of the lipid head group and didodecyphosphosphate in \( N(CH_2)_4^+ \) is shown scaled to the same phosphate surface area. The surface hydration is dominated by the phosphate group with an additional weak contribution from the tetraalkylamine counterion. The macromolecular surfaces with the same hydrated chemical groups show similar force amplitudes. The common unifying theme is that all the macromolecular systems are hydrated. Within the hydration force framework, the 3–4-Å exponential decay length \( \lambda \) reflects a water-water correlation length. Such a correlation has been observed by X-ray scattering for density fluctuations in pure water (38). The force amplitude or pre-exponential factor is caused by the structure of water in direct contact with apposing surfaces. In all cases exclusion is characterized by a 3.5–4-Å decay length associated with the pre-exponential factor for the exclusion of simple nonpolar alkyl alcohols from DNA and of polar solutes from hydrophobically modified cellulose. The pre-exponential factor for the exclusion of simple nonpolar alkyl alcohols from DNA varies linearly with the difference between the number of alkyl carbons and hydroxyl oxygens (43) in the solute, which further emphasizes the connection between molecular hydration and force magnitude.
Figure 1 Osmotic stress force curves are shown for several biological assemblies: schizophyllan (a triple helical polysaccharide composed only of glucose) in water (\(\bullet\)), \(\iota\)-carrageenan (a double helical charged polysaccharide) in 0.4-M NaCl (\(\square\)), hydroxypropyl cellulose in water at 5°C (\(\circ\)), DNA in 0.4-M NaCl (\(\triangle\)), DNA in 0.4-M \(\text{N(CH}_3\text{)}_4\text{Cl}\) (\(\triangleleft\)), DNA in 0.25-M \(\text{N(CH}_3\text{)}_4\text{Cl}\) scaled to the same phosphate surface density as DDP and egg PC. Those biomacromolecular assemblies measured in water are net neutral, either zwitterionic (egg PC) or wholly uncharged (schizophyllan and HPC). Forces between charged surfaces are insensitive to salt concentration over the last 10–15 Å, but they do depend significantly on the counterion identity. The X-ray spacings have been adjusted to give approximate surface separations from either gravimetric measurement of spacing as dependent on water content (the lipid bilayers), fiber dimensions determined by X-ray scattering (DNA, schizophyllan, and \(\iota\)-carrageenan), or the spacing between polymers when dried (HPC). Even though force amplitudes span a range of \(\sim\)1000 in pressure, the decay lengths, \(\lambda\) of the apparent exponential are similar for all these systems; the fits to the schizophyllan and DDP forces give 3.4 and 3.5 Å, respectively. The overlap of the force data for egg PC, \(\text{N(CH}_3\text{)}_4\text{Cl}\)–DDP, and \(\text{N(CH}_3\text{)}_4\text{Cl}\)–DNA scaled to the same phosphate surface area indicates that force amplitudes are determined by the hydration of groups on the surface. Data for DDP and egg PC are from Leikin et al. [2].

Organize water symmetrically, as illustrated in Fig. 3a, are repulsive since the hydrogen-bonded network between surfaces will become increasingly disrupted as surfaces move closer.

The order parameter theory gives for the osmotic pressure between two planar surfaces with a separation D and water–water correlation length \(\lambda_w\),

\[
\Pi^{\text{homo}}_\text{water} = \frac{C_r}{\sinh^2(D/2\lambda_w)} \sim C_r e^{-D/\lambda_w}, \text{ for } D \gg \lambda_w
\]

Similarly, two homogeneous surfaces with a complementary water structuring, as illustrated in Fig. 3b, will be attractive, as the hydrogen-bonding structure of the intervening water is strengthened,

\[
\Pi^{\text{homo}}_\text{water} = -\frac{C_a}{\cosh^2(D/2\lambda_w)} \sim -C_a e^{-D/\lambda_w}, \text{ for } D \gg \lambda_w
\]

The coefficients \(C_a\) and \(C_r\) are measures of the strength of water ordering on the surface.

Of course, biological surfaces are not homogeneous. Real surfaces typically can have both opposing attractive and repulsive patches. In such a case, the osmotic pressure can be represented by (2, 9, 13, 46, 47),

\[
\Pi = \frac{C_a}{\cosh^2(D/2\lambda_w)} + \frac{C_r}{\sinh^2(D/2\lambda_w)} \sim -(C_a - C_r)e^{-D/\lambda_w} + (C_a + C_r)e^{-2D/\lambda_w}, \text{ for } D \gg \lambda_w
\]

In the case, \(C_a \gg C_r\), repulsions dominate and the purely repulsive exponential form \(\Pi \sim -C_a e^{-D/\lambda_w}\) is recovered. If the attractive and repulsive patches are balanced, then the second-order term of the approximation dominates with an apparent decay length that is only half that of \(\lambda_w\), or \(\sim 2\) Å. This second-order term is always repulsive and reflects the loss of favorable water–water interactions that are caused by the presence of the other surface. Its magnitude is only dependent on the overall surface water structuring whether attractive or repulsive.

Often decay lengths do not divide so neatly into two classes of value. Three complications can cause significant variability in observed decay lengths. Surface periodicities or the separation length scale of hydrated groups on the surface must necessarily contribute to decay lengths (2). This effect will decrease the
Hydration Forces

Figure 2. The distance dependence characterizing exclusion of small solutes from macromolecular surfaces follows the same exponential behavior as the hydration force between macromolecules. The extent of exclusion can be extracted from the dependence of forces on solute concentration. \( \Pi_{\text{excess}} \) is the effective osmotic pressure applied by the solute in the bulk solution on the macromolecular phase, and \( \Pi_{1} \) is the maximal pressure from complete exclusion. \( \Pi_{\text{excess}}/\Pi_{1} = 0 \) then corresponds to complete exclusion and \( \Pi_{\text{excess}}/\Pi_{1} = 1 \) then indicates full inclusion. The distance dependent exclusion of the polar polyol adonitol (a) and glycerol (g) from hydrophobically modified hydroxypropyl cellulose (HPC) and of the nonpolar alcohols i-propanol and methyl pentanediol (MPD) from spermidine3 alkyltriamine spermidine is shown. As in Fig. 1, interaxial spacings are converted to surface separations. The apparent exponential decay length varies between 3.5 and 4 Å (solid lines indicate fits to the data).

Hydration Attraction

If \( C_{s} \gg C_{1} \), in Equation 4, then attractive forces can result. Forces measurements on DNA double helices spontaneously assembled by polycations provide evidence for attractive hydration forces (12, 13). The binding of several polycations result in spontaneous condensation of DNA. Packaging of genetic material in vivo and compaction of DNA for use in nonviral gene therapy exploit this attraction. The equilibrium spacing between helix surfaces in the condensed state varies between 6-12 Å depending on the nature of the condensing counterion indicative of both attractive and repulsive forces acting on DNA as implicit in Equation 4. Figure 4 contrasts the osmotic stress curves for DNA in the +3 alkyltriamine spermidine that results in spontaneous precipitation of DNA, in the +2 alkylamine putrescine, and in the +1 alkylamine tetramethylammonium, \( \text{N(CH}_{3}\text{)}_{4}^{+} \). Neither putrescine2- nor \( \text{N(CH}_{3}\text{)}_{4}^{+} \) condenses DNA. The \( \text{N(CH}_{3}\text{)}_{4}^{+} \)-DNA curve shows a 4-Å single exponential characteristic of purely repulsive hydration forces dominated by the first-order term of Equation 4. The putrescine2- -DNA force curve also shows a single exponential force at high osmotic pressures but with a \( \sim 2 \)-Å decay length indicative of a balance between attractive and repulsive patches that leaves the second-order term of Equation 4 dominant. The force curve in the presence of spermidine3- is different. At \( \Pi_{1} = 0 \), there is a balance between the attractive first-order term and the repulsive second-order term that results in a finite equilibrium spacing (arrow). This spacing leaves \( \sim 9.5 \) Å of water-filled space between DNA molecules. The curve eventually merges with the single exponential, \( \sim 2 \)-Å decay length observed with putrescine2- as the interhelical spacing decreases, which aligns with the prediction of the order-parameter formalism for the emergence of a second-order repulsive term in the presence of hydration attractions. The merging of spermidine and putrescine forces at high osmotic pressures is expected for DNA surfaces with chemically homologous bound cations. The second-order repulsion is simply determined by the overall water structuring on a single helix, not on the specific mixture of attractive and repulsive patches. Determining the decay length for attractions is more difficult because there are currently no experimental techniques to directly measure attractive forces between molecules. Combining single-molecule force measurements of net attraction with the osmotic stress...
The temperature-favored assembly of hydrophobically modified cellulose, hydroxypropyl cellulose (HPC), has also been probed (31). At temperatures below ~40°C, exponential repulsive forces are observed with an approximate 3-4 Å decay length. At very close spacings (~2 Å surface separation), the force amplitude increases rapidly, which is likely caused by steric interactions between hydroxypropyl groups on the surface. The force amplitude of the 3-4 Å decay length exponential decreases linearly with temperature extrapolating to 0 at ~40°C very close to the temperature at which the polymer precipitates from dilute solution, ~42°C. Above 40°C, the equilibrium spacing between HPC polymers in the absence of applied osmotic pressure continuously decreases with increasing temperature. Since the very short-ranged steric force does not change observably with temperature, this change in equilibrium spacing must result from an attractive force with an amplitude that increases with temperature. Force data above 40°C can be well fit by the sum of the rapidly changing short-ranged force and an attractive 4 Å decay length exponential. Remarkably the temperature dependence of the fitted attractive force amplitude is a linear continuation of the repulsive force amplitudes directly observed at lower temperatures. Since hydrophobic forces have long been considered to be caused by the release of water structured around nonpolar surfaces, a connection with hydration forces should not be surprising. The strengthening of attraction or weakening of repulsion as the temperature increases is caused by a difference in the temperature dependence of water structuring around polar hydroxyl groups and nonpolar methyl groups. A confounding aspect, however, is that the 2 Å second-order repulsion observed with condensed DNA is completely absent with HPC. This hydrophobic polymer simply acts as a homogeneous surface that smoothly transforms from repulsive to attractive as the temperature is increased.

Hydration Linked to Recognition Reactions and Protein Transitions

The important lesson from the direct force measurements for biology is that polar or charged surfaces will avidly retain their hydration waters in preference to other surfaces if those surfaces are not closely complementary. A key difference, therefore, between specific and nonspecific binding of proteins, for example, should be in the water sequestered by the complexes (53, 54). Differences in water between specific and nonspecific recognition complexes can be readily determined from the dependence of the relative equilibrium constant or ligand binding constant on osmotic pressure or water activity. For osmotic pressures expressed as an osmolar concentration, a standard linkage relation is applied to water in a same way it is traditionally applied to salt, ligands, or pH is,

\[
\frac{d \ln (K_{eq})}{d [\Pi_{w}]} = -\frac{\Delta N_{w}}{55.56}
\]

where \(K_{eq}\) is the equilibrium constant between two complexes and \(\Delta N_{w}\) is the difference in bound water between them.
We have focused on differences in hydration among specific, noncognate, and nonspecific complexes of three sequence specific DNA binding proteins. DNA binding proteins will typically bind to nonspecific DNA or indeed to polyanions in general with binding constants in the range from $10^2$ to $10^7$ M$^{-1}$ and to their recognition sequence with constants typically in the range $10^3$ to $10^5$ M$^{-1}$.

The water release accompanying the transfer of the type II restriction nuclease BambhI from a nonspecific sequence to the recognition sequence was measured in order to validate the osmotic stress approach (55). Crystal structures of both the specific (56) and the nonspecific (57) complexes are known. The nonspecific structure shows a gap between the protein and DNA surfaces that can accommodate $\sim$150 waters. The two interfaces are in direct contact in the specific complex (Fig. 5). Relative nonspecific-specific binding constants can be determined from the loss of specific binding as the concentration of nonspecific competitor DNA is increased, using either a gel mobility shift or a self-cleavage assay (58). Relative binding constants vary significantly with the concentration of added osmolytes. The linear dependence of $\ln(K_{rel})$ on the osmotic pressure of the added solutes observed in Fig. 5 indicates that about 135 ($\pm$15) extra water molecules are retained in the nonspecific versus specific BamHI-DNA complex in good agreement with the structural data. The insensitivity of $\Delta N_w$ to the osmolyte used set water activity indicates that this water is likely sequestered at the protein-DNA interface of the nonspecific complex. This number of waters corresponds to about one-two hydration layers at the interface.

EcoRI is another type II restriction endonuclease that recognizes the DNA sequence GAATTC. As is typical for these restriction nucleases, sequence recognition is exquisitely stringent. The change of even a single base pair is sufficient to reduce the binding constant by a factor of $10^4$ to a value more characteristic of completely nonspecific binding (59). The osmotic pressure dependence of the relative specific-nonspecific binding constant indicates that the nonspecific EcoRI-DNA complex sequesters $\sim$110 more waters (60, 61). EcoRI complexes with DNA sequences that have even one wrong base pair ("star"-sites) also sequester $\sim$110 more waters than the specific complex. This water, however, can be removed from complexes with a single wrong base pair by increasing the osmotic pressure (62). The osmotic work required to remove almost all of the water from a complex of EcoRI with the sequence TAAACCT was estimated as 3–4 Kcal/mole complex. Water could not be removed from complexes with more than one wrong base pair even at the highest osmotic pressures examined. The loss of water from the "star" sequence complexes provides a natural link between enzymatic "star" activity of EcoRI and osmotic pressure.

The binding of the Cro repressor protein to DNA is more typical of specific-DNA-protein interactions. Unlike BamHI and EcoRI, changes in the DNA recognition sequence result in gradual decreases in the binding free energy of Cro protein. This bacteriophage $\alpha$ repressor recognizes two 6 base pair regions separated by a 5-bp spacer. An examination of a set of 5 binding sequences that span a range of $\sim$2200 in relative binding constants shows that for every 10-fold decrease in binding constant about eight to nine more water molecules are incorporated by the complex (63). Water and binding energy are strongly coupled. There are also marked differences in the extent of hydration between specific and nonspecific protein-protein interfaces. In a survey of crystal structures in the Protein Data Bank, the average number of crystallographically resolved waters in specific binding interfaces (for instance, between dimers) was compared with the number of waters in nonspecific "crystal contacts" (protein crystal contacts not normally involved in biologic interactions) (64, 65). With an average of 10 waters per 1000 Å$^2$, the biologically specific interfaces are 50% less hydrated than the crystal packing interfaces. These numbers represent waters sufficiently immobilized to be visible by X-ray and underestimate the actual amount of water at protein-protein interfaces. Although neither interface is dry, the displacement of interfacial waters at the protein-protein interface was a distinguishing characteristic of specific binding.

Allosteric conformational changes are a hallmark of protein structure and function. Small changes in the structure of a binding pocket can often trigger large conformational changes elsewhere in the protein. Hydration forces are reasonable candidates for this amplification of small structural perturbations since small rearrangements of the chemical groups on surfaces can result in large changes in the hydration force between them. Once again a coupling between water and conformation should be expected if hydration forces underlie the structural change. The T-R transition of hemoglobin that accompanies oxygenation is the best known and most extensively studied allosteric reaction. The binding of 4 O$_2$ molecules is linked to the binding of $\sim$50–60 extra water molecules by the protein as surfaces that were complementary are slightly perturbed and now prefer interactions with water (66).

Conclusions and Prospectives

Although it is almost universally acknowledged that water plays an important role in biologic assembly and recognition processes (67), defining its role quantitatively, however, has proven more difficult. It has long been known that the unfavorable interaction of water with nonpolar groups underlies the hydrophobic interaction. The observation of a common force at close separation among nonpolar, polar, and charged macromolecules that can be interpreted as resulting from water structuring greatly broadens the scope of hydration-driven interactions. The observed force is unlike that expected for double-layer electrostatics or van der Waals interactions that are more commonly considered important. Removing the last one or two layers of water between macromolecules is energetically costly if the apposing surfaces are not closely complementary. The force characteristics observed, particularly the observed decay lengths, seem remarkably simple and robust. The transmission of surface water structuring into the solution through correlated water-water interactions can account for the decay lengths. The order parameter theory provides a satisfactory initial framework for understanding and calculating hydration.

Hydration Forces

The structural data. The insensitivity of versus specific BamHI-DNA complex in good agreement with added solutes observed in nonspecific–specific binding constants can be determined from linear dependence of $\ln(K_{rel})$ on the osmotic pressure of the significantly with the concentration of added osmolytes. The or a self-cleavage assay (58). Relative binding constants vary

The water release accompanying the transfer of the type II restriction nuclease BambhI from a nonspecific sequence to the recognition sequence was measured in order to validate the osmotic stress approach (55). Crystal structures of both the specific (56) and the nonspecific (57) complexes are known. The nonspecific structure shows a gap between the protein and DNA surfaces that can accommodate $\sim$150 waters. The two interfaces are in direct contact in the specific complex (Fig. 5). Relative nonspecific-specific binding constants can be determined from the loss of specific binding as the concentration of nonspecific competitor DNA is increased, using either a gel mobility shift or a self-cleavage assay (58). Relative binding constants vary significantly with the concentration of added osmolytes. The linear dependence of $\ln(K_{rel})$ on the osmotic pressure of the added solutes observed in Fig. 5 indicates that about 135 ($\pm$15) extra water molecules are retained in the nonspecific versus specific BamHI-DNA complex in good agreement with the structural data. The insensitivity of $\Delta N_w$ to the osmolyte used set water activity indicates that this water is likely sequestered at the protein-DNA interface of the nonspecific complex. This number of waters corresponds to about one-two hydration layers at the interface.

EcoRI is another type II restriction endonuclease that recognizes the DNA sequence GAATTC. As is typical for these restriction nucleases, sequence recognition is exquisitely stringent. The change of even a single base pair is sufficient to reduce the binding constant by a factor of $10^4$ to a value more characteristic of completely nonspecific binding (59). The osmotic pressure dependence of the relative specific-nonspecific binding constant indicates that the nonspecific EcoRI-DNA complex sequesters $\sim$110 more waters (60, 61). EcoRI complexes with DNA sequences that have even one wrong base pair ("star"-sites) also sequester $\sim$110 more waters than the specific complex. This water, however, can be removed from complexes with a single wrong base pair by increasing the osmotic pressure (62). The osmotic work required to remove almost all of the water from a complex of EcoRI with the sequence TAAACCT was estimated as 3–4 Kcal/mole complex. Water could not be removed from complexes with more than one wrong base pair even at the highest osmotic pressures examined. The loss of water from the "star" sequence complexes provides a natural link between enzymatic "star" activity of EcoRI and osmotic pressure.

The binding of the Cro repressor protein to DNA is more typical of specific-DNA-protein interactions. Unlike BamHI and EcoRI, changes in the DNA recognition sequence result in gradual decreases in the binding free energy of Cro protein. This bacteriophage $\alpha$ repressor recognizes two 6 base pair regions separated by a 5-bp spacer. An examination of a set of 5 binding sequences that span a range of $\sim$2200 in relative binding constants shows that for every 10-fold decrease in binding constant about eight to nine more water molecules are incorporated by the complex (63). Water and binding energy are strongly coupled. There are also marked differences in the extent of hydration between specific and nonspecific protein-protein interfaces. In a survey of crystal structures in the Protein Data Bank, the average number of crystallographically resolved waters in specific binding interfaces (for instance, between dimers) was compared with the number of waters in nonspecific "crystal contacts" (protein crystal contacts not normally involved in biologic interactions) (64, 65). With an average of 10 waters per 1000 Å$^2$, the biologically specific interfaces are 50% less hydrated than the crystal packing interfaces. These numbers represent waters sufficiently immobilized to be visible by X-ray and underestimate the actual amount of water at protein-protein interfaces. Although neither interface is dry, the displacement of interfacial waters at the protein-protein interface was a distinguishing characteristic of specific binding.

Allosteric conformational changes are a hallmark of protein structure and function. Small changes in the structure of a binding pocket can often trigger large conformational changes elsewhere in the protein. Hydration forces are reasonable candidates for this amplification of small structural perturbations since small rearrangements of the chemical groups on surfaces can result in large changes in the hydration force between them. Once again a coupling between water and conformation should be expected if hydration forces underlie the structural change. The T-R transition of hemoglobin that accompanies oxygenation is the best known and most extensively studied allosteric reaction. The binding of 4 O$_2$ molecules is linked to the binding of $\sim$50–60 extra water molecules by the protein as surfaces that were complementary are slightly perturbed and now prefer interactions with water (66).

Conclusions and Prospectives

Although it is almost universally acknowledged that water plays an important role in biologic assembly and recognition processes (67), defining its role quantitatively, however, has proven more difficult. It has long been known that the unfavorable interaction of water with nonpolar groups underlies the hydrophobic interaction. The observation of a common force at close separation among nonpolar, polar, and charged macromolecules that can be interpreted as resulting from water structuring greatly broadens the scope of hydration-driven interactions. The observed force is unlike that expected for double-layer electrostatics or van der Waals interactions that are more commonly considered important. Removing the last one or two layers of water between macromolecules is energetically costly if the apposing surfaces are not closely complementary. The force characteristics observed, particularly the observed decay lengths, seem remarkably simple and robust. The transmission of surface water structuring into the solution through correlated water-water interactions can account for the decay lengths. The order parameter theory provides a satisfactory initial framework for understanding and calculating hydration.
Figure 5  The nonspecific DNA complex of the restriction nuclease BamHI retains significant water at the protein-DNA interface compared with the specific complex. Neutral osmolytes (triethylene glycol, α-methyl glucoside, and glycine betaine) are added to the solution to change water activity. Increasing the osmotic pressure of the bulk solution increases the free energy difference between specific and nonspecific binding. The linear dependence corresponds to 135 more water molecules sequestered by the nonspecific complex compared with the specific complex. The insensitivity to the nature of the osmolyte indicates the water is sterically sequestered. The inset shows the structures of the nonspecific (pdb identification-1esg) and specific (1bhm) complexes of BamHI looking down the DNA axis. The DNA is in tan, and the BamHI protein in blue. The nonspecific complex has a gap at the DNA-protein interface large enough to accommodate ∼150 water molecules that is not present in the specific complex. 

forces. The frustrating gap in our understanding is in connect-
sing surface water structuring with the pre-exponential factors, C0 and C1 of Equation 4. Making this link is necessary to de-
velop predictive tools for calculating hydration forces, but it will
require the combined contributions of theory, careful character-
ization of surface water structuring, and further measurement of intermolecular forces.

References

3. Lekin S, Rau DC, Parangian VA. Direct measurement of forces between self-assembled proteins: temperature-dependent exponen-
5. M. Mironov Tj. Short-range interactions between lipid bilayers mea-
6. M. Mironov Tj, Simon SA. Hydration and steric pressures be-
7. Petracek H, Tritam-Nagle S, Gaurisch K, Harries D, Parangian VA. Structure and fluctuations of charged phospholipid/serine bi-
9. Rand RP, Fuller NL, Parangian VA. Rau DC. Variation in hydra-
11. Rau DC, Parangian VA. Direct Measurement of forces be-
12. Rau DC, Parangian VA. Direct measurement of temperature-depen-
13. Rau DC, Parangian VA. Direct measurement of the intermolec-
Hydration Forces


52. Sidorova NV, Murayevs S, Rau DC. Differences in hydration coupled to specific and nonspecific competitive binding and to specific binding of the restriction nuclease BamHI. J. Biol. Chem. 2006;281:35656-35666.
Water is an integral part of DNA structure, and it plays an important role in dictating conformational preferences and binding properties of nucleic acids. Although in recent years a wealth of information on the physical properties of water in the vicinity of nucleic acids has accumulated, our understanding of how water interacts with nucleic acids and how it mediates DNA recognition by small ligands and DNA-binding proteins remains limited. In this article, we review the current state of DNA hydration research. In particular, we discuss structural, dynamic, and thermodynamic aspects of nucleic acid hydration. We also present an overview of the structural and thermodynamic role of water in modulating protein–DNA recognition and highlight the importance of hydration as a major contributor to the energetics of nucleic acid recognition. We emphasize the need for additional physico-chemical studies of DNA hydration by experimental and theoretic scientists.

**Liquid Water**

Water is the least understood liquid with respect to its macroscopic properties; yet, because of those same properties, it is also the only milieu known to support life (1). Some of the more important properties of water that make life possible are the negative change in volume associated with the ice-to-water transition, the density maximum at 4 °C, the anomalously high melting and boiling points, the high mobility of H3O+ and OH− ions, and its translucency to visible light. The ability of a water molecule to engage in hydrogen bonding is the key to understanding the macroscopic anomalies of liquid water. A water molecule can act simultaneously as both donor and acceptor of hydrogens, which leads to the formation of extensive hydrogen-bonded networks in liquid water. In crystalline ice, each water molecule is hydrogen bonded to four nearest neighbors that are arranged spatially with local tetrahedral symmetry. A comparison of the enthalpies of fusion and sublimation of ice reveals that liquid water at 0 °C is only ~15% less hydrogen bonded than ice. At 100 °C, liquid water still retains half of its hydrogen bonds.

The current view of liquid water is that of highly dynamic, random, three-dimensional networks of hydrogen-bonded water molecules with a local preference for tetrahedral geometry but with a large proportion of strained and broken hydrogen bonds (1). The abnormal properties of water are thought to result from the competition between bulky water structures networked by strong hydrogen bonds and more compact arrangements of water molecules with a prevalence of strained and broken hydrogen bonds (1). The distribution of water structures can be described by a pressure- and temperature-dependent partition function that takes into account the free energy differences between the various hydrogen-bonded arrangements. This distribution is altered locally under the influence of a solute; however, the change in distribution is paid for by respective changes in free energy, entropy, and enthalpy. These thermodynamic changes depend on the chemical nature of the solute (charged, polar, nonpolar) and, correspondingly, the nature of solute-solvent interactions.

**Hydration as a general phenomenon**

Water is not just a medium that dissolves a solute. Interactions between solute and water also dictate the conformational preferences of macromolecules, simultaneously guiding and actively participating in all aspects of molecular recognition. The modifying effect of hydration can have both structural and thermodynamic components. The structural effect is exerted by ordered waters that serve as localized hydrogen donors or acceptors for polar and charged groups of a macromolecule or their complexes and may be conceptualized as a structural extension of the macromolecule. The thermodynamic contribution of hydration is caused by solute-induced changes in the energetics of all affected waters. In this context, structural waters
Water of hydration is chemically identical to water in the vicinity of a solute. From a thermodynamic standpoint, hydration can be viewed as binding of water molecules to the hydration sites of a solute. The energetics of this association is modulated by the type of solute-solvent interactions (electrostatic, hydrogen bonding, van der Waals) and by solute-induced solvent reorganization. The latter occurs even in the absence of appreciable solute-solvent interactions because the equilibrium distribution of hydrogen-bonded water networks of the bulk becomes disrupted at the solute surface.

Water represents an integral part of DNA structure with the energetics of hydration being an important determinant of the conformational preferences of nucleic acids and the affinity and specificity of DNA interactions with other molecules. In the sections that follow, we discuss the current state of knowledge on the impact of hydration on nucleic acid structure and recognition.

Physical properties of DNA hydration

Water of hydration is chemically identical to water in the bulk. The differences between these two populations of water involve only their physical properties. Consequently, many physical methods have been employed to characterize water of DNA hydration.

Gravimetric measurements, in which the weight of water adsorbed by DNA films is determined as a function of relative humidity, have a long and venerable history (2, 3). Falk et al. (2) have concluded that, at 92% humidity, double-stranded DNAs are hydrated with about 20 water molecules per nucleotide. More recently, a comparative gravimetric investigation of the poly(dA)poly(dT) duplex and the poly(dT)poly(dA)poly(dT) triplex revealed that, at 98% humidity, the duplex and the triplex are hydrated by 21 ± 1 and 17 ± 2 waters per nucleotide, respectively (3). These findings are in quantitative agreement with the results of infrared spectral studies of DNA films performed as a function of relative humidity (3-5). Results of a detailed infrared study suggest that the initial hydration of the sugar-phosphate backbone and heterocyclic bases (at 64% relative humidity) includes about six water molecules (5). At this stage, the less-hydrated A-conformation DNA begins to form. An increase in humidity leads to additional binding of four to five more water molecules. At this level of hydration, formation of A-DNA is complete. Finally, above 95% humidity, the number of water molecules in the hydration shell of DNA becomes equal to ~20 per nucleotide, which results in the final stabilization of B-DNA.

One manifestation of strong solute-solvent interactions is the inability of affected waters to freeze when the temperature falls well below the freezing point. Nuclear magnetic resonance (NMR), infrared spectroscopy, and low-temperature calorimetry have been employed to characterize the number of nonfreezing waters in the hydration shell of DNA (6-9). Based on their infrared measurements of DNA films, Falk et al. (6) have concluded that about 10 water molecules per nucleotide are incapable of freezing with an additional 3 waters that show a tendency to supercool at a high cooling rate. Subsequent low-temperature NMR measurements in dilute DNA solutions have revealed a significantly greater number of nonfreezing water molecules that ranges from 27 to 29 molecules per nucleotide (7). The observed discrepancy may reflect partially a higher sensitivity of NMR over the early infrared spectroscopic techniques for characterizing water states.

Systematic differential scanning calorimetric measurements have revealed the sequence and conformation dependences of the number of nonfreezing waters in DNA hydration shells (8, 9). These studies have suggested that duplex DNA contains 11. CsCl density gradient ultracentrifugation of genomic DNA have concluded that, at 92% humidity, double-stranded DNAs have revealed a significantly greater number of nonfreezing waters per nucleotide than the single strand (10 vs. ~7 waters), whereas AT-rich domains of DNA are more hydrated than GC-rich domains (14 vs. 8 nonfreezing waters per nucleotide). These calorimetric results seem to be in qualitative agreement with the results of ultracentrifugation and osmotic stress studies (10, 11). Results of osmotic stress measurements have suggested that the triplex-to-duplex plus a single-strand transition of poly(dT)poly(dA)(poly(dT)) is accompanied by releasing one water molecule per triplet, whereas the duplex-to-single strand transition of E. coli DNA and poly(dA)poly(dT) causes the release of four water molecules (10). It has been noted, however, that at reduced humidity, GC base pairs exhibit a greater predisposition to adopt the A-conformation than do AT base pairs (12, 13). Consequently, the results of these low water activity studies may predominantly reflect reduced hydration of A-relatives to B-DNA rather than the differential hydration of GC and AT base pairs.

Volumetric investigations have provided a wealth of information on the sequence-, composition-, and structure-dependent hydration of nucleic acids and their low-molecular-weight analogs (14). The partial molar volume, V′, and adiabatic compressibility, K′, have proven to be sensitive to solute hydration (14). Generally, when comparing solutes that are structurally and compositionally similar, lower values of V′ and K′ (less positive or more negative) correlate with stronger and/or more extensive hydration of the solute. The partial molar volumes, V′, and adiabatic compressibilities, K′, of lithium, sodium, potassium, rubidium, cesium, ammonium, and tetramethylammonium salts of various polymeric nucleic acids have been determined at 25 °C with rigorous consideration of the Donnan membrane equilibrium effect (15). These measurements have revealed that, for any salt, the poly(dIdC)poly(dIdC) duplex is the most hydrated B-form DNA duplex followed by the poly(dGdC)poly(dGdC) and poly(dAdT)poly(dAdT) DNA duplexes. Furthermore, it has been observed that B-DNA generally exhibits lower values of V′ and K′ than A-RNA, with the latter including poly(rA), poly(rU), poly(rG), poly(rC), and poly(rU)poly(rC). Finally, single-stranded poly(U) is more hydrated than double-stranded poly(A)poly(rU) (normalized per base), an observation consistent with an increase in hydration accompanying the helix-to-coil transition. This notion is in agreement with a decrease in compressibility accompanying helix-to-coil transitions of polymeric DNA and RNA.
duplexes (16), although it contradicts the results of osmotic stress and low temperature calorimetric investigations (9, 10). The discrepancy may reflect differential probing of hydration by different techniques, which emphasizes again that hydration is defined operationally and it depends on the experimental observable.

In summary, the picture emerging from these studies suggests that DNA is an extensively hydrated macromolecule; the very structure of DNA is dictated by its interactions with water. The aggregate results suggest that 10 to 38 waters per phosphate interact with DNA and that these waters can be distinguished from bulk water by various physical observables. DNA hydration, as characterized by physical methods, has been shown to be sequence-, composition-, and conformation-dependent. However, different physical parameters are sensitive to different subpopulations of waters of hydration. As such, different parameters may be complementary but not directly comparable with each parameter providing its own unique window into a particular aspect of DNA-solvent interactions.

**Structural Aspects of DNA Hydration**

Information about ordered waters around nucleic acids comes predominantly from single-crystal X-ray studies of DNA and RNA oligonucleotides, although fiber diffraction data also have provided valuable insights. X-ray crystallography enables one to detect a subpopulation of the more localized water molecules predominantly from the first hydration shell of a biopolymer. Less localized waters or waters from more distant hydration shells are bound less tightly and generally do not contribute to measurable electron densities. Analysis of X-ray diffraction patterns in terms of hydration requires a combination of high quality data and reliable methods of refinement. Waters of hydration are in dynamic equilibrium with bulk water and generally exhibit low site occupancies and high temperature factors. Consequently, X-ray crystallography detects not solvent molecules themselves but loci that are occupied frequently and that are referred to commonly as hydration sites. The presence of cations in the vicinity of DNA may cause electron densities to be mistaken easily for those of water molecules.

With these notes of caution, X-ray crystallographic studies have provided important insights into the patterns of hydration of the A-, B-, and Z- conformations of DNA. In a seminal work, based on 15 B-DNA, 22 A-DNA, and 22 Z-DNA structures, Schneider et al. (17) have determined the average number of water molecules located in the first hydration shells of phosphates and bases of A-, B-, and Z-form DNA. It has been found that the sum of the waters in the hydration shells of phosphates and bases coincides with the net number of ordered waters in A- and B-DNA. This agreement is consistent with a picture in which the hydration shells of phosphate groups and bases in DNA do not overlap. By contrast, the sum of phosphate and base waters in Z-DNA (6.8) is larger than the total number of ordered waters (5.3), which suggests an overlap between the hydration shells of the backbone and the bases. The latter reflects the presence of continuous water networks that bridge phosphates and bases, which is a hydration signature of Z-DNA; Z-DNA is favored by low water activity conditions.

In agreement with conventional wisdom, in all three forms, water exhibits the highest affinity for the anionic O3P and O2P oxygens (see Fig. 1) (17, 18). The next highest affinity is observed for the hydrophilic exocyclic and endocyclic nitrogens of the bases (17, 18). Each of the two partially charged phosphate oxygens, O1P and O2P, has three hydration sites in the first layer, which form a so-called “cone of hydration.” By contrast, O3P and O5P phosphate ester oxygens and the ring oxygen O4′ are hydrated poorly. No ordered water molecules exist within 3 Å from O3P and O5P oxygens. In the right-handed A- and B-DNA, O5P is excluded sterically from water. All water contacts to O3P are longer than 3.1 Å, which suggests only weak hydrogen bonding of waters that are bound primarily to the partially charged oxygens O1P and O2P.

The O4′ oxygen of the furanose ring, which is often shielded by the attached base or by stacking on a neighboring base, can participate in hydration networks, although its involvement depends on its accessibility to water (18, 19). In A- and B-DNA, O4′ usually shares water molecules with the minor groove hydropathic base atoms from a previous residue. In the Z- form, only the guanine sugar O4′ atom is accessible for hydrogen bonding, because the O4′ atoms of cytosine sugars are oriented toward the guanine ring of the next residue.

In B-DNA, water molecules that solvate neighboring phosphate groups can be linked only through second-layer waters (17, 19). By contrast, in A- and Z-DNA, water molecules frequently bridge anionic phosphate oxygen atoms of different residues (17-19). In A- and Z-DNA, water molecules that bridge anionic phosphate oxygens form apparent strings of water along the duplex backbone. In fact, it has been proposed that the more economical hydration of phosphate oxygens in A- and Z-DNA relative to B-DNA is the driving force behind the B-to-A and B-to-Z transitions observed in conditions of reduced water activity (20).

The periodic order of double-stranded nucleic acids facilitates arrangement of water molecules into regular hydration networks in and at the grooves. Except for B-DNA with its strong, sequence-dependent pattern of minor groove hydration, polar atoms in the DNA grooves are well hydrated at about half the level of the O1P and O2P oxygen atoms (17-19). In the minor groove of AT-rich domains of B-DNA, water molecules that bridge the adenine N3 and thymine O2 atoms are, in turn, connected via other waters, there by creating the spine of hydration. A similar spine of hydration has been reported for the minor groove of Z-form d(GCCCGC), where water molecules form hydration networks that bridge O2 cytosine atoms on opposite strands (19, 22). The major groove of A-DNA is narrower than that of B-DNA. This disparity is in inverse proportion to minor groove hydration, polar atoms in the A-DNA octamer d(GGTATACCC), and decamer d(GGGTGATACCC), the water molecules in the major groove are arranged in fused pentagons and hexagons that connect the phosphate backbone with the bases (18).
The spine of hydration in the central AA TT part of d(CGCGAATTCCGG)\(_2\) is, perhaps, the most distinctive feature of the Dickerson-Drew B-DNA dodecamer (see Fig. 2) (22). The spine represents a string of waters at the floor of the minor groove spanning O2 of thymine and N3 of adenine, which are brought into close proximity by helix rotation. The spine of hydration is stabilized additionally by hydrogen bonding to the sugar O4' atoms at the bottom of the minor groove (23). These first hydration shell waters are coordinated tetrahedrally and interconnected by second layer waters. A more recent high-resolution X-ray crystallographic structure of the Dickerson-Drew dodecamer revealed a third hydration layer that consists of a second spine that runs parallel to the inner spine at the periphery of the minor groove (24, 25). With the outer spine, the zigzag-shaped inner spine represents the lower part of four nearly planar fused hexagons that dissect the minor groove along the central hexamer portion of the dodecamer duplex (25). The spine of hydration is disrupted sterically in GC-rich domains of B-DNA because of the presence of the guanine N2 amino group in the minor groove.

**Dynamics of DNA Hydration**

Water molecules in the solvation shell of DNA are able to exchange with the bulk solvent with relative ease with the rate of exchange depending on the nature of DNA-water interactions. Insight into the kinetics of solvent exchange and the dynamics of water of hydration is derived from NMR and time-resolved fluorescent techniques. These techniques are mutually reinforcing; they are sensitive to different time domains, different subpopulations of water, and different physical characteristics related to water motion. NMR methods are most effective to characterize water molecules with lifetimes of 0.1 ns and longer. These relatively long-lived waters may play important structural roles. On the other hand, time-dependent fluorescent techniques are sensitive to short-lived, nonstructural waters of hydration with lifetimes in the femtosecond to picosecond range.

The use of NMR in hydration studies is based on two complementary approaches. In one approach, nuclear Overhauser effect (NOE) cross-peaks between DNA protons and hydration water are used to gain insight into the dynamics of localized water...
molecules near nonexchangeable DNA protons. The second approach uses the nuclear magnetic relaxation dispersion (MRD) spectrum of DNA to characterize the relaxation properties of perturbed waters without specifying their relative positions. Intramolecular magnetization relaxation and fast proton exchange by labile DNA protons may complicate interpretation of DNA water cross-peaks in NOE measurements and also may affect interpretation of \( n^2 \) MRD results. Such complications are avoided in \( \text{O}^1 \) and \( \text{H}^2 \) MRD measurements that are uniquely related to water relaxation (26).

Most NMR studies on DNA hydration have investigated the spine of hydration in the minor groove of the Dickerson-Drew dodecamer and related sequences (27). NOE studies have revealed the presence of long-lived water molecules in the minor groove of AT-rich domains of B-DNA with residence times of \( \sim 1 \text{ ns} \) (26–28). Consistent with these results, MRD measurements with and without the minor groove binding agent netropsin also have located long-lived waters with a residence time of 0.9 ± 0.1 ns in the minor groove of the Dickerson-Drew dodecamer (26, 28). NOE measurements also have detected waters with lifetimes in excess of 0.5 ns in oligonucleotides with a central 5′-TTAAN-3′ domain, when \( N \) is either T or C (29, 30). By contrast, no long-lived waters in the minor groove are detected when the flanking nucleotide \( N \) is G or A (29, 30). These observations suggest subtle sequence-dependent effects of minor groove hydration that may correlate with the groove geometry.

NMR studies of water dynamics in the major groove of B-DNA are sparse. Although NOE cross-peaks between water and thymine methyl protons and guanine 8H and adenine 8H protons in the major groove of the Dickerson-Drew dodecamer have been identified, the estimated lifetimes on the order of 200 to 500 ps are short compared with that of water in the minor groove (26, 28). MRD measurements at 4\(^\circ\)200 to 500 ps are short compared with that of water in the minor groove. Most experimental techniques provide an average residence time for the entire population of fast-lived DNA water molecules.

**Computer Simulations**

Theoretical approaches have emerged as a powerful means to characterize nucleic acid hydration at a level not restricted by the limitations of experimental techniques. In particular, molecular dynamics (MD) computer simulations have provided a wealth of information on the structure and dynamics of water in the vicinity of nucleic acids (12). Recent advances also identified short-lived water molecules with residence times of \( \sim 30 \text{ ps} \) that are not associated with the minor groove of B-DNA (28). Surprisingly, this MRD-based estimate is in good agreement with fluorescent-detected hydration lifetimes discussed below. It represents an average residence time for the entire population of fast-lived DNA water molecules.

Time-resolved emission spectroscopy (TRES), also referred to as time-resolved Stokes shift spectroscopy, enables one to derive information about the dynamics of biopolymer-solvent interactions on the femtosecond to nanosecond time scales, provided that suitable solvatochromatic fluorescent probes have been identified. Such probes should exhibit significant Stokes shifts that change with solvent polarity and should have fluorescent lifetimes on the order of the dynamic solvent exchange process or longer. TRES detects solvent dynamics that influences the energy difference between the excited and the ground states of the fluorophore and is insensitive to dynamic processes that are significantly slower than the fluorescence lifetime.

The fluorescence lifetimes of the four natural DNA bases within the femtosecond range are too short to be useful as spectroscopic probes to characterize the solvation dynamics of DNA (31). The minor groove binder Hoechst 33258, the fluorescent adenine analog 2-Aminopurine (2A.p), and the base pair mimic coumarine 102 abasic site are better probes because they exhibit longer lifetimes and better solvatochromatic properties. Of the two base pair analogs, the coumarine 102-abasic site base pair mimic disrupts the native DNA structure more than the adenine analog 2A.p. However, the spectroscopic characteristics of the former are superior to those of 2A.p for probing solvation dynamics.

Using 2A.p and Hoechst 33258 as probes, Pal and Zewail (32) have identified two groups of weakly bound water molecules in the vicinity of DNA with residence times of \( \sim 1 \text{ ps} \) and 12 to 19 ps. Subsequently, the Berg group (33) using the coumarine 102-abasic site base pair mimic has discovered power law solvation dynamics that range over six orders of magnitude from the low femtoseconds to the high nanoseconds. The 1-ps and 19-ps solvation dynamics detected by Pal and Zewail represent a part of the overall solvation dynamics detected by the Berg group (33). These studies provided unique insights into the ultrafast solvation dynamics at the site of the modified base. However, it is yet to be understood how these results relate to the hydration dynamics of unperturbed DNA. A large body of evidence suggests that even minor chemical modifications, such as the adenine to 2A.p replacement, may impact significantly the structure and thermodynamics of DNA. By implication, it is not unreasonable to anticipate that the solvation dynamics at the probe site may also be perturbed by the chemical modification.
Nucleic Acid Hydration

A- and B-conformations contain, respectively, 17.7 and 19.3 tightly bound waters. GC base pairs in A- and B-conformations interact with water more strongly than A-T base pairs in both A- and B-conformations. In B-DNA, this disparity reflects partially the wider minor groove of GC base pairs and their weaker interaction with DNA. B-DNA contains in its hydration shell one more water molecules in the first hydration layer interact weakly with DNA. GC base pairs interact with water more strongly than AT base pairs in both A- and B-conformations. In B-DNA, this disparity reflects partially the wider minor groove of GC base pairs and their more extensive hydration (three solvation sites per base pair). This sequence-dependent pattern of hydration agrees with the more extensive hydration (three solvation sites per base pair).

Protein–DNA Recognition

Water plays a major role in modulating DNA recognition by drugs and proteins. In protein-DNA recognition, water plays a structural role in addition to its thermodynamic role. In fact, it has been suggested that water is a critical element of the recognition code, mediating interactions that would be less favorable in its absence (33). Interfaces of specific and nonspecific protein-DNA complexes can be significantly hydrated, often with water molecules that form bridges between amino acid residues and nucleic acid bases. These water molecules mediate the specificity of protein-DNA recognition by screening unfavorable electrostatic interactions and by facilitating formation of water-mediated networks of protein-DNA hydrogen bonds. For example, the crystal structure of the paired homeodomain dimer complexed with DNA displays 18 ordered waters that mediate contacts between the protein and the DNA major groove (36). Molecular dynamics simulations of the estrogen receptor binding to consensus and nonconsensus DNA sequences have indicated that binding specificity and stability is conferred by a network of direct and indirect (water mediated) protein-DNA hydrogen bonds (37). Significantly, the fluctuating network of hydrogen bonds between the receptor and the nonconsensus DNA facilitates penetration of water molecules to the protein-DNA interface.

The trp repressor represents a striking example of the structural aspect of hydration: all but one interaction between the protein and the operator DNA are mediated by water molecules (38). A comparison of high-resolution structures of the free and bound states of the trp operator DNA has revealed that the hydration sites in the two structures are the same (39). Thus, conserved water molecules mediate trp repressor-operator contacts, consistent with a picture in which protein and DNA recognize each other’s hydration patterns. A subsequent analysis of 11 protein-nucleic acid complexes has shown that the positions of polar protein atoms hydrogen bonded directly to DNA groups generally correspond to hydration sites that would normally be occupied by water molecules in unbound DNA (40). This observation is consistent with a picture in which the hydration sites of free DNA mark the protein binding sites at the protein-DNA functional interface. A similar conclusion has been reached based on grand canonical Monte Carlo simulations of the BamHI complexes with noncognate and cognate DNA sequences (41). Based on these results, it has been concluded that interfacial waters can serve as a “hydration fingerprint” for a given DNA sequence that guides its recognition by DNA-binding proteins.

Hydration has been suggested to play a functional role in which it facilitates discrimination by a DNA-binding protein between the specific and the nonspecific sites by linear diffusion or “sliding” (42, 43). This possibility was emphasized by von Hippel who wrote, “Can one really think of the protein-DNA interface of a nonspecific complex as retaining full hydration, and what role does the expulsion of water bound to polar groups at the interface play in stabilizing the specific complexes that form when the DNA target site is reached?” (42). A partial answer to this question stems from the fact that little, if any, dehydration accompanies the formation of a nonspecific protein-DNA complex as suggested by near-zero changes in heat capacity $\Delta C_H$ (43–45). By contrast, very large and negative changes in $\Delta C_H$ that accompany specific protein-DNA association suggest extensive dehydration of specific protein-DNA complexes (42, 44, 45).

Differential changes in hydration associated with the binding of several DNA-binding proteins to cognate and noncognate sites have also been detected by osmotic stress measurements. For example, the binding of the restriction endonuclease EcoRI to its specific DNA sequence is accompanied by a release of $\sim$110 more waters than binding to a nonspecific sequence (46). Based on the independence of this number on the osmolyte type, it...
has been proposed that the nonspecific complex sequencers 110 waters in a space between the interacting protein and the DNA surface that is sterically inaccessible to solutes. More recently, osmotic stress measurements have been used to detect ordered (crystallographically detectable) and unordered (crystallographically undetectable) waters in a space between the interacting protein and the DNA surface that is sterically inaccessible to solutes. From the time of the discovery of the double-helix structure of DNA, it has become increasingly clear that solute–solvent interactions represent the very foundation of the conformational preferences, stability, and biological functions of nucleic acids. In recognition of this fact, many experimental and theoretic studies explore the hydration properties of nucleic acids. The structural picture of DNA and RNA hydration that is emerging from X-ray and NMR studies is robust. However, hydration is not limited to structurally ordered water molecules; it involves a large number of unordered waters. The latter population of water molecules is more challenging to detect and to characterize, although their thermodynamic impact on nucleic acid structure and recognition may be as important as, or even exceed that, of ordered waters. More concerted experimental and theoretic studies are required to fully understand the role of water in fine tuning the structure and recognition of nucleic acids. The emerging insight may well be the key to understanding the molecular mechanisms of gene expression and control.

Concluding Remarks

From the time of the discovery of the double-helix structure of DNA, it has become increasingly clear that solute–solvent interactions represent the very foundation of the conformational preferences, stability, and biological functions of nucleic acids. In recognition of this fact, many experimental and theoretic studies explore the hydration properties of nucleic acids. The structural picture of DNA and RNA hydration that is emerging from X-ray and NMR studies is robust. However, hydration is not limited to structurally ordered water molecules; it involves a large number of unordered waters. The latter population of water molecules is more challenging to detect and to characterize, although their thermodynamic impact on nucleic acid structure and recognition may be as important as, or even exceed that, of ordered waters. More concerted experimental and theoretic studies are required to fully understand the role of water in fine tuning the structure and recognition of nucleic acids. The emerging insight may well be the key to understanding the molecular mechanisms of gene expression and control.

References


Roles of Water in Biological Recognition Processes

Martin F. Chaplin, London South Bank University, London, United Kingdom

doi: 10.1002/9780470048672.wecb636

Water is central to biology, providing much more than the means to allow the movement of biologic molecules. It is an excellent solvent, a substrate and facilitator for reactions, and it provides a means for ionizations and extends the intricate three-dimensional surface geometry and complex charge distribution of biomolecules out towards the bulk solution. Water’s hydrogen bonds and dipoles are strong enough to present structural information concerning surfaces to distant diffusing molecules and to fine tune their orientation to match potential recognition sites, but they are weak enough to break to make way for direct binding processes. Water occupies potential interaction sites until they are recognized by suitable molecules, with the release of the water providing free energy to help compensate for the entropy loss on binding. Also, water molecules may determine the conformation of the biological molecule and directly assist and take part in the recognition process. Among the techniques available, X-ray crystallography has provided the most information on these roles for water.

Aqueous Hydrogen Bonding in Biology

Water is suited highly to its role as the medium of life and clearly has guided molecular evolution. It is the third smallest molecule known, and so it can penetrate the smallest of structural crevices. It is polar and capable of hydrating and shielding ions and dissolving a wide range of materials, both polar and small nonpolar. It can bind and release hydrogen ions to create a rich ionic environment with long-range interactions. Most importantly, water can form highly directional hydrogen bonds that allow interactions with electronegative atoms and link together to form chains and clusters. All biomolecules function within this aqueous environment; with much of biology being guided by water that facilitates highly specific recognition processes (1).

The water molecule

The water molecule (H₂O) consists of an electronegative oxygen atom holding tightly to the electrons from two hydrogen atoms, which leaves their positively charged nuclei partially exposed (Fig. 1a). Thus, water has a dipole that will interact with charged molecules and ions to solvate and to shield them from other charged species. Water ionizes to give hydrated protons and hydroxide ions. Importantly it can also carry excess protons and act as a proton wire to transfer protons rapidly from one site to another in biomolecular systems (2).

The disparity in the mass of the oxygen and hydrogen atoms allows the hydrogen nuclei to vibrate with considerable movement, which takes them closer to interact with other atoms and endows them with more freedom of movement. Because of water’s small size, a high density of water molecules exists in solutions that produce a high density of molecular interactions, which make large and important enthalpic and entropic contributions. In many circumstances, the free energies of interactions are relatively small because of the compensation between these terms; stronger binding to water is associated with both lower enthalpy and lower entropy (3). In other cases, such interactions can provide substantial entropic or enthalpic drive to direct a recognition process.

Hydrogen bonding

Water molecules interact with each other and with electronegative atoms by means of roughly, tetrahedrally positioned hydrogen bonds (Fig. 1b). These are modest, mainly electrostatic, directed interactions. Individually, they are strong compared with thermal fluctuations, although far weaker than covalent bonds, giving water its powerful hydration properties. The small molecular size and the extensive vibration of the hydrogen nuclei in water causes individual hydrogen bonds to have only short lifetimes (À ps), although they often will reform after...
breaking. Thus, water possesses significant thermodynamic hydration power sufficient to affect the structure of the hydrated biomolecules, but the rapid kinetics for hydrogen bond formation and breakage causes the hydrating water molecules to exchange rapidly, which otherwise might be taken for weak binding. This counterintuitive behavior is key to much of the use and the properties of water within biomolecular interactions. In particular, it allows water to orient molecules strongly according to neighboring surface characteristics but to move rapidly out of the way when required to allow biomolecular binding.

The heterogeneous nature of water is responsible for local correlations between the orientations of the molecules. In turn, it endows liquid water with its high dielectric and its consequent ability to shield electrostatic interactions. Also, it allows water molecules to transfer information concerning the structure of surfaces outward towards the bulk of the solution. The strength of water’s hydrogen bonds depends on the strength and the direction of the other hydrogen bonding that involves the same water molecules. The preference of water molecules for balanced numbers of donor and acceptor hydrogen bonds results in both cooperative and anticooperative effects: an accepted hydrogen bond strengthening a water molecule’s hydrogen bond donating ability but weakening its ability to accept a second hydrogen bond (Fig. 1b) (4). Some strong hydrogen bonds in biological systems are formed when water donates a hydrogen bond to charge-dense oxygen atoms in carboxylate and phosphate groups, or when water accepts a hydrogen bond from positively charged amino groups or oxonium ions.

The high density of hydrogen bonding sites in liquid water, together with their flexibility, allows water to act as a lubricant for translational movement between biomolecular surfaces while retaining hydrogen bonding and consequent control over the orientation of the moving molecules.

The heterogeneous nature of water

Under circumstances that favor hydrogen bonding, the density of water expands toward that of fully tetrahedral hydrogen-bonded ice (5). Such a structure has more negative enthalpy and lower entropy. When ice melts, this expanded structure collapses somewhat to the liquid state, but it still contains much of its hydrogen bonding. Liquid water is heterogeneous with fluctuating volumes of lower density, in which more hydrogen bonds are changing back and forth, into volumes of higher density where more broken hydrogen bonds and greater van der Waals interactions exist (3). Thus, aqueous systems are in constant flux with the local equilibria between the low-density and high-density states shifted by different surfaces, groups, and ions (6). Where the less dense structuring is found, water is more structured and viscous, whereas denser patches with greater dangling hydroxyl groups are more reactive, less viscous, and contain more rapidly diffusing molecules. The characteristics of the biomolecular structure are responsible primarily for controlling the structuring in the surface water layers. Thus, the ease of diffusion for visiting molecules is controlled both toward and away from areas of biomolecular surfaces.

The hydrophobic effect

When water lies next to hydrophobic surfaces, it cannot form as many hydrogen bonds as when it is in the bulk of the liquid. It can maximize its van der Waals contacts and hydrogen bonding to itself and to the bulk by forming loose, transient networks of pentagonal rings over some areas. Otherwise, these surfaces are energetically unfavored. Water molecules next to such surfaces can translate sideways and away from the surface more easily than they can rotate, and many hydroxyl groups will be left without hydrogen-bonding partners. Hence, biomolecular hydrophobic-binding sites usually are hydrated poorly, with the water molecules they contain unable to form their optimum number of hydrogen bonds, and often they possess moderately high entropy when compared with bulk water molecules. When displaced by a complementary-shaped hydrophobic ligand, the entropy loss on binding between the biomolecules will be compensated by the binding free energy due to the complementary surface dispersion enthalpy and both enthalpic and entropic contributions from the released water.

Figure 1 (a) The structure of an H2O molecule in liquid water, which shows the average bond length and angle, the approximate outline shape, and the surface electrostatic potential. (b) A typical tetrahedral water cluster found in liquid water that shows typical dimensions of its hydrogen bonds. Donating (d) and accepting (a) hydrogen bonds are indicated. Individual bond lengths and angles vary considerably around those given.
Water binding

The strength of binding of water to and between biomolecules is difficult to determine because of the complex nature of hydration interactions and the static and dynamic effects of the local environment. Binding consists of enthalpic and entropic terms and includes the effects of the rearrangement, removal, or addition of water within its surroundings. Entropy changes include hydrogen bonds, electrostatic terms, and dispersion terms, whereas entropy changes include conformational restrictions in the biomolecules and changes in both the rotational and the translational freedom of the water molecules reflecting differences in the organization between the bulk and interfaces. The process of binding (as exemplified by protein–ligand binding) can be described by the following equilibria:

\[ P + H_2O_{bulk} = Paq_1 + H_2O_{aq1, aq2} \]  
\[ L + H_2O_{bulk} = L_{aq1} + H_2O_{aq1, aq2} \]  
\[ Paq_1 + L_{aq1} = PL_{aq2} + H_2O_{aq2, aq3} \]

where the subscripted terms “bulk” and “aq” represent different amounts of water. Although the organization of the water at the interfaces differs on binding, most remaining interfacial water molecules retain their original anchoring hydrogen bond(s). The last equation (Equation 3) determines the binding free energy, which in turn differentiates recognition in terms of the specificity of different ligands. As a clearly an entropic cost (\( T \Delta S \) is negative) exists in binding \( P \) to \( L \), then compensation must exist in terms of any combination of entropy gain (\( T \Delta S \) is positive) by the water released, entropy gain by remaining water at the interfaces (but usually entropy is lost here), enthalpy gain (\( \Delta H \) is negative) in the water released, enthalpy gain by remaining water at the interfaces, and enthalpy gain in the interactions between \( P \) and \( L \). It should be noted that apart from carboxylate and phosphate oxygen atoms, water energetically prefers to hydrogen bond to itself rather than to biomolecules, if it can.

It is known that water mediates protein folding by orienting the polypeptide strands and manages their interactions as they move from being solvent separated to forming the direct peptide hydrogen bonds found in their secondary structures (7). Although it has not been shown with the same degree of proof, it is likely that similar processes operate in macromolecular binding. Water has both dynamic and structural roles. The ease with which the water molecules move relative to each other and the biomolecular surfaces determines the dynamics, whereas the strength of the hydrogen bonding linking areas on this surface and between surfaces determines its structural effects. Dynamically, it can control and guide the process by its proof, it is likely that similar processes operate in macromolecular binding. Water molecules coat biomolecular, ionic, and solute surfaces in aqueous solution, minimizing the system free energy using van der Waals contacts, electrostatic interactions, and hydrogen bonds to compensate partially for the entropic and enthalpic cost of their removal from the bulk phase. They even enter hydrophobic spaces, which maximizes their van der Waals surface contacts and hydrogen bonds as far as possible.

Water molecules hydrogen bond to several biomolecular groups. Of particular note are the donated hydrogen bonds to the oxygen atoms in carboxylate, phosphate, and peptide carbonyl oxygen atoms and the accepted hydrogen bonds from peptide amines, other neutral and positively charged amines, and alcohol groups. When binding within or between biomolecules, water usually forms three or four hydrogen bonds and rarely (if ever) forms just two hydrogen bonds of the same form, accepting or donating, because of their anticooperativity (4).

Usually, only bridges that consist of single water molecules are considered, because bridging by chains of water molecules are more difficult to determine in structural studies because of their kinetic instability. However, such extensive bridges are important in the overall thermodynamics and stereospecificity of binding. Together, these aqueous interactions probably are decisive in guiding conformational rearrangements, differentiating potential binding sites, and selecting those eventually chosen (8). Normally, a mixture of direct and water-mediated links exist in about equal amounts with the direct links between the biomolecules being relatively more directed and inflexible and kinetically more difficult to reverse without the help from neighboring hydration.

The role of water in mediating binding differs between the biomolecules because of variations in the strength of the natural hydration of the interfaces, which approximates roughly to DNA > proteins > ligands.

Protein–protein recognition

The protein surface is varied in shape, structure, and charge, so it is no surprise that its hydration is varied equally with both static interactions and weak interactions in about equal measure. Water molecules form an extensive network around proteins, which extend its surface characteristics out into the bulk. Often, this extended network seems rather thin to many analytical methods and current modeling, but it is thought likely that its influence extends somewhat further to several water layers. The amount of structured water around a protein controls the protein’s dynamics, which in turn may control its binding to other biomolecules. Water is important for the structure of individual proteins, their linking to form quaternary structures, and binding to form molecular clusters. Although

Biological Recognition

Our appreciation of biological recognition is different today from that of a few years ago. Then, the focus was concerned primarily with the complementary molecular surfaces (e.g., the lock-and-key mechanism), together with appropriately placed hydrogen bonding and electrostatic interactions between hydrophilic groups. Water was assumed simply to make way for the binding. Today, we know that water controls the kinetics and thermodynamics totally and must be included in any accurate description of these processes.

Biomolecules and water

Water molecules coat biomolecular, ionic, and solute surfaces in aqueous solution, minimizing the system free energy using van der Waals contacts, electrostatic interactions, and hydrogen bonds to compensate partially for the entropic and enthalpic cost of their removal from the bulk phase. They even enter hydrophobic spaces, which maximizes their van der Waals surface contacts and hydrogen bonds as far as possible.
Roles of Water in Biological Recognition Processes

such structuring may help recognition by hydrophilic sites, it will mask complementary structures where dehydration is required. In general, the hydrogen bonding to hydrating water molecules in binding sites is poor so that it may be removed easily both kinetically and thermodynamically.

Proteins may make specific links through their side chains or less specific links of the nature of peptide linkages. Using water molecules to connect these groups, rather than having direct interactions, gives a greater degree of freedom in terms of the relative protein–protein interaction. This analysis indicated that water molecules contributed to the close packing of atoms that ensured complementarity between the protein surfaces and that provided suitably situated polar interactions. Both studies showed that some interfaces were essentially dry with water molecules distributed around the periphery of the contact, whereas others were hydrated within the central region of the interfaces. As an example, the E9 deoxyribonuclease (DNase) and immunity protein Im9 protein–protein complexes, including 24 protease–inhibitor, 19 antibody–antigen, and 11 that are involved in signal transduction. This analysis indicated that water molecules contributed to the close packing of atoms that ensured complementarity between the protein surfaces and that provided suitably situated polar interactions. Both studies showed that some interfaces were essentially dry with water molecules distributed around the periphery of the contact, whereas others were hydrated within the central region of the interfaces. As an example, the E9 deoxyribonuclease (DNase) and immunity protein Im9

Protein–ligand recognition

The binding sites of proteins are, or are surrounded by, relatively rich areas for binding water. Many biological ligands, on the other hand, contain relatively few groups that can hydrogen bond to water. An extreme example of these ligands is phosphothioate A7 (13), which has a hydrophobic catalytic site held some distance from the substrate phospholipid surface and requires rigidly ordered water molecules to surround the active site and to maintain this interface of binding.

A network of 392 crystal ligand–protein structures (14) showed an average of 4.6 ligand-bound water molecules, with 76% of them bridging between the ligand and the protein. On average, three interactions existed that involved each of these linking water molecules with the protein–ligand complex: two to the protein and one to the ligand. A separate analysis of 251 protein–ligand active-site complexes gave an average of 34 water molecules and 118 aqueous hydrogen bonds per active site, with average coordination of just over two strong hydrogen bonds per water molecule (15). In addition, these water molecules were found to lie close to carbon-bound hydrogen atoms, which increased their van der Waals contacts. Most aqueous hydrogen bonds were to the protein rather than to the ligand, with an intermediate number of water–water hydrogen bonds. Using water to link the ligand and the protein can broaden the specificity due to the extra stereochemical flexibility. Thus, the peptide–binding protein OsP4A uses several flexible adaptively water molecules to hydrogen bond and to shield charges when binding lysine frayed tripeptides KXX, where K represents lysine and X is any one of the 20 common amino acids (16). Before such binding, these bridging water molecules are held by two or more hydrogen bonds to the unliganded protein, and they are conserved generally in the liganded protein crystals.

The hydrophobic interaction is an important factor in molecular recognition (17). Thus, the streptavidin–binding site for biotin contains a five-membered cyclic pentameric ring of water molecules stabilized by additional hydrogen bonds to coplanar amino acid side chains from the streptavidin and enclosed top and bottom by hydrophilic groups. When these highly ordered water molecules are expelled to make way for biotin, there is an entropy gain that more than compensates for the entropy loss on the strong streptavidin–biotin binding.

Peptide inhibitors are used in highly active antiretroviral thera-

DNA–DNA recognition

Although the structure of the DNA double helix is well known to be caused by base pairing and base stacking, the form of the helix is also highly dependent on its state of hydration. Hydration is crucially important for the conformation and use of nucleic acids and for its recognition by itself and other molecules (21). The strengths of many aqueous interactions are greater than those for proteins because of their highly ionic character. The DNA double helix can take up several conformations with differing hydration. The predominant natural DNA, B-DNA, has a wide and deep major groove as well as a narrow and deep minor groove and requires the greatest hydration. Partial dehydration converts it to A-DNA (with a narrow and deep major groove)
by decreasing the free energy required for A-DNA deformation and twisting, which is employed usefully by encouraging supercoiling but eventually leads to denaturation.

Hydration is greater and is held more strongly around the phosphate groups that run along the inner edges of the major grooves, but it is more ordered and more persistent around the bases with their more directional hydrogen-bonding ability and restricted space. Because of the regular structure of DNA, hydrating water forms chains along the double helix in both the major and the minor grooves. The cooperative nature of this hydration aids both the zipping (annealing) and unzipping (unwinding) of the double helix. Water motion within the grooves is slowed down compared with the bulk water, with the greatest reduction within the more restricting minor groove (22). On separation of the double helix (i.e., melting), about four water molecules per base pair are released despite extra hydration sites being released by the previously hydrogen-bonded base pairing (23), which confirms the importance of the cooperative nature of the water binding within the grooves.

**DNA–protein recognition**

The major problem in DNA protein recognition is how the protein can rapidly find a particular DNA base sequence from the enormous choice available when the specificity of the bases is somewhat hidden by the hydrogen bond linked and stacked base pairs. In an analysis of 109 protein-DNA complexes, an average of about nine water molecules per complex were bridged between the protein and the DNA, with another 125 linking to one or the other (24). Most water molecules were useful in screening unfavorable electrostatics from phosphates to allow binding (Fig. 3). The most common link found was where the water molecule donated hydrogen bonds to both a side chain and accepting phosphate oxygen atoms (25). Another analysis of 39 crystallographically characterized protein-DNA complexes indicated that 46% of bridging waters (32 per complex) linked the protein to the nucleotide bases and the remainder was useful in screening out the unfavorable electrostatics, being linked very strongly to the DNA phosphate groups (25).

Figure 2  Cartoon that shows water-assisted asparagine recognition and aspartate discrimination by asparaginyl-tRNA synthetase. (a) Among the bound water molecules in the unliganded enzyme, one (Wat1) is bound between the Gin259 peptide carboxyl and the Gin230 carboxylate and another (Wat4) between the side chains of Tyr333 and Arg364. (b) These water molecules link to the amide group on the asparagine-AMP substrate on binding and two water molecules are released (Wat2 and Wat3). Aspartate-AMP cannot bind at the same site as the Wat1 water molecule is not able to donate three hydrogen bonds. Some hydrogen atoms have been added to clarify the hydrogen bonding. Figure partially redrawn from Reference 17 with permission.

The restriction endonuclease MspI makes specific contacts between donating protein-charged amino or guanidinium groups and accepting phosphate oxygen atoms. A further analysis of 39 crystallographically characterized protein-DNA complexes indicated that 46% of bridging waters (32 per complex) linked the protein to the nucleotide bases and the remainder was useful in screening out the unfavorable electrostatics, being linked very strongly to the DNA phosphate groups (25).

Water-mediated links are also important in nuclear hormone receptors such as the estrogen receptor-DNA complex. A major question concerns how a protein detects speedily a specific DNA sequence-binding site from the many closely related sites available. Restriction endonucleases have to find rapidly their specific hydrolytic sites in DNA molecules among the many similar sites, which differ by as little as a single base pair. They do this first by scanning the DNA rapidly at a distance, held at low affinity by favorable electrostatic interactions but separated by a significant but lubricating hydration layer. This process is then followed by partial dehydration and binding at their specific sites (27). Thus, the complexes of the restriction endonucleases BamHI and EcoRI with nonspecific sections of DNA have about 100-150 more bound water at the interface than the specific complexes (28, 29), which distances the protein from the DNA. The large amount of these water molecules released subsequently to the bulk liquid on specific binding contributes favorably to the binding thermodynamics. The specificity is almost total with a single substitution of thymine at the beginning of the specific nucleic acid sequence GAAATTC (to give TAATTC) removing nearly all binding (5000 × less) by EcoRI (29).

The restriction endonuclease MspI makes specific contacts with all eight bases in the four base pair recognition sequence (CCGG), by six direct and five water-mediated hydrogen bonds and 13 water-mediated links to the phosphates (30). Number...
Roles of Water in Biological Recognition Processes

Figure 3 Cartoon that shows how water molecules shield the phosphate charges from the protein by donating hydrogen bonds (shown as dashed lines) to acceptor atoms on both the DNA and the protein, to attenuate their electrostatic repulsion. A need exists for these water molecules to accept hydrogen bonds from other water molecules (as shown) or for protein groups to stabilize such links.

van der Waals contacts exist as water molecules contribute to the close packing, which mediates shape complementation. A key feature of direct protein DNA links in bound complexes is that they can change to and from water-mediated links to allow movement within binding cleft (31) while retaining contact information.

Investigative Methods

No clear-cut method exists that gives complete or unambiguous information concerning the biomolecular binding processes in aqueous solution. The main problem is that although water’s hydrogen bonds are significant thermodynamically, water molecules are very mobile individually, as described earlier. Water molecules, therefore, have significant thermodynamic effects on recognition that are difficult to pin down by diffraction or NMR, because of their constant movement. A hydrogen bond only has a lifetime measured in picoseconds, and even an individual H₂O molecule exists for less than a millisecond. In addition, explicit hydration is not included easily within molecular simulations because of the complexity of the inter–water and water–surface interactions. Many computer models for water have been developed, but they have very limited use in forecasting accurately the properties of bulk water. However, their predictions are probably more accurate close to surfaces than at intermediate distances where greater complexity exists because of the bulk aqueous milieu. Therefore, if care is taken, simulations can be used to help determine the structure and the properties at biologic interfaces (32).

The most useful data in many of the above studies is from X-ray and neutron diffraction, which provides the average positions of some water molecules in the free and the bound complexes. Although the structural information concerns crystals only, it is believed that the stereochemistries of many biomolecules within crystals are often close to those found in biological situations. Both methods detect water molecules in favored sites but fail to detect exactly where the most water molecules are present at any given time. Also, X-ray data cannot provide direct information that concerns the hydrogen bonding as the hydrogen atoms cannot be observed. The data does, however, provide a solid foundation and can be used as a starting point in molecular simulations.

NMR can be used to determine the presence and the movement of water molecules next to biomolecular surfaces (33). Sometimes it is particularly useful to detect the links to buried water and to water molecules that mediate between biomolecules.

Occasionally, vibrational spectroscopy can provide information that concerns specific water molecules or clusters when they are distinct sufficiently from the wide range of such structures found in the bulk liquid (34).

Calorimetry determines the heat generated (ΔH) during binding and solvation processes. Together with the binding constant for the binding equilibrium (providing ΔG), the entropy changes during binding (ΔS) can also be determined. These values are interesting and useful but fail to provide precise or accurate molecular details concerning the binding even when a range of similar molecules are compared.

Determination of the osmotic pressure of solutions in the presence of other solutes and biologic-binding processes can indicate clearly changes in the amount of water bound or, more exactly, the amount of water that can no longer interact freely with the solutes. Such methods have been useful particularly to determine the DNA–protein interactions (39).
Conclusion: Knowns and Unknowns

The more information is discovered concerning the role of water in biological recognition processes, the more its importance is confirmed. Where diffraction data is available we have, or can get, a good appreciation of the biological recognition process and the importance of the water molecules at the interface. However, currently we lack the means to determine the importance of water molecules removed slightly from the interface or details of how they guide the binding process, if indeed they do. While retaining the structural base from diffraction studies, the development of more accurate molecular simulations together with more powerful computers and concentrating on the interfaces and neighboring water without washing too much resource on inconsequential distant bulk water (for example, Reference 35) is clearly a way forward.

References

Roles of Water in Biological Recognition Processes

Further Reading


See Also

Biointeractions at Solid-Water Interfaces
NMR to Study Molecular Recognition
Water in Protein Folding, Role of
Water, Properties of, Water Channels
Water, Properties of
Alfons Geiger and Dietmar Paschek, Physikalische Chemie, Technische Universität Dortmund, Germany
doi: 10.1002/9780470048672.wecb627

Water is well known for its unusual properties, which are the so-called “anomalies” of the pure liquid, as well as for its special behavior as solvent, such as the hydrophobic hydration effects. During the past few years, a wealth of new insights into the origin of these features has been obtained by various experimental approaches and from computer simulation studies. In this review, we discuss points of special interest in the current water research. These points comprise the unusual properties of supercooled water, including the occurrence of liquid–liquid phase transitions, the related structural changes, and the onset of the unusual temperature dependence of the dynamics of the water molecules. The problem of the hydrogen-bond network in the pure liquid, in aqueous mixtures and in solutions, can be approached by percolation theory. The properties of ionic and hydrophobic solvation are discussed in detail.

It is widely assumed that during the 4 billion years of evolution on our planet, life has adjusted to all properties of water and has taken advantage of any of the numerous unusual features of this liquid, which has been called “life’s matrix” (1–3). Water is made of the two most abundant cosmic elements besides the inert helium (4), and it has served as one of the initial compounds for the energetically induced reactions to form the first organic molecules. This reaction may have happened like in the Miller-Urey primordial earth atmosphere under the influence of electric discharge (5) or in the porous interior of icy comets under the influence of cosmic radiation (6). The abundance of water in our galaxy has been estimated as several ten-thousand earth oceans per sun, and it is distributed on planets, moons, comets, and dust grains mostly in the form of crystalline and amorphous ice. In the past few years, many new aspects of the microscopic structure, dynamic behavior, and role of water in biologically relevant molecular processes were obtained from computer simulation studies, which have been performed to understand various new experimental observations. Therefore, in this short review, we will strongly focus on the picture of water that developed from such computational studies. This review can by no means be complete but lists some points of special interest in the current water research. Because of the importance of water, numerous reviews and monographs on different aspects of water can be found (7–13).

Water appears in various condensed forms; 15 different crystalline ice structures are reported, as well as at least three amorphous (noncrystalline) ices and a similar number of metastable liquid water forms (14). This structural diversity has its origin in the elementary building blocks of water: the hydrogen bonds and the tetrahedral arrangement of H-bonded neighbors [which is often called the “Walrafen pentagon” (15)]. Both building blocks provide large structural flexibility. The strength as well as the directionality of the H-bonds is intermediate between van der Waals interactions and covalent bonds, which allows easy distortion of the perfect tetrahedral local arrangement and a versatile adjustment of the water structure to the changing thermodynamic conditions or the presence of solutes.

As we largely refer in this short review to simulation results, the following should be noted beforehand: The first molecular dynamics simulation studies of water by Rahman and Stillinger (16, 17) were received by the community of water researchers with great enthusiasm, as they demonstrated that the structural, dynamic, and thermodynamic properties of this complex liquid could be reproduced simultaneously to an astonishingly high degree by the use of a simple pair interaction potential, which was made up from Lennard-Jones and Coulombic contributions. This finding offered an unprecedented opportunity to study structural and dynamic details on the molecular level. To improve the reliability of the obtained simulation results by a better quantitative agreement between simulation model and real water, many different interaction potentials were developed since then. Numerous comparative studies have been published. A recent compilation can be found in Reference 18; other examples, which focus on special properties, are given in the following sections. Despite these efforts, no current model is fully satisfactory, but the interpretation of simulation results in comparison with real water can be improved by taking into account the shift of the phase diagram between model liquid and real water.
Anomalies of Water and Polyamorphism

The exceptional rank of water is manifested by its unusual properties compared with most other liquids, the so-called "anomalies." These anomalies comprise thermodynamic as well as structural and dynamic properties, especially their pressure and temperature dependence. The key to understanding these properties lies in two competing influences on the local structure: the attempt to build low-density tetrahedral structures (with low energy) versus the tendency toward closer packing (with higher entropy) (19).

The density maximum at 4°C and the decrease of volume on melting of ice are well-known anomalies. More aspects of the extraordinary behavior of water have been brought into the focus of many researchers by the seminal articles of Angell on supercooled water (20, 21). In contrast to "ordinary" liquids, the isothermal compressibility and the heat capacity of water increase drastically during supercooling. This finding indicates strongly increasing volume and entropy fluctuations during cooling. A spectacular explanation for this behavior was delivered by a computer simulation study, which gave evidence for the existence of a (second) critical point of water buried in the deeply supercooled liquid region (19, 22) [see also the reviews by Stanley and Debenedetti (23, 24)]. This second critical point is considered the endpoint of an equilibrium line between two forms of (metastable) liquid water: a low- and a high-density liquid.

The two different liquids have their counterparts in the amorphous solid state: the experimentally well-studied high-density amorphous (HDA) and low-density amorphous (LDA) ice forms (25). However, it is still unknown exactly how the different amorphous ice forms and supercooled liquid water are connected with the one second critical point. No "no man's land" region largely prohibits direct experimental access to the low temperature liquid because of the inevitable onset of crystallization (24) in this region. Therefore, computer simulation studies, in which crystallization does not take place, have been used extensively to establish the existence of a liquid-liquid transition. The location of the corresponding second critical point strongly depends on the interaction potential that was used in these simulations (26-28). It may be shifted to negative pressures, which are correlated with the prediction of a van der Waals-like model developed by Poule et al. (19), in which such a shift occurs with decreasing hydrogen bond strength. In such a scenario, the experimentally observed diverging fluctuations in supercooled water at ambient pressure do not develop by approach to the critical point; instead, these fluctuations develop by the approach to the spinodal line that emerges from the critical point at negative pressures. This finding could explain the early observation of Angell (20, 21), which suggested that all temperature-response functions and temperature coefficients diverge at the same temperature in ambient pressure supercooled water. The unavoidable crystallization occurs after passing the spinodal as it encounters a phase transition to the low-density liquid state, which has a local structure that is very similar to crystalline ice (29).

Some indirect experimental evidence exists for the liquid-liquid critical point hypothesis from the changing slope of the melting curves, which was observed for different ice polymorphs (30, 31). A more direct route to the deeply supercooled region, by confining water in nanopores to avoid crystallization, has been used more recently by experimentalists. These researchers applied neutron-scattering, dielectric, and NMR-relaxation measurements (32-35). These studies focus on the dynamic properties and will be discussed later. They indicate a continuous transition from the high to the low-density liquid at ambient pressure. The absence of a discontinuity in this case could be explained by a shift of the second critical point to positive pressures in the confinement. This finding correlated with simulations, which yield such a shift when water is confined in a hydrophilic nanopore (36).

A through the presented scenarios are still under discussion, the existence of a first-order-like transition between metastable high- and low-density supercooled water with a second critical point at negative pressures in bulk water and positive pressures in confinement is strongly suggested (20). Alternatively, singularity-free scenarios are discussed to explain the properties of supercooled water (24, 29).

As indicated above, the study of amorphous solid water in bulk and in confinement is an important source of information for the understanding of the liquid. In fact, water was the first liquid to show "polyamorphism": the mentioned existence of high density and low-density amorphs. A morphous solid water can be produced experimentally along very different routes by vapor deposition, by pressurizing crystalline ice, or by fast temperature quench of tiny droplets. Also, different subsequent annealing procedures have been used. Recently, also a very high-density form (VHDA) of amorphous water has been observed and shown to be distinct from HDA (37). Neutron scattering data revealed that the transformation between HDA and VHDA is related to an increasing population of "interstellar" water molecules (38). Simulation studies indicate that VHDA (not HDA) should be considered as the amorphous solid counterpart to the high-density liquid water phase at ambient conditions (39, 40). The question whether the HDA to VHDA transition is also first-order-like (as LDA to HDA) is not yet resolved (41, 42). The important influence of the preparation method in the first liquid to show "polyamorphism": the mentioned existence of high density and low-density amorphs. A morphous solid water can be produced experimentally along very different routes by vapor deposition, by pressurizing crystalline ice, or by fast temperature quench of tiny droplets. Also, different subsequent annealing procedures have been used. Recently, also a very high-density form (VHDA) of amorphous water has been observed and shown to be distinct from HDA (37). Neutron scattering data revealed that the transformation between HDA and VHDA is related to an increasing population of "interstellar" water molecules (38). Simulation studies indicate that VHDA (not HDA) should be considered as the amorphous solid counterpart to the high-density liquid water phase at ambient conditions (39, 40). The question whether the HDA to VHDA transition is also first-order-like (as LDA to HDA) is not yet resolved (41, 42). The important influence of the preparation method

Hydrogen Bond Network

In the perfect crystalline structure of "ordinary" (hexagonal) ice, each water molecule is H-bonded to four tetrahedrally
arranged neighbors. From the comparison of the enthalpies of sublimation, melting, and evaporation, it can be concluded that about 80% of all H-bonds survive the melting process. Despite such a wide range of possible definitions of intact versus broken H-bonds (46), it is therefore generally accepted that the water molecules in the liquid form at any instant a random, quasi-infinite, space-filling network (48). This network is subject to constant restructuring ("transient gel" (49)), the lifetime of the individual bonds are in the subpicosecond range (46, 50, 53). Computer simulations revealed that this network could be described quantitatively by combinatorial calculations and percolation theory (52).1

The physical mechanisms, which are connected with this primarily topological phenomenon of the existence of a percolating H-bonded network, are still not analyzed in detail. Nonetheless, several observations have been compiled recently that show a correlation between the existence of a spanning network and properties of physical and biological relevance (53–56). These observations concern the occurrence of phase separation in mixtures as well as the conformational transition and function of biomolecules. Computer simulations revealed that the phase separation in a water/tertarypropanol mixture is preceded by the formation of mesoscopic structures, but “sparse” H-bonded clusters in the organic rich phase, which grow to be space filling at phase separation with a fractal dimension $d_f = 2.5$, as expected for a percolation cluster in an infinite three dimensional system (55, 57). Such percolating networks have also been detected by neutron-scattering experiments in completely miscible aqueous solutions (58).

The space-filling network, which is identified in pure water at ambient conditions, even exists in supercritical water; the percolating line of demixing phase transitions is an extension of the boiling line (55). The close relation between demixing phase transition and percolation transition of physical clusters has also been used in simulations to localize the liquid-liquid transition region in supercooled water. The lowest density amorphous water phase (solid or liquid) has been characterized by the presence of a percolating network of well-ordered (ice-like tetrahedral), four-coordinated water molecules, whereas in high-density amorphous water phases, a percolating network of tetrahedrally bonded molecules is missing (54).

The formation of spanning H-bonded water networks on the surface of biomolecules has been connected with the widely accepted view that a certain amount of hydration water is necessary for the dynamics and function of proteins. Its percolative nature has been suggested first by Careri et al. (59) on the basis of proton conductivity measurements on lysozyme; this hypothesis was later supported by extensive computer simulations on the hydration of proteins like lysozyme and SNase, elastine like peptides, and DNA fragments (53). The extremely interesting question of protein hydration is a huge field by its own but beyond the scope of this article.

### Dynamics of Water Molecules

The molecular motion in water has been studied for decades with all available modern spectroscopic and scattering methods, including neutron scattering, nuclear magnetic and dielectric relaxation, infrared spectroscopy, and light scattering. Each applied method probes different aspects of the motional behavior on different length and time scales. A NMR furnishes diffusion coefficients and integrals over orientational correlation functions, quasielastic neutron scattering reveals information on the short-time translational and rotational motion. The results that were obtained for a wide range of temperature and pressure conditions have been interpreted in the frame of translational and rotational diffusion models. The temperature dependence of characteristic parameters like reorientation and residence times has been discussed in detail (12, 60–62). That water is so fluent is an apparent contradiction to the fact that the space-filling network of hydrogen bonds is made up of bonds that have an interaction energy strength well above the thermal energy $k_B T$. This puzzle was resolved by showing the importance of network defects: The presence of an excess (fifth) neighbor in the first-neighbor shell of water molecules allows the intermediate formation of bifurcated H-bonds, which provides a low-energy barrier path for reorientation and coupled translational motion (63–65). Consequently, a decrease of the local water density (which makes the presence of an excess neighbor less probable) decreases the mobility of water molecules. For example, this effect has been observed by NMR experiments in the hydration shell of convex hydrophobic particles, in which the molecular mobility is decreased (66). However, it does not decrease to such an extent that one could speak of “icebergs,” as this is still done occasionally (see the section entitled “Hydrophobic hydration and interaction”). In cold water, the increasing expansion of water reduces the mobility of the water molecules in addition to the pure thermal activation, which leads to a strong non-Arrhenius temperature dependence of reorientation times, diffusivity, and viscosity (20) (see below).

Implications of the existence of a liquid-liquid phase separation for the dynamic behavior of water have been discussed by Angell et al. (67, 68), who postulated a crossover from a so-called fragile to a strong glass-forming liquid behavior because of a transition into the region of the low-density liquid at deep supercooling. Possible mechanisms were discussed that dominate the molecular mobility in the different temperature ranges (69), which lead to different temperature dependences: At high temperatures, as mentioned above, the switching through bifurcated H-bonds is most effective and is connected with a low activation energy. At lower temperatures beyond the density maximum, a strong non-Arrhenius behavior with increasing apparent activation energy is produced by the development of a more perfect local order, which enforces an approach to structural arrest of the water molecules in the cages of their neighbors (70). This arrest is then overcome at even
lower temperatures by jump diffusion (71, 72); in other models by the collective relaxation of the cage of neighbors (73). Finally, in the locally well-ordered, low-density liquid, when approaching the glass transition, the formation of Frenkel-type defect pairs may enable a diffusion behavior that parallels the “strong glass former” Arrhenius line of the Angell plot (69).

The expected crossover could not be studied experimentally in pure water because of the onset of crystallization at strong supercooling. Recently, the possibility to supercool water to a much larger extent than bulk water, when it is confined to small pores, has been exploited (34). From dielectric spectroscopy and quasielastic neutron-scattering experiments on water confined in the nanopores of clays and silica glass, a transition (crossover) from a strongly activated non-Arrhenius motional behavior to a low activation energy Arrhenius line at even lower temperatures has been observed (32, 33). This observation correlates with the expected fragile-to-strong transition when crossing from “normal” to low-density water. Using such experiments in an extended pressure range, the position of the second critical point could be estimated for the confined water. This fragile-to-strong dynamic crossover was also observed for the hydration water of biomolecules (proteins and DNA) (74, 75). Most interestingly, this crossover occurs at the same temperature as the so called “protein glass transition,” which suggests that this transition in the dynamics of the protein is the result of the approach to the above-mentioned extension of the liquid-liquid equilibrium line of the solvent (the so-called Widom line of the second critical point of water). It has to be mentioned here that some controversy still surrounds the origin of this abrupt change in the temperature dependence of the mobility of the water molecules: This behavior has also been attributed to the limitation of the spatial extension of fluctuations in confinements (76).

Hydrophobic Hydration and Interaction

The “hydrophobic effect” is manifested thermodynamically by the low solubility (large positive solvation free energy) that nonpolar molecules or aggregates experience in water (for more extensive reviews, see References 77-79). The hydrophobic effect is of great relevance for a variety of phenomena, which includes protein folding as well as the structural organization of amphiphilic aggregates. The latter are forming micelles of various topology, as well as lyotropic mesophases and lipid membranes.

Surprisingly, the low solubility of small-sized particles does not stem from a weak interaction of particles with their surrounding water environment (77). For example, the heat of solvation of methane in water at ambient temperature is of similar magnitude as the heat of vaporization of pure liquid methane (80). The positive solvation free energy of small apolar particles at low temperatures is the consequence of negative solvation entropy, which overcompensates for the negative solvation enthalpy. It is widely believed that this “entropy penalty” is caused by the orientation order introduced to the hydration shell of water molecules as they try to maintain an intact hydrogen bond network (77). Parallel to the entropy decrease observed for low temperatures, theoretical and experimental studies also indicate a slowing down of the translational and reorientational dynamics of water in the hydration shell of an apolar moiety (66, 82-83). A rather thermodynamic signature of hydrophobic hydration is the large positive solvation heat capacity. The heat capacity increase is attributed to the temperature-induced mutual interactions among the solvent molecules in the hydration shell (84). It is considered to be caused by the progressing disintegration of the hydrogen bond network around the solute with increasing temperature (77, 84). Because the solvation of small apolar molecules is accompanied by an entropy decrease of the solvent, the formation of contact pairs of apolar particles is a way to reduce this “entropy penalty.” The tendency to form apolar contact pairs in solution is termed “hydrophobic interaction” and essentially controlled by the solvent. Because the association of small apolar particles is entropically favorable, a temperature increase leads to more stable apolar contacts. “Hydrophobic interaction” is a classic example of an “entropic force.”

Contrasting the behavior close to small apolar solutes, water behaves differently at an extended (planar) interface. Here, the thermodynamic features are mostly governed by water’s interfacial tension, which is essentially entropic in nature (weakening with increasing temperature). Consequently, at some length-scale a “crossover” has to occur (85, 86) from an entropy to an enthalpy dominated solvation behavior. Recent studies indicate that this transition appears at a length-scale significantly below 1 nm (87, 88).

The thermodynamic signatures of small apolar particle hydration can be modeled by simple two-state models (89-92) that solely focus on water’s hydrogen bonding as supposedly dominating effect. Stronger hydrogen bonds close to an apolar particle are counterbalanced by fewer possible hydrogen bonds. Silverstein et al. (92) consistently related experimental data that described water’s hydrogen bond equilibrium with hydrophobic solvation calorimetric data. Their calculations suggest that at lower temperatures, the hydrogen bonds are more intact than in the bulk, whereas at high temperatures, hydrogen bonds are more broken. The model moderately readsout older theories by Franks and Evans in their so called “iceberg” model (93), in which the hydrophobic particles where thought to be stabilizing structured ice-like entities in water. However, because the entropy change experienced by a water molecule in a hydrophobic hydration shell is about five times smaller than a crystal-like environment (77), the “iceberg” model too strongly exaggerates the degree of ordering that is present in a hydrophobic hydration shell (77). Simple two-state models seem to fail in reliably predicting absolute solvation free energies (77) because altering hydrogen bonding does not provide sufficient information to determine the entropic “free volume contribution” (94, 95).

A conceptually complementary approach to describe hydrophobic effects has been introduced by Pratt and coworkers (78, 96). Their information theory (IT) model is based on an application of Widom’s potential distribution theorem (97) combined with the perception that the solvation free energy of a small hard sphere, which is essentially governed by the probability to find an empty sphere, can be expressed as a limit of the distribution of water molecules in a cavity of the size...
of the hydrophobic particles. Because the distribution functions are essentially determined by density fluctuations of water at the molecular scale, the IT model relates the hydration and interaction of hydrophobic particles with the temperature dependence of water's thermodynamic response functions, such as expansivity and compressibility.

An instructive, simplified computer model for water is the "Mersenne–Benz" (MB) model of Silverstein et al. (98). It has been shown to capture qualitatively the anomalous thermodynamic behavior of water as well as the thermodynamic features of the hydrophobic effect. The MB-model notably captures the effect of increasing particle size. A hydrophobic particle of about twice the size of a water molecule is found to increase the free energy by a different mechanism, namely by increasingly breaking the hydrogen bonds. Similar to the thermodynamic behavior observed for planar interfaces, the mechanism increases the enthalpy but has only little effect on the entropy and heat capacity. The simulations indicate that at a large, extended interfacial surface, water is geometrically unable to form its maximal number of hydrogen bonds to other water molecules. Thus, enthalpically costly “dangling” hydrogen bonds form pointing toward the interface (99, 100).

Realistic three-dimensional computer models for water were proposed already more than 30 years ago (14). However, even relatively simple effective water model potentials based on point charges and Lennard-Jones interactions are still very expensive computationally. Significant progress with respect to the models ability to describe water’s thermodynamic, structural, and dynamic features accurately has been achieved recently (101–103). However, early studies have shown that water models essentially capture the effects of hydrophobic hydration and interaction on a near quantitative level (81, 82, 104). Recent simulations suggest that the exact size of the solvation entropy of hydrophobic particles is related to the ability of the water models to account for water’s thermodynamic anomalous behavior (105–108). Because the “hydrophobic interaction” is inherently a multibody interaction (105), it has been suggested to compute pair- and higher-order contributions from realistic computer simulations. However, currently it is inconclusive whether three-body effects are cooperative or anticooperative (109).

An analysis of computer simulations of water at different pressures by Hummer et al. (110) suggested that hydrophobic contact pairs become increasingly destabilized with increasing pressure. The proposed scenario could explain the pressure denaturation of proteins as a swelling in terms of water molecules that enter the hydrophobic core by creating water-separated hydrophobic contacts. Additional support for the validity of Hummer’s IT-model analysis has been achieved by pressure-dependent computer simulation studies of isolated pairs of hydrophobic particles, as well as rather concentrated solutions of hydrophobic particles (111, 112). Recently, the pressure-induced swelling of a polymer composed of apolar particles at low temperatures can be observed (113).

Ion Hydration

As already suggested by Max Born in 1920 (114), the large negative solvation free energies of ions in aqueous solution can be explained by mostly purely electrostatic effects, which assume water to be represented by a dielectric continuum (115). Small changes of the ion diameter are found to affect the solvation free energies strongly. The measured solvation free energies roughly scale with the third power of the ion diameter, as proposed by the Born theory (115). However, a structuring effect on the first hydration shell water is obvious and has been experimentally determined by X-ray and neutron scattering techniques (116–119). In case of an anion, the first shell water molecules form a hydrogen bond type configuration in which the OH-bond points toward the anion (118), although on average it does not point exactly to the center of the ion (117). For the case of cations, the water molecules are found to be pointing with their oxygen toward the ion. The water dipole axis, however, seems to exhibit an average tilt of about 30 degrees with respect to the ion-water-oxygen connecting vector (117, 119).

Recent first principles simulations of aqueous salt solutions suggest that this might be an artifact caused by averaging a rather broad tilt angle distribution (120). In those simulations, the dipole vector that points directly toward the ion is the most likely configuration of a broad distribution. Earlier classic MD simulations had revealed a more filled “cone-pair”-type of ion-water bonding (121), which was possibly a consequence of the tetrahedral charge distribution of the employed water model (17) and is perhaps an artifact.

Ionic Influence on Hydration Water and Bulk Water

Salts are known to influence several properties of aqueous solutions in a systematic way (122, 123). The effect of different anions and cations seems to be ordered in a sequence; this theory was already proposed by Hofmeister in 1888 (124) from a series of experiments on the salts ability to precipitate “hen-egg white protein.” Numerous other properties of aqueous salt solutions are also found to be systematically salt dependent, such as the surface tension or the surface potential (122). However, the exact reason for the observed specific cation and anion sequences is still not fully understood (125). Model calculations (126), as well as nuclear magnetic relaxation experiments (127), propose a delicate balance between ion adsorption and exclusion at the solute interface. This balance is tuned by the solvent (water) structure modification according to the ion hydration (128, 129) and hence is possibly subject to molecular details.

In principle, two different mechanisms have been proposed on how the ions influence protein stability. Firstly, it has been suggested that a modification of water’s structure is the origin of the Hofmeister sequence (130). It has been hypothesized that some ions “kosmotropes” enhance the structure that surrounds the ions, which leads to a strengthening of the hydrophobic effect and thereby stabilizes the proteins (131). However, the ions that break the structure that surrounds the ions (“chaotropic”)
The evolutionary process. As we have tried to show, many of the properties, many of which have been exploited by nature during liquid water provides a unique wealth of unusual liquid/solvent properties, many of which have been exploited by nature during the evolutionary process. As we have tried to show, many of water’s unusual properties, or “anomalies,” stem from water’s tendency to form a roughly four-coordinated hydrogen bond network. Computer simulations indicate that many effects can be explained in great detail by Scraton-Yates ion molecular models. A pattern reveals that in water, hydrogen bond-based local-order (entropy) and interaction energies have the beneficial tendency to compensate each other, which is important for many solvation processes and for water structuring and phase behavior. The small size of the water molecule and its effective hydrogen-bond formation make it a particularly helpful agent in the process of protein folding; it is deemed essential to enable protein motions.

Conclusion

Liquid water provides a unique wealth of unusual liquid/solvent properties, many of which have been exploited by nature during the evolutionary process. As we have tried to show, many of water’s unusual properties, or “anomalies,” stem from water’s tendency to form a roughly four-coordinated hydrogen bond network. Computer simulations indicate that many effects can be explained in great detail by Scraton-Yates ion molecular models. A pattern reveals that in water, hydrogen bond-based local-order (entropy) and interaction energies have the beneficial tendency to compensate each other, which is important for many solvation processes and for water structuring and phase behavior. The small size of the water molecule and its effective hydrogen-bond formation make it a particularly helpful agent in the process of protein folding; it is deemed essential to enable protein motions.

References


Water Properties of...


See Also

Hydration Forces: Water and Biomolecules
Water and Protein Folding
Computational Chemistry in Biology
Computational Modelling of Protein Folding
Several pathways are compartmentalized in the endoplasmic reticulum (ER). These intraluminal activities require the passage of substrates, cofactors, and products through the ER membrane. The arguments for a general permeability of the ER membrane are contradicted by strong biochemical, pharmacological, clinical, and genetic evidence, which indicates that the lipid bilayer has a barrier function and that specific transport activities are needed in the membrane. Consequently, the ER lumen can be regarded as a separate metabolic compartment. This article overviews the best characterized intraluminal processes in which the compartmentation is important either by defining an intraluminal milieu, by limiting the rate of the reaction, by determining the specificity, or by creating a common substrate pool because of the colocalization.

The best known functions of the ER require a high membrane surface and/or a separate, specific microenvironment within the organelle. Although many enzymes hosted by the ER use its membranous structure only as a scaffold, others are compartmentalized within the ER: i.e., their active site is localized in the lumen. The activity of these enzymes usually is dependent on the special composition of the luminal compartment. The enzymes often receive their substrates and cofactors from or release their products to the cytosol, therefore, the transport of these compounds across the ER membrane is indispensable. This article focuses on this latter group of the ER enzymes, the functioning of which makes the ER a separate metabolic compartment of the eukaryotic cell.

Biologic Background

The ER is a continuous network of membranous tubular and lamellar structures in the cytosol (1, 2). The membranes that build up the organelle can constitute more than 95% of the total cellular membranes, and the total volume of the organelle can compose about 10% of cell volume, e.g., in hepatocytes. Although the ER is a single, spatially continuous compartment, it can be divided structurally and functionally into different subdomains. The ER forms contact sites beside the nuclear envelope with practically all the other organelles and the plasma membrane; these functional regions might have specific composition and features, but they are restructured continuously to adapt the actual cellular requirements. The ER network can be divided into rough (RER) and smooth (SER) domains on the basis of the presence or absence of ribosomes on the outer surface of the membrane, respectively. Functional and morphological differences are linked: Whereas the ER is responsible for the synthesis, posttranslational modification, and folding of secretory and membrane proteins, the SER is responsible for lipid biosynthesis, biotransformation, and the production of small molecules to be secreted.

The ER hosts several enzymes: Some of them are facing toward the cytosol and use the ER membrane as a scaffold, e.g., cytochrome P450 isozymes, NADPH:cytochrome P450 reductase, or hydroxymethyl-glutaryl-CoA reductase. Others are compartmentalized in the ER: that is, their active center is localized in the lumen. The reactions catalyzed by these enzymes belong to various pathways of biochemistry and cell biology, such as carbohydrate and lipid metabolism, biotransformation, signaling, steroid metabolism, and protein processing. Table 1 presents a list of some important intraluminal enzymes. Usually, because their substrates derive from the cytosol, the corresponding membrane transport processes are required for their action.
Table 1. Some important activities compartmentalized within the ER lumen

<table>
<thead>
<tr>
<th>Pathway/process</th>
<th>Enzyme/protein</th>
<th>Role</th>
<th>Transport involved</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>carbohydrate metabolism</td>
<td>glucose-6-phosphatase</td>
<td>glucose production (last common step of glycogenolysis and gluconeogenesis)</td>
<td>G6P, P1, glucose</td>
<td>8,14,24-27,31-33,39, 41,42,55</td>
</tr>
<tr>
<td></td>
<td>hexose-6-phosphate dehydrogenase</td>
<td>NADPH generation for reductases; antioxidant defense; base synthesis?</td>
<td>G6P</td>
<td>15,16,18,34, 43,44</td>
</tr>
<tr>
<td>posttranslational processing of secretory proteins</td>
<td>various</td>
<td>folding and posttranslational modification</td>
<td>translocan protein channel</td>
<td>4-7,51-53</td>
</tr>
<tr>
<td>oxidative protein folding</td>
<td>ERO1p (ER oxidoreductin)</td>
<td>electron transfer from PDI to FAD</td>
<td>FAD</td>
<td>5,13</td>
</tr>
<tr>
<td></td>
<td>PDI</td>
<td>disulfide bond formation and isomerization, folding dehydroascorbate reduction</td>
<td>dehydroascorbates, GSH</td>
<td>10,11,22,23, 44,46,54</td>
</tr>
<tr>
<td>posttranslational modification</td>
<td>prolyl-3 and 4-hydroxylases, lysyl hydroxylase</td>
<td>hydroxylation</td>
<td>(dehydro)ascorbate</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>UDP-glucose glucosyltransferase</td>
<td>quality control of secretory proteins</td>
<td>UDP-glucose</td>
<td>61</td>
</tr>
<tr>
<td>biotransformation</td>
<td>UDP-glucuronosyl-transferase isozymes</td>
<td>glucuronidation of endo- and xenobiotics</td>
<td>UDP-glucuronate acid glucuronidases</td>
<td>19,20,21,47, 48</td>
</tr>
<tr>
<td></td>
<td>β-glucuronidase</td>
<td>deglucuronidation</td>
<td>glucuronidases</td>
<td>28,49,50</td>
</tr>
<tr>
<td></td>
<td>carboxyl reductase</td>
<td>reduction of xenogenous aldehydes and ketones (coupled with the G6PT-H6PDH system)</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>steroid metabolism</td>
<td>11β-hydroxysteroid dehydrogenase type 1</td>
<td>prereceptorial activation of glucocorticoids (coupled with the G6PT-H6PDH system)</td>
<td>15,16,18,34, 37,48,44</td>
<td></td>
</tr>
<tr>
<td>lipid metabolism</td>
<td>diacylglycerol acyl transferase</td>
<td>diacylglycerol synthesis for VLDL</td>
<td>acylcarnitine carnitine</td>
<td>63</td>
</tr>
<tr>
<td>antigen presentation</td>
<td>MHC I class</td>
<td>oligopeptide binding</td>
<td>TAP</td>
<td>38,39</td>
</tr>
</tbody>
</table>
Arguments for and against the general permeability of the ER membrane

It has been questioned whether the ER lumen can be considered as a separate metabolic compartment or whether the lumen is just a specific localization for some enzymatic reactions. The inward and outward traffic of a wide variety of molecules led some scientists to the conclusion that the ER membrane is permeable to any low-molecular-weight compounds and acts only as a molecular sieve to keep the luminal proteins together. Several arguments can support this hypothesis. The composition of the ER membrane is different from that of the plasma membrane; it has a low cholesterol and high protein content. The less-ordered membrane structure might increase leakiness for small molecules (3). The RER is abundant in translocon protein channels, which can allow the passage of ions and small-molecular-weight molecules (4–6). Other transiently open transmembrane channels might also allow nonselective fluxes. Moreover, a high number of structurally unrelated molecules, i.e., xenobiotics or bioisosteres, have been shown to enter the ER lumen. The total enzymatic activity can be revealed on the intact ER membrane, which excludes the substrates from the lumen. The characteristic micro-environment of the ER lumen hardly is just as important a determinant of the ER metabolism as the pattern of the compounds that cannot enter or leave the lumen.

In vivo evidences

Recent in vivo observations support the hypothesis that the ER membrane has a barrier function. The two best studied gradients across the ER membrane are related to the calcium homeostasis and the oxidative protein folding. Compared with the cytosol, the ER lumen is characterized by a magnitude level of free calcium ion (29) that is four orders higher and nearly a hundred times lower than the ratio of glutathione (GSH) and oxidized glutathione disulfide (GSSG) (30). Although the generation of these gradients is dependent on largely the poor permeability of the ER membrane to calcium ion and glutathione, the gradients necessarily are created and maintained by continuous active processes, such as the pumping of calcium (29) and the luminal oxidation of thiols (12), respectively. Therefore, it can be also concluded that high-capacity active processes surpass the velocity of the passive transmembrane fluxes.

The strongest evidence for the separation of the cytosolic and endoplasmic compartments is provided by the genetic analyses of two ER-related human syndromes, namely glycogen storage disease type 1 (GSD 1) and cortisone reductase deficiency (CRD). GSD 1 refers to the congenital deficiency of glucose 6-phosphatase (G6Pase) activity, which causes a complex metabolic disorder, including the abnormal storage of glycogen in the liver (31). In addition to the defects of G6Pase enzyme (GSD 1a), the mutations in a different gene were proven to cause similar metabolic disturbances combined with some additional symptoms (32) — a disease called GSD 1b. The protein encoded by this other gene turned out to be the G6PT, which is needed to access the luminal G6Pase to its substrate (8, 24, 32). GSD 1b pathology was mimicked fully by G6PT knockout mice (33). The existence of GSD 1b proves that glucose 6-phosphate (G6P) cannot enter the ER lumen unless it is mediated by a specific transporter.

CRD is characterized by the insufficient reduction of cortisone to cortol by 11β-hydroxysteroid dehydrogenase type 1 (11βHSD1) in the ER lumen. The analyses of the mutation suggest that the CRD phenotype is caused by the combined defects of 11βHSD1 and hexose 6-phosphate dehydrogenase (H6PDH) (34). This finding indicates that the luminal generation of NADPH by H6PDH (rather than the cytosolic NADPH production) drives cortisone reduction in physiologic conditions. In other words, luminal cortisone reduction is dependent on local NADPH generation because the cytosolic and endoplasmic pyridine nucleotide pools are separated. Although recent investigations revealed that the combined mutations do not necessarily cause CRD (35, 36), the importance of the collaboration between the two enzymes was validated additionally by the results obtained in H6PDH knockout mice (37).

Another proof against the general permeability of the ER membrane and for the necessity of specific transporters is derived from the field of immunology. TAP peptide transporters (TAP1 and TAP2), which were characterized at the molecular level, belong to the ABC transporter superfamily (38). They translocate oligopeptides (approximate length of 8–16 amino acid residues) into the ER lumen and the target compartment is needed for the transport of the presented peptides to the pathway of antigen presentation. A recent study revealed that the particular function of TAP is the presentation of peptides with a high density of MHC class I molecules (39). Thus, the luminal environment of the ER lumen could be maintained if the passive diffusion driven by the gradients equalized effectively the concentrations on both sides of the membrane.
acid residues) produced by the proteasome in the cytosol to the ER lumen, where they bind to the major histocompatibility complex (MHC) Class I molecules. MHC–peptide complexes leave the ER by vesicular transport to the cell surface for recognition by cytotoxic T lymphocytes. The loss of the TAP function leads to the impairment of antigen presentation, as it is observed commonly in tumors and virus-infected cells that escape immune surveillance (39). This condition shows clearly that the basal permeability of the ER membrane does not permit the appearance of these small oligopeptides in the lumen.

Chemistry

In this section, the elucidated role of the ER membrane (as a barrier with selective transport) is summarized in the best studied metabolic systems of the ER.

Glucose 6-phosphatase system

The means of glucose production in liver and kidney is derived either from hydrolysis of G6P glycogen breakdown or gluconeogenesis. The reaction is catalyzed by G6Pase, which is an integral membrane protein of the ER with an intraluminal active center (8). Theoretically, the system composes three transporters (for G6P, glucose, and phosphate) associated with the enzyme activity functionally (8). G6PT has been identified at the molecular level (24, 32), and the protein-mediated glucose transport across the ER membrane has been characterized recently (40). The intact membrane barrier is an important determinant of the physiologic characteristics of G6Pase. Permeabilization of microsomal vesicles or in situ ER membranes (41, 42) only doubles the rate of G6P hydrolysis, and it causes a 1015-fold enhancement of the hydrolysis of mannose 6-phosphate and various other sugar phosphates by the same enzyme (14). It is important that the ER membrane separates the unselective G6Pase from intermediates of glycolysis/gluconeogenesis and amino sugar metabolism. However, the high substrate specificity of G6Pase in physiologic conditions is not an intrinsic property of the enzyme but relies on the selective transport of G6P into the lumen of ER.

GSD 1a (G6Pase deficiency) and GSD 1b (G6PT deficiency) cause the same metabolic derangements based on insufficient G6Pase activity (31, 33). This fact is inconsistent evidence with the theory of an unspecifically permeable ER membrane.

Hexose 6-phosphate dehydrogenase and 11β-hydroxysteroid dehydrogenase type 1

Interconversion of cortisone and cortisol catalyzed by 11βHSD1 is fully reversible in vitro. The physiologic direction in vivo is cortisone reduction, which is driven by a high NADPH/NADP+ ratio in the ER lumen. The high ratio can be maintained only by local NADP+ reduction catalyzed by certain luminal dehydrogenases, such as the H6PDH (15, 16, 24), because the permeability of the ER membrane to pyridine nucleotides is negligible (17, 18).

It has been reported that extravascular NADP+ and NADPH can penetrate the ER membrane in long incubations and that the cytosolic [NADPH]/[NADP+] ratio can affect 11βHSD1 activity (43). Nevertheless, the H6PDH knockout mice lack 11βHSD1-mediated glucocorticoid generation (37), which proves that the enzyme cannot rely on cytosolic NADPH resources and that a separate luminal pyridine nucleotide pool exists. It also shows that clearly the high luminal [NADPH]/[NADP+] ratio is dependent on H6PDH activity.

G6P is transported to H6PDH from the cytosol across the ER membrane. In fact, similarly to G6Pase, the substrate specificity of H6PDH is dependent on largely its localization in the ER. The enzyme has dehydrogenase activity on various hexose 6-phosphates, such as G6P, galactose 6-phosphate, or 2-deoxyglucose 6-phosphate and on simple sugars such as glucose and, it has dual nucleotide specificity for NADP+ and NAD+ (44). Nevertheless, under physiologic conditions in the ER lumen, the native substrates for H6PDH are believed to be G6P and NADP+ (44). These findings indicate that the enzyme has no access to the cytosolic NAD+, sugars, and sugar phosphates, except G6P, which is transported by the specific G6PT (8).

Oxidative protein folding and antioxidant metabolism

Luminal proteins and luminal domains of membrane proteins of the ER contain remarkably more disulfide bridges and less thiol groups than the cytosolic proteins. In accordance with this low [protein thiol]/[protein disulfide] ratio, the ER lumen shows also a characteristically low (about 1:1) [GSH]/[GSSG] ratio, which is nearly 100:1 in the cytosol (30). Such a high potential difference and concentration gradient would be expected if the membrane was permeable to both GSH and GSSG. The finding that indicates that isolated hepatic microsomes still contain GSH and GSSG (22) suggested strongly that the ER membrane represents a barrier for these molecules. Indeed, the membrane is impermeable to GSSG, whereas GSH has a slow protein-mediated transport (23). Therefore, the oxidizing environment in the compartment can be maintained by local oxidation of GSH that yields GSSG, which is entrapped in the lumen.

Luminal thiol oxidation is facilitated by ascorbate (vitamin C) (45) or FAD (12, 13), so the physiologic role of their transport has been proposed. ER membrane is permeable selectively to dehydroascorbate, the oxidized form of ascorbate (10, 11). Luminal reduction of dehydroascorbate to ascorbate is associated with thiol oxidation and leads to ascorbate entrapment (46). FAD uptake and a consequent thiol oxidation have also been found in yeast and in liver microsomes (12, 13). In contrast to FAD, pyridine nucleotides—NADP(H), NAD(H)—of similar size and structure cannot enter the ER lumen at a significant rate, which is indicated by the high latency of intraluminal H6PDH and 11βHSD1 (15) and by direct transport measurements (18).

Glucuronidation

The quantitatively most significant second-phase reaction of hepatic biotransformation is glucuronidation that takes place in the ER. The transfer of glucuronosyl group from glucuronic acid residues produced by the proteasome in the cytosol to the ER lumen, where they bind to the major histocompatibility complex (MHC) Class I molecules. MHC–peptide complexes leave the ER by vesicular transport to the cell surface for recognition by cytotoxic T lymphocytes. The loss of the TAP function leads to the impairment of antigen presentation, as it is observed commonly in tumors and virus-infected cells that escape immune surveillance (39). This condition shows clearly that the basal permeability of the ER membrane does not permit the appearance of these small oligopeptides in the lumen.
UDP-glucuronate (UGA) to appropriate functional groups of the substrates is catalyzed by UDP-glucuronosyltransferases (UGTs). These enzymes are integral membrane proteins of the ER with their active center localized in the lumen (47). UGA is synthesized in the cytosol, and the produced glucuronides are pumped out of the cell by plasma membrane transporters. Thus, conjugation with glucuronate requires UGA import (19) and glucuronide export across the ER membrane. It has been demonstrated, using a photoaffinity-labeling technique, that UGA, but not UDP-glucose, has access to the active center of UGTs in intact microsomal vesicles (48). The high latency (more than 90%) of UGTs observed in both microsomal vesicles and isolated permeabilized hepatocytes indicates that the transport processes (presumably UGA uptake) are rate limiting (20, 21). Activity of the luminal β-glucuronidase, a glucuronide-cleaving enzyme, is also limited by substrate (glucuronide) transport, although it has moderate (approximately 40%) latency (49).

A protein-mediated glucuronide transport across the ER membrane has been demonstrated (50), and a competition for the transport has been found between glucuronides of similar size (28). The pattern of interactions suggested the presence of multiple glucuronide transporters with overlapping specificities in the ER membrane (28).

Translocon and permeability of (R)ER membrane

A bundle of channels may contribute theoretically to a general permeability or leakiness of the ER membrane, which has been demonstrated in case of the ribosome-bound translocon complex, which cotranslationally imports nascent peptides into the ER lumen (51). The average diameter of the translocon tunnel is approximately 20 Å, which is wide enough to allow the transport of Ca²⁺ (52) and small water-soluble molecules (4–6). Therefore, the presence of translocon complexes might be responsible for the higher permeability of the RER membrane versus the SER membrane.

Unspecific permeability is prevented in the translationally active translocon because the tunnel is occupied by the peptide chain being polymerized. Similarly, the pore is blocked by BiP proteins, prominent intraluminal chaperones, after dissociation of the ribosome from the complex (53). In fact, it has been argued that the BiP locks form a smaller barrier for uncharged polar molecules than for charged ones. Furthermore, the dissociation of ribosomes from translocon complexes is delayed after the termination of protein synthesis. When a nontranslating ribosome is associated with the translocon complex, they form a transitional low-selectivity channel between the cytosol and the ER lumen (5).

Chemical Tools and Techniques

**In vivo studies**

Measurement of the concentrations in the cytosolic and luminal compartments in vivo has been achieved in the case of calcium ions but remains a merely theoretical possibility for most organic molecules except glutathione (54). In fact, a remarkable transmembrane gradient has been detected in case of both Ca²⁺ and glutathione, which supports the barrier function of the ER membrane. The maintenance of high concentration differences would require continuously intensive pump activities and lead to unreasonable heat generation if the membrane is highly permeable.

**Studies in cellular systems**

The potential role of transport across the ER membrane was investigated in cells by using selective and general membrane permeabilizing toxins, antibiotics, or detergents. Intact ER can be exposed to the incubation medium in cells whose plasma membrane has been permeabilized selectively. Alternatively, the intraluminal enzymes can be directly exposed by general (including plasma and ER membranes) permeabilization. Experiments using these models revealed that the transport of substrates or cofactors across the ER membrane is rate limiting for glucosidase (42), UGT (20), and β-glucuronidase (49) activities.

The transport of exogenous compounds has been investigated recently by determining the amount of modified intraluminal proteins upon addition of nonphysiologic biotin-derivatives capable of chemical protein modification (7). The ER membrane has been shown to be permeable to three different biotin-derivatives, and it has been concluded that it does not form a barrier to small molecules; that is, either the lipid bilayer or the integral membrane proteins allow their diffusion unspecifically.

**Studies in subcellular systems**

Isolated ER-derived vesicles (microsomes) maintain their original orientation (i.e., their intravesicular surface corresponds to the intraluminal side of the ER membrane (55, 56)) and keep some of their low-molecular-weight luminal components during the long-lasting preparation procedure. For example, both oxidized and reduced forms of glutathione (22) and of pyridine nucleotides (17) are present in liver microsomes, which suggests strongly that the ER membrane acts as a barrier to these molecules.

Microsomes are used widely for the determination of "isotope space" and for in vitro transport assays, such as "rapid filtration" and "light scattering" measurements. Isotope space reveal the distribution of the studied compound between intra- and extravesicular water compartments; hence, the measurements provide information on the permeability but not on the transport kinetics. The total water space and the extravesicular space can be measured using radiolabeled analog of water or a completely impermeable compound, respectively (57). Rapid filtration is based on the quick separation of vesicle-associated molecules from the medium and on the calculation of the intraluminal content after various incubation periods. This method allows the application of physiologic concentrations but offers a limited time resolution (58). On the contrary, the light scattering technique requires high concentrations but provides a real-time monitoring of traffic. This indirect detection takes advantage of the transport-associated osmotic shrinkage and the swelling of the vesicles.
microsomal vesicles (58). The permeability to several molecules of a wide size-range has been compared with these methods, and very slow permeation, or complete impermeability, has been demonstrated in certain cases.

A sophisticated approach using active site-directed photoaffinity substrate analogs has also been applied to study the inward transport in isolated microsomal vesicles. These experiments provided convincing and elegant evidence for the translocation of UDP-glucuronate (48) and of FAD (12) by photo incorporation of the probes into the luminally oriented enzymes in intact microsomes. The drawback of the method is that it is not suitable to determine of the rate or capacity of transport.

Enzyme latency is an experimental manifestation of compartmentation, which means that the activity of certain intraluminal enzymes is increased remarkably when the membrane is permeabilized either by detergents or by channel-forming antibiotics (e.g., amphotericin). It is based on the rate-limiting transport of substrates and/or effectors across the intact ER membrane. Some activities in the ER are more than 90% latent; i.e., they increase more than 10-fold during permeabilization (20, 21).

Conclusion

The ER membrane represents a real barrier between the cytosol and the lumen for water-soluble, charged small molecules. Although this barrier function is compromised at a certain extent by the existence of nonspecific pores (e.g., the translocon protein channel), the ER lumen can be considered still as a separate metabolic compartment. This barrier generates a characteristic microenvironment in the lumen with a higher Ca²⁺ concentration than the cytosol. This barrier function is maintained, which means that the activity of certain intraluminal enzymes is increased remarkably when the membrane is permeabilized either by detergents or by channel-forming antibiotics (e.g., amphotericin). It is based on the rate-limiting transport of substrates and/or effectors across the intact ER membrane. Some activities in the ER are more than 90% latent; i.e., they increase more than 10-fold during permeabilization (20, 21).

References


Further Reading


See Also

Compartment Metabolic Diseases, Chemical Biology of Metabolism, Cellular Organization of Oxidative Post-Translational Modifications Small Molecule Transport
Cellular Organization of Metabolism

Peter D. Karp, SRI International, Menlo Park, California
doi: 10.1002/9780470048672.wecb321

The field of pathway bioinformatics is concerned with the representation and the manipulation of metabolic information within computers. By organizing genome information into pathways, pathway bioinformatics places genes and their products into a mechanistic framework. This article describes how metabolic pathways are represented in a computer, and it describes the BioCyc (SRI International, Menlo Park, CA) collection of pathway/genome databases for several hundred organisms. Each BioCyc database describes the genome and the metabolic network of a single organism. This article describes computational algorithms for computing with pathway data. Pathway visualization algorithms help scientists comprehend this complex information space and facilitate analysis of large-scale omics datasets. Pathway analysis algorithms predict the metabolic network of an organism from its genome, identify the genes coding for missing enzymes in metabolic pathways, and enable the comparison of metabolic networks from multiple organisms. They also allow the prediction of the metabolic capabilities of an organism and identify potential drug targets within the metabolic network.

Together, Pathway Tools and BioCyc permit extremely fast and accurate modeling of the metabolic network of an organism from its genome sequence. Previously, hundreds of person-years of laboratory work were required to characterize an organism’s metabolic map. Now, given an annotated genome sequence for the organism, its metabolic network can be predicted computationally within a few days. Manual review of that computational prediction will yield a more accurate result within a few weeks. Computational reconstructions of metabolism are not perfectly accurate; thus, for increased model fidelity, we recommend following the computational reconstruction with manual curation of the metabolic network. A manual curation effort surveys the past biochemical literature for an organism, tracks newly emerging literature on an ongoing basis, and updates the metabolic model within a PGDB to reflect those findings.

Introduction

The bioinformatics subfield of pathway bioinformatics is concerned with developing computer representations of the metabolic network of an organism, with developing databases of metabolic information, and with developing algorithms for computing with metabolic information. This article will discuss the approaches for each of these problems in the Pathway Tools and BioCyc (SRI International, Menlo Park, CA) projects. Pathway Tools is a set of algorithms for computational analysis of metabolic data, and it includes computer representations of the metabolic network (1). BioCyc is a collection of Pathway/Genome Databases (PGDBs) for several hundred organisms (2). The BioCyc databases were constructed using Pathway Tools, and they can be queried and analyzed using Pathway Tools.

Representing Metabolic Knowledge in PGDBS

Metabolites, reactions, and pathways

Two alternative ways exist in which one might choose to represent the metabolic network in a computer: by listing of all metabolic reactions that occur in the cell or by partitioning that reaction set into a carefully delineated set of metabolic
Database relationships and attributes

The previous two subsections described the important types of objects in a PGDB. Here, we describe how these objects are linked together by biologically meaningful database relationships. PGDB relationships knit together the objects in a PGDB by defining how these objects are related. For example, queries can follow the relationship from a gene to the protein that it codes for, from a protein to a reaction that it catalyzes, and from a reaction to a metabolic pathway in which it is a component. To answer questions such as “find all metabolic pathways in which the products of a gene play a role.”

Every PGDB object has a stable unique identifier (ID), that is, a symbol that identifies that object uniquely within the PGDB. Example unique IDs include TRP (an identifier for a metabolite), RXN0-2382 (an identifier for a reaction), and PWY0-1280 (an identifier for a pathway). Relationships within a PGDB are implemented using IDs. For example, to state that the TRP (tryptophan) object is a reactant in the reaction RXN0-2382, a field of RXN0-2382 called LEFT (meaning “reactants”) contains the value TRP. Many PGDB relationships exist in both forward and backward directions; for example, the TRP object contains a field called APPEARS-IN-LEFT-SIDE-OF that lists all reactions in which TRP is a reactant. The fields LEFT and APPEARS-IN-LEFT-SIDE-OF are called inverses.

Figure 1 shows the relationships that link together levels 1–3 of the metabolic network representation. Note how inverse fields allow the user to query relationships in any direction within a PGDB; for example, given a metabolite we can query for the pathways in which it is involved, and given a pathway we can query for its metabolites.

Figure 2 shows the relationships that link together the genome and the proteome. For clarity, many related objects are omitted from Figs. 1 and 2, such as the many other Escherichia coli genes that are components of its chromosome, the other reactions that are components of TRYPSYN-PWY, and the reactants and products of RXN0-2382.

The representations in Figs. 1 and 2 must be connected because enzymes in the proteome catalyze reactions in the reactome. Thus, PGDBs contain relationships that link enzymes with the reactions they catalyze. However, these relationships are indirect, passing first through an intermediary object called an enzymatic reaction, as shown in Fig. 3. This arrangement allows us to capture the many-to-many relationship that exists between enzymes and reactions— one reaction can be catalyzed by multiple enzymes, and multifunctional enzymes catalyze multiple reactions. The purpose of the enzymatic reaction is to encode information that is specific to the pairing of the enzyme with the reaction, such as cofactors, activators, and inhibitors. Consider a bifunctional enzyme with two active sites, in which one of the active sites is inhibited by pyruvate, and the second active site is inhibited by lactate. We would represent this situation with two enzymatic reactions that link the enzyme to the two reactions it catalyzes, and each enzymatic reaction would specify a different inhibitor.
Cellular Organization of Metabolism

Figure 1 Relationships that link levels 1–3 of the metabolic network. The metabolite tryptophan (ID TRP) is a reactant of the reaction whose ID is RXN0-2382, which in turn is a member of the pathway whose ID is TRYPSYN-PWY. The field IN-PATHWAY is the inverse of the field REACTION-LIST.

Figure 2 Relationships that link the genome and proteome of a PGDB. The E. coli chromosome contains thousands of genes, one of which is EG11025 (trpB). Its product is the TrpB protein, whose ID is TRYPSYN-BPROTEIN. That monomer forms a homomultimer represented by the object CPLX0-2401.

Metabolic Databases

The preceding conceptual structure underlies all PGDBs created by Pathway Tools. Those PGDBs fall into several categories. The BioCyc collection of PGDBs is a collaboration between the Bioinformatics Research Group at SRI International and the Computational Genomics Group at the European Bioinformatics Institute (2). In addition, many other PGDBs have been created by other users of Pathway Tools. Some are listed in a table on the BioCyc home page (http://BioCyc.org). These PGDBs can be accessed through the web sites operated by their creators, and in some cases they are available through the BioCyc web site. In addition, some PGDBs can be downloaded for local use within Pathway Tools through SRI’s online registry of PGDBs (http://BioCyc.org/registry.html).

The overall framework of BioCyc is to define a single foundational database of experimentally elucidated pathways from many organisms (MetaCyc; SRI International, Menlo Park, CA) that is used to predict the metabolic pathways of other organisms from their sequenced genomes. Each prediction is modeled as a single organism-specific PGDB. Thus, the BioCyc organism-specific PGDBs each model the metabolism of a single organism in detail, whereas MetaCyc captures well-defined pathways from many organisms but does not define a comprehensive model of the pathways of any organism (except for E. coli, because MetaCyc contains all metabolic pathways from the EcoCyc [SRI International, Menlo Park, CA] PGDB).

BioCyc is divided into three tiers that reflect the degree of manual curation of these databases. The Tier 1 PGDBs EcoCyc and MetaCyc have undergone more than two person decades of curation each. By curation, we mean effort on the part of biologists to read the biomedical literature and to enter information from publications into these PGDBs.

Tier 1: EcoCyc

EcoCyc (4) describes the genome, the metabolic pathways, and the transcriptional regulatory network of E. coli K-12. EcoCyc curators enter newly discovered functions of E. coli genes into EcoCyc, as reported in the literature. They also enter E. coli metabolic pathways and information about E. coli operon organization, promoter locations, and control of those promoters by binding of transcription factors to nearby DNA sites. EcoCyc contains a written summary of the function of every E. coli gene for which experimental information is available. The information in EcoCyc was obtained from the more than 14,000 publications cited by EcoCyc.

Tier 1: MetaCyc

MetaCyc (5) is a multiorganism encyclopedia of metabolic pathways and enzymes. Like EcoCyc, it contains literature-derived information on experimentally elucidated metabolic pathways and enzymes. MetaCyc version 10.5 (October 2006) contains all...
197 metabolic pathways from EcoCyc and all EcoCyc metabolic enzymes. MetaCyc includes another 600 metabolic pathways from other organisms. Approximately half of the pathways in MetaCyc are from microorganisms, and approximately one-third of the pathways are from plants, with the remainder that comes largely from animals. The metabolic pathways in MetaCyc were elucidated experimentally in more than 700 organisms, and the information in MetaCyc has been drawn from more than 10,000 publications. MetaCyc contains extensive mini-review summaries and literature citations in its pathways. It also contains enzyme entries to explain the biologic functions of pathways and enzymes as well as to make this information accessible to scientists who are not experts in each pathway and enzyme.

The Tier 2 and Tier 3 PGDBs were derived computationally by applying the following sequence of computational operations to the annotated genomes of each organism, as described in more detail in the next section, and in Reference 2.

1. The annotated genome of each organism was converted to PGDB format.
2. The PathoLogic program predicted the metabolic pathway complement of each organism.
3. The PHFiller program predicted which genes within the organism will code for missing enzymes within the predicted metabolic pathways.
4. An operon predictor was executed for the bacterial genomes.
5. A cellular overview diagram was computed for each organism.
Tier 2

The Tier 2 PGDBs were created computationally by the preceding methodology, and then some amount of manual curation was applied to these PGDBs. For example, after being created computationally, the HumanCyc (SRI) International, Menlo Park, CA) PGDB (6) received extensive curation to assign human metabolic enzymes manually to their associated reactions; to enter 10 metabolic pathways and their enzymes from the literature into HumanCyc; and to enter associated summaries, literature citations, and other information such as enzyme regulators, cofactors, and subunit structure.

Tier 3

The Tier 3 PGDBs were created computationally by the preceding methodology, with no subsequent manual curation. We encourage scientists to adopt Tier 2 and Tier 3 PGDBs for ongoing curation and refinement. No single group can curate all the world’s genomes, so we encourage experts of the biology of an organism to assume responsibility for updating its PGDB to reflect existing and emerging information in the literature, on an ongoing basis.

Computing with the Metabolism of a PGDB

Once the metabolic network of an organism has been encoded using the preceding representation, many types of computational analyses are enabled.

Querying and visualization of metabolism

We are confronted immediately with the need to allow users to access information within metabolic databases. Pathway Tools provides several types of queries for each datatype within a PGDB. Users can query pathways, enzymes, metabolites, and proteins by exact name or by substring search. Additional queries supported include querying reactions by their Enzyme Commission (EC) number; querying metabolites by chemical substructures expressed in the SMILES language; querying pathways and reactions according to their substrates; and querying enzymes by molecular weight, pl, and by the small molecules that activate and inhibit them.

In presenting the answer to a query, the complexity of metabolic information demands the development of visualizations of the data that speed their comprehension by the user. Thus, an important aspect of the bioinformatics of metabolism is the visualization of metabolic information. Pathway Tools contains several visualizations of metabolism, all of which are generated automatically. It can produce drawings of individual metabolic pathways and of clusters of related pathways called superpathways. These drawings can be generated at multiple levels of detail so that the user can choose to show or to hide information such as enzyme and gene names, names of intermediate or side metabolites, and the chemical structures of metabolites. The drawings depict substrate-level regulation of the enzymes within a pathway, and all components of the drawing are clickable by the user. For example, clicking on a metabolite takes the user to a page that shows the metabolite structure and lists all its synonyms, all reactions and pathways in which it is a substrate, and all enzymes whose activities it regulates. Pathway Tools also generates information pages for enzymes and for biochemical reactions.

Pathway Tools can generate a visualization of the entire metabolic network of an organism, which we call the cellular overview diagram (7). This diagram is generated automatically from any PGDB, and it depicts all metabolic pathways in the PGDB as well as reactions not assigned to any pathway and all transporters identified in the PGDB. The overview diagram can be used to visualize omics datasets in a mode of operation called the Omics Viewer (7). The input to the Omics Viewer is a combination of gene expression data, proteomics data, metabolomics data, or other measurements that associate numbers with genes, reactions, or metabolites. The numbers are mapped to colors that are painted onto the elements of the cellular overview to allow the power of the human visual system to be used to interpret large-scale datasets in a pathway context. For example, a dot in the diagram that represents a single metabolite would be assigned a color that indicates the measured concentration of that metabolite in a metabolomics experiment. Finally, Pathway Tools can generate a poster-size version of the cellular overview complete with labels for entities in the diagram.

Prediction of metabolic pathways and pathway hole fillers

Pathway Tools predicts the metabolic pathway complement of an organism by assessing what known pathways from the MetaCyc PGDB are present in the annotated genome of a new organism. This inference is performed in two steps. First, enzymes in the annotated genome are assigned to their corresponding reactions in MetaCyc, which defines the reactome of the organism. The assignment proceeds by matching both the gene-product names (enzyme names) and the EC numbers assigned to genes in the genome. For example, the fabD gene in Bacillus anthracis is annotated with the function “malonyl CoA-acyl carrier protein transacylase.” That name was recognized by Pathway Tools as corresponding to the MetaCyc reaction whose EC number is 2.3.1.39. Therefore, Pathway Tools imported that reaction and its substrate into the B. anthracis PGDB, and it created an enzymatic-reaction object to link that reaction to that B. anthracis protein.

Once the reactome of the organism has been established, Pathway Tools imports into the new PGDB all MetaCyc pathways that contain at least one reaction in the organism’s reactome. Once imported, Pathway Tools attempts to prune out those pathways that are likely to be false-positive predictions. That pruning process considers both the fraction of reaction steps in the pathway that have assigned enzymes and how many of the reactions with assigned enzymes are unique to that pathway (as opposed to being used in additional metabolic pathways in that organism). The remaining pathways are those predicted to occur in the organism under analysis.
A final inference tool provided by Pathway Tools is called the pathway hole filler. A pathway hole is a reaction in a metabolic pathway for which no enzyme has been identified in the genome that catalyzes that reaction. Typical microbial genomes contain 200–300 pathway holes. Although some pathway holes are probably genuine, we believe that most probably result from the failure of the genome annotation process to identify the genes that correspond to those pathway holes. For example, genome annotation systems systematically under-annotate genes with multiple functions, and we believe that the enzyme functions for many pathway holes are unidentified second functions for genes that already have one assigned function.

The method used by the pathway hole filling program PH-Filler (8) is as follows. Given a reaction that is a pathway hole, the program first queries the UniProt database to find all known sequences for enzymes that catalyze that same reaction in other organisms. The program then uses the BLAST tool to compare that set of sequences against the full proteome of the organism in which we are seeking hole fillers. It scores the resulting BLAST hits by considering information such as genome localization, that is, is a potential hole filler in the same operon as another gene in the same metabolic pathway? At a stringent score-cutoff, our method finds potential hole fillers for approximately 45% of the pathway holes in a microbial genome.

Analysis and comparison of metabolic networks

Once the metabolism of an organism is captured in a computable form, we can write programs to characterize the size and structure of the metabolic network of an organism (9). For example, version 10.5 of the EcoCyc PGDB contains 176
pathways of small-molecule metabolism, which contain 702 component reactions. A total of 245 reactions of small-molecule metabolism are not assigned to a specific pathway. The cellular overview diagram can be used for comparative purposes by coloring those metabolic reactions shared between two organisms by using the desktop version of Pathway Tools. The Web version of Pathway Tools provides a suite of comparative analysis tools. For example, Fig. 4 shows comparisons of the overall pathway complements of E. coli and B. anthracis, which is broken down according to the Pathway Tools ontology of pathways. Figure 5 shows a detailed comparison of the pathways of biosynthesis of fatty acids and lipids in these two organisms.

A second form of pathway analysis is computing the potential outputs that the metabolic network might produce when supplied with a set of input metabolites (10). A third computational analysis method predicts choke points in the metabolic network, which are enzymes that if inhibited would be likely to create a major bottleneck in the metabolic network, and they are therefore likely to be good targets for developing antimicrobial drugs (11). In addition, it is possible to compute the equilibrium flux rates through an entire metabolic network (12).

Computational access to PGDBs
In addition to the user-friendly graphical interfaces to PGDBs provided through the Web and desktop versions of Pathway Tools, we provide the following modes of access to PGDBs to facilitate the construction of programs that explore pathway data computationally.

Programmatic access through application program interfaces (APIs)
Programmers can access and update PGDB data directly by writing programs in the Java, Perl, and Common Lisp languages (13).

Downloadable files in multiple formats
Pathway Tools can export PGDBs into several different file formats that are described at http://bioinformatics.ai.sri.com/ptools/flatfile-format.html. These formats include column-delimited tables, SBML (see http://sbml.org/), BioPAX (see http://biopax.org/), Genbank, FASTA, and attribute-value.

Relational database access via biowarehouse
For those who want to query PGDB data through a relational database system, the attribute-value files exported by Pathway Tools can be loaded into SRI’s BioWarehouse system (14). BioWarehouse is an Oracle or MySQL-based system for integration of multiple public bioinformatics databases. PGDB data can be queried through BioWarehouse alone or in combination with other bioinformatics databases such as UniProt, Genbank, NCBI Taxonomy, ENZYME, and KEGG.
Queries using the pathway tools query language, BioVelo

Pathway Tools provides a powerful and easy-to-use query language for querying PGDBs, called BioVelo. See http://biocyc.org/query.html for details.

Acknowledgment

I thank Carol Fulcher and Alexander Shearer for comments on this manuscript. This work was supported by grants GM077678, GM077905, and GM70065 from the National Institutes of Health.

References

Cell membranes are two-dimensional fluids that exhibit a wide range of dynamic behaviors. Recent technical advances have enabled unprecedented views of membrane dynamics in living cells. In this technical review, we provide a brief overview of three well-studied examples of membrane dynamics: lateral diffusion of proteins and lipids in the plane of the membrane, vesicular trafficking between intracellular compartments, and exchange of proteins on and off membranes. We then discuss experimental approaches to monitor membrane protein and lipid dynamics, and we place a special emphasis on the use of genetically encoded fluorescent probes and live cell-imaging techniques.

The concept that membranes are fluid, dynamic structures is now over 35 years old (1, 2). In this review, we describe three of the best-studied examples of cell membrane dynamics—lateral diffusion of proteins and lipids within the plane of the bilayer, membrane trafficking between intracellular compartments, and exchange of proteins on and off membranes—along with the recent technical advances that have enabled researchers to visualize these motions directly within living cells. In particular, we describe how the use of green fluorescent protein (GFP) from Aequorea Victoria and other, more recently developed labeling technologies can be used to mark molecules to study inside cells. We also summarize common fluorescence microscopy techniques for live cell imaging, including conventional techniques such as wide-field and confocal microscopy, and more specialized techniques such as total internal reflection fluorescence (TIRF) microscopy. Finally, we cover advanced methods used to study cell membrane dynamics, including single particle tracking (SPT), fluorescence recovery after photobleaching (FRAP), photoactivation, fluorescence correlation spectroscopy (FCS) and image correlation spectroscopy (ICS).

Types of Cell Membrane Dynamics

Lateral diffusion

Membranes are two-dimensional fluids whose protein and lipid components continuously exchange positions because of Brownian motion, a process commonly referred to as lateral diffusion. Lateral diffusion enables proteins and lipids to explore their environment, which encourages interactions between molecules. Thus, the speed of lateral diffusion is one of the limiting factors that regulate intermolecular interactions, and, consequently, cellular function. Diffusional mobility of molecules can be estimated based on their size and the viscosity of the lipid bilayer and surrounding aqueous phase. As a result, deviations from this behavior can provide important insights into the environment that proteins and lipid encounter in biologic membranes.

How rapidly diffusion occurs is characterized by the diffusion coefficient \( D \), a parameter that provides a measure of the mean of the squared displacement \( x \) of a molecule per unit time \( t \). For diffusion in two dimensions such as a membrane, this is given by \( \langle x^2 \rangle = 4Dt \). The Saffman-Delbrück model of Brownian motion in biologic membranes describes the relationship between membrane viscosity, solvent viscosity, the radius \( R \) and height of the diffusing species, and \( D \) for both lateral and rotational diffusion of proteins in membranes (3, 4). This model predicts for example that for lateral diffusion, \( D \) should be relatively insensitive to the radius of the diffusing species, scaling with \( \log (1/R) \).

Interestingly, the diffusional behavior of membrane proteins measured experimentally by FRAP, FCS, or single particle tracking in cells is more complex than predicted by this model. This technique is described best for the case of cell surface proteins, as assessed by FRAP. Such measurements indicate that diffusion is typically much slower than one would expect based on membrane viscosity. In cell membranes, typical values of \( D \) for transmembrane proteins are approximately 0.05 \( \mu \)m^2/s or less, which is much slower than observed in artificial membranes composed of purified lipids. In addition, a significant fraction of proteins is often immobile over the timescale of diffusion experiments (4, 5). Furthermore, diffusional mobilities vary among proteins, and sometimes they differ for the same protein expressed in different cell lines (4, 5). Deviations from pure diffusion are more readily apparent when the trajectories.
of single molecules are analyzed (Fig. 1). Individual molecules exhibit a range of diffusive behaviors, characterized as immobilization, transient confinement, free diffusion, and in some cases directed motion (6).

Features of cell membranes that contribute to this complex diffusional behavior include the presence of membrane domains, interactions of proteins with the cytoskeleton and extracellular matrix, and molecular crowding because of the high protein concentrations found in cellular membranes (4, 5). The diffusional mobility of proteins and lipids can also be modulated actively.

For example, crosslinking of certain cell surface molecules causes dramatic changes in their mobility, which reflects interactions with other cellular components. A well-studied example of such a protein is FcεR1, the high affinity IgE receptor, which undergoes a substantial loss of mobility on the formation of small aggregates of IgE-FcεR1 (7). Using an advanced technique known as FCS cross correlation analysis, it is now possible to detect transient interactions of membrane proteins that occur during cell signaling by virtue of their codiffusion (8).

Until recently, experiments that probe protein diffusion relied on fluorescent antibody-based probes, and thus were limited to plasma membrane proteins with extracellular epitopes. In contrast, the mechanisms that regulate diffusion of intracellular membrane proteins remained unexplored because of their inaccessibility to labeling. With the development of genetically encoded fluorescent probes, such studies have become tractable because proteins targeted to a particular organelle can be fluorescently labeled directly with GFP. In addition, improved technology has now made it possible to monitor the movement of multiple types of proteins or lipids, tagged with different markers simultaneously. Some examples of the types of questions it is now possible to address include:

How is protein diffusional mobility regulated in intracellular compartments?

Some of the first investigations of diffusion in organelle membranes such as the Golgi complex and endoplasmic reticulum suggest that unlike the plasma membrane, protein diffusion in intracellular membranes approaches theoretic limits (9). These findings suggest that intracellular membranes contain considerably fewer barriers to free diffusion than the cell surface. Diffusion studies can also be used to test whether protein immobilization is responsible for retaining proteins within a specific organelle, the effect of unfolding on protein mobility, and how various perturbations influence both the luminal and membrane environment experienced by proteins (10).

Is the mobility of proteins influenced by how they are attached to membranes?

Lateral diffusion is predicted to be relatively insensitive to the size of the transmembrane domain of the diffusing species. This prediction suggests that membrane anchorage should not strongly influence diffusion. However, membrane anchorage would be expected to influence the types of interactions that occur between a given protein and the complex environment of the cell. For example, the diffusion of peripheral membrane proteins localized to the inner leaflet of the plasma membrane by lipid anchors has been shown to be much faster than transmembrane proteins and nearly as fast as that of lipid probes in some instances (11). This finding suggests that because of their lipid anchors, these molecules do not experience the same barriers to diffusion as transmembrane proteins. It also likely reflects the ability of some proteins to undergo rapidly reversible binding to membranes (12, 13).

Do cholesterol-enriched lipid domains organize proteins into functional complexes?

Interest in the role of membrane domains in regulating protein and lipid diffusion has recently been revitalized by the lipid raft model, which proposes that cell membranes are divided into cholesterol and sphingolipid-enriched and-depleted microdomains. As a result, much effort has been made recently to relate lipid and protein diffusion in cells and model membranes (14) and to determine the effect of cholesterol levels on protein and lipid diffusion by both FRAP and FCS (11, 15). The role of cholesterol-dependent domains in regulating events such as T-cell signaling has also been investigated by examining closely the mechanisms of diffusional trapping of proteins at the single molecule level by TIRF microscopy; unexpectedly, protein-protein interactions seem to play a more important role than cholesterol-dependent domains in this process (16).

Membrane trafficking

Although integral membrane proteins and lipids are laterally mobile within the plane of a given cell membrane, they cannot exchange between different membrane compartments unaided because of the high energetic barriers to exposing their hydrophobic portions to water. The movement of transmembrane proteins between cellular compartments is accomplished by the formation of closed bilayer structures known as transport vesicles that bud off from one compartment and fuse with another. The term “membrane trafficking” refers to the movement
structures involved in this process (21, 22).

Membrane trafficking involves many highly regulated events, which include cargo selection, recruitment of coat and accessory proteins, vesicle budding, intra-organellar transport, targeting to and fusion with the target membrane, and recycling of cellular machinery for another cycle of transport (17). Recent work has now begun to focus on the temporal and spatial regulation of these events in living cells (18–20). Such approaches are especially powerful when coupled with cell biological manipulations such as drug treatments, the introduction of mutant proteins into cells, or knockdown of protein expression to probe the cellular mechanisms that underlie membrane trafficking (20). Examples of the types of questions being addressed are listed in the following sections.

What are the kinetics of intracellular transport through the secretory pathway?

The transmembrane protein VSVG is a classic tool used to study protein transport through the secretory pathway. Some early studies of vesicular transport in living cells visualized the movement of a GFP-tagged version of VSVG by time-lapse confocal microscopy (21, 22). From these data, it was possible to derive a kinetic model of the movement of this well-studied protein between intracellular compartments, while at the same time providing information about the nature of the vesicular structures involved in this process (21, 22).

What pathways do pathogens use to enter cells?

Certain viruses and bacterial toxins are endocytosed by a specialized pathway that involves invaginations of the plasma membrane known as caveolae. Recent studies have traced out this caveolar endocytic pathway by jointly tagging the cellular machinery with GFP and labeling viruses or bacterial toxins with fluorescent dyes, which enables their dual visualization in real time (23, 24). Such work has revealed several unanticipated properties of this pathway, such as the ability of viruses to induce formation of actin comets that propel virus-containing vesicles (23) and the stable, immobile nature of the caveolar coat that encases these vesicles (24).

How do cargo molecules and trafficking machinery progress through the endocytic pathway?

Movement of cargo through the endocytic pathway could occur by vesicular transport between stable compartments. Alternatively, it could involve progressive maturation of endosomal membranes. Recent work has tested these models by using fast, live cell imaging in combination with tools to quantify dynamic changes in the levels of the small GTPases Rab5 and Rab7 on endosomal structures during transport of fluorescently labeled cargo (25). These data suggest a model in which conversion in Rab proteins is a mechanism by which cargo progresses between early and late endosomes. Other studies have focused on how specific cargo molecules that enter the cell via identical pathways, such as cholera toxin and SV40, ultimately are sorted to various intracellular destinations. Sorting requires the pH-dependent release of cholera toxin from caveolar domains in endosomes; SV40 remains immobilized under these conditions (24).

Cycling of proteins on and off membranes

Although integral membrane proteins must rely on membrane trafficking as their sole mechanism for traversing the cell, many peripheral membrane proteins can move on and off membranes in a reversible manner by shifting between a membrane-bound and a soluble state. This reversible binding is critical for the function of coat proteins involved in the formation of transport vesicles (clathrin, COP1, and COPII), proteins that play roles in organelle identity and membrane trafficking (Rabs and Arf proteins), and a wide variety of signaling proteins (Ras, Raf, and protein kinase C) (26–28). Biochemically, such exchange can be detected by fractionating cells into membrane and soluble fractions, and quantifying the amount of protein found in each. Using live cell imaging approaches, it is possible to begin to monitor these events in real time to address questions such as the following.

On what cellular membranes does exchange occur?

One major advantage to studying membrane/cytosol exchange of proteins in living cells is that it is possible to compare the recruitment of proteins with different organelles simultaneously. For example, such studies have been instrumental in showing that Ras activation, as reported by the membrane recruitment of the FP-tagged Ras binding domain of Raf from the cytosol, occurs on Golgi membranes as well as the cell surface, and that the kinetics of recruitment are different for the two compartments (29).

How rapidly do proteins cycle on and off membranes?

Compared with membrane trafficking, which occurs with characteristic rates of 3% per minute for the case of VSVG movement from the endoplasmic reticulum to the Golgi complex (22), the cycling of proteins on and off membranes can occur over very rapidly. In our own work (30), we have observed in phatural bleaching of GFP fusion proteins that a mutant form of HRas that lacks both palmitoylation sites is able to constitutively cycle on and off membranes of the Golgi complex with half-times of ~5 s (Fig. 2).

Is exchange constitutive or regulated?

Membrane/cytosol exchange is sometimes constitutive, but more often it occurs in a tightly regulated manner. Many proteins are recruited to membranes in response to the transient generation of protein or lipid binding sites on the membrane, which are recognized by modular protein interacting or lipid interacting domains contained within the recruited protein. Alternatively, membrane binding can be regulated by loss of membrane-binding sites or perturbation of binding motifs within

These vesicular structures between organelles and/or the plasma membrane, and includes the processes of endocytosis and exocytosis.

Membrane trafficking involves many highly regulated events, which include cargo selection, recruitment of coat and accessory proteins, vesicle budding, intra-organellar transport, targeting to and fusion with the target membrane, and recycling of cellular machinery for another cycle of transport (17). Recent work has now begun to focus on the temporal and spatial regulation of these events in living cells (18–20). Such approaches are especially powerful when coupled with cell biological manipulations such as drug treatments, the introduction of mutant proteins into cells, or knockdown of protein expression to probe the cellular mechanisms that underlie membrane trafficking (20). Examples of the types of questions being addressed are listed in the following sections.

What are the kinetics of intracellular transport through the secretory pathway?

The transmembrane protein VSVG is a classic tool used to study protein transport through the secretory pathway. Some early studies of vesicular transport in living cells visualized the movement of a GFP-tagged version of VSVG by time-lapse confocal microscopy (21, 22). From these data, it was possible to derive a kinetic model of the movement of this well-studied protein between intracellular compartments, while at the same time providing information about the nature of the vesicular structures involved in this process (21, 22).

What pathways do pathogens use to enter cells?

Certain viruses and bacterial toxins are endocytosed by a specialized pathway that involves invaginations of the plasma membrane known as caveolae. Recent studies have traced out this caveolar endocytic pathway by jointly tagging the cellular machinery with GFP and labeling viruses or bacterial toxins with fluorescent dyes, which enables their dual visualization in real time (23, 24). Such work has revealed several unanticipated properties of this pathway, such as the ability of viruses to induce formation of actin comets that propel virus-containing vesicles (23) and the stable, immobile nature of the caveolar coat that encases these vesicles (24).

How do cargo molecules and trafficking machinery progress through the endocytic pathway?

Movement of cargo through the endocytic pathway could occur by vesicular transport between stable compartments. Alternatively, it could involve progressive maturation of endosomal membranes. Recent work has tested these models by using fast, live cell imaging in combination with tools to quantify dynamic changes in the levels of the small GTPases Rab5 and Rab7 on endosomal structures during transport of fluorescently labeled cargo (25). These data suggest a model in which conversion in Rab proteins is a mechanism by which cargo progresses between early and late endosomes. Other studies have focused on how specific cargo molecules that enter the cell via identical pathways, such as cholera toxin and SV40, ultimately are sorted to various intracellular destinations. Sorting requires the pH-dependent release of cholera toxin from caveolar domains in endosomes; SV40 remains immobilized under these conditions (24).

Cycling of proteins on and off membranes

Although integral membrane proteins must rely on membrane trafficking as their sole mechanism for traversing the cell, many peripheral membrane proteins can move on and off membranes in a reversible manner by shifting between a membrane-bound and a soluble state. This reversible binding is critical for the function of coat proteins involved in the formation of transport vesicles (clathrin, COP1, and COPII), proteins that play roles in organelle identity and membrane trafficking (Rabs and Arf proteins), and a wide variety of signaling proteins (Ras, Raf, and protein kinase C) (26–28). Biochemically, such exchange can be detected by fractionating cells into membrane and soluble fractions, and quantifying the amount of protein found in each. Using live cell imaging approaches, it is possible to begin to monitor these events in real time to address questions such as the following.

On what cellular membranes does exchange occur?

One major advantage to studying membrane/cytosol exchange of proteins in living cells is that it is possible to compare the recruitment of proteins with different organelles simultaneously. For example, such studies have been instrumental in showing that Ras activation, as reported by the membrane recruitment of the FP-tagged Ras binding domain of Raf from the cytosol, occurs on Golgi membranes as well as the cell surface, and that the kinetics of recruitment are different for the two compartments (29).

How rapidly do proteins cycle on and off membranes?

Compared with membrane trafficking, which occurs with characteristic rates of 3% per minute for the case of VSVG movement from the endoplasmic reticulum to the Golgi complex (22), the cycling of proteins on and off membranes can occur over very rapidly. In our own work (30), we have observed in photobleaching of GFP fusion proteins that a mutant form of HRas that lacks both palmitoylation sites is able to constitutively cycle on and off membranes of the Golgi complex with half-times of ~5 s (Fig. 2).

Is exchange constitutive or regulated?

Membrane/cytosol exchange is sometimes constitutive, but more often it occurs in a tightly regulated manner. Many proteins are recruited to membranes in response to the transient generation of protein or lipid binding sites on the membrane, which are recognized by modular protein interacting or lipid interacting domains contained within the recruited protein. Alternatively, membrane binding can be regulated by loss of membrane-binding sites or perturbation of binding motifs within...
Cellular Membranes, Dynamics of

![Image](http://example.com/image1.png)

**Figure 2** Visualization of reversible protein binding to the Golgi complex. (a) COS-7 cells that express a GFP-tagged HRas palmitoylation mutant were imaged over time before and after photobleaching fluorescent molecules localized to the Golgi complex (circle). Fluorescence recovers rapidly and completely within 1 min. Bar, 10 µm. (b) Kinetics of recovery for the GFP-HRas palmitolyation mutant after photobleaching the entire Golgi complex. Data are from a representative experiment similar to that shown in panel A. The rapid recovery kinetics are highly suggestive of reversible membrane binding. (Reproduced from Reference 30. Copyright 2005 Rockefeller University Press.)

proteins. For example, recent evidence suggests that KRas, which contains a cluster of basic amino acids that facilitates its binding to the plasma membrane, can be released via phospho-
rylation of residues in the polybasic domain. This release in turn allows for the regulated redistribution of KRas to mitochondria, where it promotes apoptosis (31).

Chemical Tools and Techniques to Study Cell Membrane Dynamics

**Probes**

Fluorescent probes of membrane dynamics fall into two general classes: exogenous and genetically encoded. Exogenous probes, which include small organic dyes and quantum dots, usually are targeted to the protein of interest via immunolabel-
ing approaches. They are used commonly to label cell surface proteins, although they can be introduced into cells by microin-
jection or by permeabilizing cells. Genetic approaches make use of intrinsically fluorescent proteins (FP) or genetic tags that generate binding sites for small dyes. Here, molecular biology tools are used to engineer DNA constructs that contain fusions of the protein of interest and the genetic tag. The DNA is then introduced into cells by transfection or transduction (Fig. 3).

**Small organic dyes**

Small organic dyes are available with a wide range of spectral properties encompassing the entire visible range, including the near UV and far red (32). Prior to the advent of FP, small organic dyes were the workhorses for fluorescent labeling of plasma membrane proteins, either by using dye-labeled anti-
bodies (Fig. 3a) or by labeling proteins directly that bind to plasma membrane proteins or lipids such as bacterial toxins or growth factors. Indeed, most early work that defines the diffu-
sional mobility of proteins on the plasma membrane relied on immunodetection of cell surface proteins with antibody frag-
ments labeled covalently with fluorescent dyes (5). The use of small organic dyes to label proteins fluorescently remains a valuable tool to study the endocytic itineraries of ligands, viruses, and bacterial toxins (24). However, dye-labeled proteins have historically been much less useful in studies of intracel-
tular proteins in living cells, because this task requires their introduction by either microinjection or permeabilization of the plasma membrane. The susceptibility of many fluorescent dyes to photobleaching also limits their use in live cell imaging ap-
plications.
Probes of lipid dynamics

Synthetic lipid analogs such as DiI and BODIPY-labeled lipids are classic reporters of lipid dynamics (32, 33). Fluorescent glycolipid analogs have also been used for many years, especially in studies of endocytosis and lipid trafficking (34). However, it is important to note the limitations of these probes. Many make use of short-chain analogs to enable their delivery into cells. In addition, some fluorophores used to label these lipids are used to replace a fatty acid chain. As a result, these lipids may not exhibit behavior similar to that of their native counterparts. To address these limitations, polyene lipids, which have a structure similar to that of natural lipids, have been developed recently as an alternative to classic fluorescent lipid analogs (35). Fluorescent cholera toxin analogs have joined the growing list of fluorescent lipid probes (36). Of these, dihydroergosterol is most likely to be a useful structural analog of cholsterol.

An alternative approach to the use of fluorescent lipid analogs to study lipid dynamics is the use of fluorescent reporter proteins that bind specific lipids. Such experiments are also of interest from the biologic standpoint of protein function. For example, plekstrin homolog (PH)-domain-containing proteins bind phosphoinositide lipids (26, 27), whereas cholera toxin binds ganglioside GM1 (20). It should be noted however that protein reporters that interact with a specific lipid might not be "neutral" from a biologic point of view, as they may compete for binding with endogenous molecules.

GFP and its derivatives

Many recent advances in the study of cell membrane dynamics (and indeed in much of cell biology) have been driven by the discovery of green fluorescent protein from Aequorea victoria (37). GFP is exceptionally large (27 kDa) and thus has the potential to disrupt the structure, function, and/or localization of the protein to which it is linked (Fig. 3b). Second, the expression of fusion proteins typically relies on transfection, which can be problematic in some primary cell lines and can lead to overexpression artifacts. Third, certain fluorescent proteins form obligate dimers. Finally, spectral overlap limits the number of different fluorescent proteins that can be used in the same experiment, although this problem is diminishing as new variants are developed.

Chemical labeling of fusion proteins

Recently, methods have been developed to label proteins site-specifically with small molecules for live cell imaging studies as an alternative to the use of FP fusion proteins (42-44). The strategy of these approaches relies on genetically incorporating a "receptor" domain that can serve as a specific binding site for a small molecule to the protein of interest. After expressing the fusion protein in cells, it can then be labeled using cell-permeant small molecule probes. This general scheme offers several major advantages over FP fusion proteins, which include 1) the potential to label proteins with relatively small, and thus in principle, minimally perturbing tags; 2) the possibility of taking advantage of small molecule probes with a wide range of chemical properties; and 3) the ability to control the time at which the proteins of interest are tagged, which allows for temporal regulation of labeling.

The tetracysteine-biarsenical system is one of the first examples of such technology (45). Here, a 12-residue sequence that includes four cysteines is incorporated into the protein of interest to enable binding of membrane permeant biarsenic dyes (FlAsH and R6GSH) (42). In addition to the small size of the tetracysteine motif, advantages of this approach include the possibility to perform correlative electron microscopy analysis and pulse-chase labeling (42). Furthermore, the fluorescence intensity of the biarsenic dyes increases substantially on binding, which decreases background fluorescence. However, the biarsenic dyes exhibit nonspecific binding to cysteine-rich proteins and require a reducing environment for labeling. Compared with the FP, this and other methods that combine genetic tags with small molecules are still in their infancy, but with additional iterations of refinement, they are likely to become useful in the future.

Quantum dots

Quantum dots are fluorescent semiconductor nanocrystals that have been incorporated recently into the toolbox of fluorescent labeling techniques (42). When coated appropriately, quantum dots can be conjugated with streptavidin or antibodies for protein labeling applications in cells (46). Quantum dots are exceptionally bright and photostable, have a broad absorbance spectrum, and can be tuned to emit at specific wavelengths depending on their size. Because of their brightness and photostability, they can be very useful for detecting low abundance proteins and are attractive probes for long-term single particle tracking studies. However, their applications in live cells are hampered somewhat by their large size (~10 to 30 nm) and "blinking" behavior. For example, the size of quantum dots has limited their use primarily to studies of extracellularly localized plasma membrane proteins because their introduction into cells...
tissue imaging a technique of choice for fluorescence microscopy. The forms of fluorescence microscopy used most commonly for live cell imaging are wide field and confocal microscopy. In wide field microscopy, fluorescence is typically excited with an arc lamp and emission is collected using a CCD camera. Such systems can be configured readily for imaging live cells over time. An advantage of this approach is its relative simplicity compared with other imaging modalities.

However, it collects light emitted from the entire depth of the specimen including out-of-plane fluorescence.

In confocal microscopy, out of plane fluorescence is eliminated by the incorporation of a pinhole in the light path, which enables the collection of three-dimensionally resolved images. This task is accomplished by using a single pinhole for the case of laser scanning confocal microscopy or with a series of rotating pinholes in spinning disk confocal microscopy. Because the intensity and position of the laser used to excite samples can be modulated rapidly in many laser scanning confocal microscopes, it is possible to use them for specialized applications such as photobleaching and photobleaching (see below). Spinning disk confocal microscopes are not well suited for photobleaching or photoactivation applications, but they have a greater image acquisition rate than laser scanning confocals.

In addition to confocal microscopy, a technique known as multiphoton microscopy can also be used to generate three-dimensionally resolved fluorescence images. In this case, excitation is limited to those fluorophores that are present within the small region where the laser is focused and thus are of sufficiently high power to excite a single fluorophore to absorb two or more photons simultaneously (47). Other advantages of multiphoton microscopy include reduced photobleaching outside of the focal plane as well as increased penetration into samples because of the use of longer wavelength excitation, which makes this a technique of choice for tissue imaging in vivo.

In some instances, it is of interest to focus exclusively on events that occur at the plasma membrane. Here, a technique known as TIRF microscopy is particularly valuable (48). TIRF uses an evanescent wave generated by a process referred to as total internal reflection to excite the sample. Total internal reflection occurs when light traveling from a medium of high refractive index arrives at an interface with a medium of lower refractive index above a so-called critical angle. Under these conditions, the light is reflected back into the high refractive index material, and an evanescent field is generated in the lower refractive index medium. The field decays exponentially away from the interface, with a typical depth between 50 to 150 nm. Because experimentally, this interface represents the surface at which cells attach to a coverslip, it is possible to selectively excite and visualize only those fluorescent molecules found at this surface. TIRF can also be combined with the advanced methods described in more detail below, which include single molecule imaging (16), FCS (49), and FRAP (50). Thus, TIRF offers an attractive imaging modality for probing plasma membrane dynamics using a variety of techniques. For a more in-depth discussion of the pros and cons of each of these approaches in live cell imaging studies, we refer the reader to several recent reviews (51, 52).

Advanced techniques used to study membrane dynamics

FRAP
FRAP (also known as fluorescence photobleaching recovery or FPR) has been used for many years to study lateral diffusion of plasma membrane proteins as characterized by their diffusion coefficient and mobile fraction (5, 53). In these experiments, a population of fluorescence molecules is bleached irreversibly by exciting with an intense, focused laser spot. Recovery of fluorescence in the bleached region is then monitored over time to determine to what extent and how rapidly the bleached molecules are replaced by unbleached molecules from other regions within the cell (Fig. 4a). Two fundamental assumptions that underlie these experiments are that the bleaching event does not damage the labeled protein or surrounding region of the cell, and that the bleaching is irreversible. In other words, that recovery occurs because of diffusional exchange and not by recovery of fluorescence of an individual fluorophore. FRAP curves are typically fit by equations for free diffusion plus an immobile fraction (53). A variation on this approach is to measure FRAP as a function of spot size, which is a technique that is sensitive to the cycling of proteins on and off membranes (13).

Until the development of the fluorescent proteins, FRAP measurements were confined to measurements of protein or lipid diffusion in the plasma membrane using fluorescently labeled Fabs or fluorescent lipid analogs that could be added to cells exogenously. For such experiments, little need existed for spatially resolved measurements because the fluorescence signal was localized to the plasma membrane. Thus, most classic FRAP studies made use of a spot photobleaching apparatus that consisted of an epi-fluorescence microscope, computer-controlled shutter, laser used both for bleaching and for low-level excitation of the sample, and detector to collect fluorescence emission coupled with electronics to resolve the measured fluorescence intensity (54). In recent years, commercially available confocal microscopes have incorporated FRAP protocols, which allows for imaging-based FRAP measurements of intracellular fluorescent proteins. These two developments have brought FRAP into the mainstream of cell biological techniques, especially those related to questions of membrane dynamics.

Examples of useful applications of confocal FRAP include selective photobleaching and fluorescence loss in photobleaching (FLIP) (18, 55). In contrast with conventional spot photobleaching FRAP, in confocal FRAP it is possible to visualize both the bleach region and the surrounding area of the cell. In addition, confocal FRAP techniques typically use...
Selective photobleaching. Fluorescent molecules in an individual compartment (as illustrated here for the Golgi complex, circled) are photobleached. The sample is then monitored over time to determine whether fluorescence can recover (d) or not recover (e) from other regions of the cell. (Adapted from Reference 62 with permission from Elsevier.)

Figure 4 Principles of photoactivation and photobleaching experiments. For purposes of illustration, regions of bright fluorescence are shaded gray, and areas that contain little or no fluorescence are shown in white. (a) FRAP. Here, molecules in a region of interest (box) are photobleached, and their exchange with fluorescent molecules from the surrounding region is monitored over time. (b) Photoactivation. Similar to FRAP in principle except here the molecules in the region of interest (box) are converted to a different state. The redistribution of photoactivated molecules can then be monitored selectively. (c) FLIP. Repeated photobleaching of a region of interest (square) is performed, while monitoring the loss of fluorescence from other regions of the cell. (d, e) Selective photobleaching. Fluorescent molecules in an individual compartment (as illustrated here for the Golgi complex, circled) are photobleached. The sample is then monitored over time to determine whether fluorescence can recover (d) or not recover (e) from other regions of the cell. (Adapted from Reference 62 with permission from Elsevier.)

Large and/or complex regions of interest and may incorporate repetitive bleaching protocols instead of a single bleaching event. For example, in selective photobleaching experiments, an entire subcellular compartment such as the Golgi complex is photobleached to examine the mechanism and kinetics of recovery from elsewhere in the cell (Fig. 4d-e). Using this approach, it is possible to monitor the kinetics of coat protein cycling on and off membranes (28) as well as to assess the kinetics of vesicular and nonvesicular transport in the secretory and endocytic pathways (18) (Fig. 2). In FLIP, a single region of interest is bleached repetitively, which allows for recovery of fluorescent material to occur in between each repetition. This technique causes a gradual depletion of fluorescence material in regions of the cell that are in communication with the bleach region (Fig. 4c). Thus, the rate and the extent of loss of fluorescent material from the area outside of the bleach region depends on the degree of connectivity between compartments (18).

Photoactivation

A thorough time-lapse imaging is often sufficient to monitor vesicular trafficking and transient membrane binding events, for other experiments it is advantageous to "mark" a particular group of fluorescent molecules and watch their redistribution over time. The generation of photoactivatable fluorescent proteins (39) has now made it possible to "highlight" a population of molecules for optical tracking experiments (55). As a result, movement of the photoactivated molecules can be visualized directly in a "pulse-chase" experiment (Fig. 4a). Many of the same concepts described above for FRAP experiments can also be applied to photoactivation.

Single particle tracking

Single particle tracking is a technique that visualizes directly the movements of individual molecules, small groups of molecules, or even viruses by either fluorescence microscopy or light microscopy (54, 56). For such experiments, individual proteins can be expressed as GFP fusion proteins, labeled with fluorescent antibodies, or labeled with quantum dots. Alternatively, they can be immunolabeled with probes that can be detected by light microscopy such as 30–60 nm gold particles or latex beads. In the limit of sparse labeling, the trajectory of individual tagged molecules can then be tracked with high temporal and spatial accuracy using highly sensitive cameras (54). Although the resolution of fluorescence microscopy is ∼250 nm, the centroid of a single molecule can be determined with an accuracy of ∼10 nm (54). Typical rates of image acquisition are ∼30 frames/sec, although much faster acquisition rates can be obtained with specialized cameras. The resulting trajectories can be plotted directly to show the movements of each molecule, or can be analyzed even more to determine the mean squared displacement as a function of time (Fig. 1). For example, during free diffusion, the molecule moves randomly, which results in a characteristic linear relationship between mean squared displacement and time. A hallmark of confined diffusion is that the mean squared displacement of the particle is at first linear, then plateaus, which reflects the limited movement of the molecule within a confined region of the membrane. A given molecule may shift between several modes of motion during the observation period. Thus, the percentage of molecules that undergo each type of motion and/or the fraction of time they exhibit a particular behavior is often reported.

An obvious advantage of single molecule tracking is that it allows for highly detailed analyses of the movements of individual molecules that are obscured in population-based measurements such as FRAP and photoactivation experiments. However, SPT
experiments are not trivial to perform and are subject to several potential artifacts. The first challenge is to demonstrate that single molecules are being studied. Crosslinking can occur in studies that use antibody-labeled gold beads or quantum dots as probes, and lead to changes in mobility. In addition, large probes can potentially interact with the extracellular matrix. Photobleaching of organic dyes or FP-fusions can be rapid, which limits visualization times. However, photobleaching can also be used to confirm that single molecules are being visualized, because they will undergo a single-step photobleach. Finally, because of the intrinsic variability in trajectories, careful analysis is required to distinguish between motions that arise from free diffusion and confined diffusion.

FCS and ICS

FCS is not an imaging technique per se, although it is often performed using a microscope-based system. Instead, FCS measures the movement of individual fluorescent molecules through a defined observation volume, recorded as fluctuations in fluorescence over time (57). Such measurements require a sensitive photodetector, a dilute sample (~om) and a sampling volume with femtoliter dimensions. The observation volume can be generated using a laser focused to a diffraction-limited spot with a confocal pinhole placed in front of the detector, multiphoton illumination, or TIRF excitation in conjunctions with a pinhole. A notable strength of FCS is its sensitivity to fluorescence fluctuations over a wide range of timescales, from fast kinetics corresponding to photophysical properties of fluorescent proteins to the diffusion of proteins in cell membranes (58). In addition, FCS measurements can be collected within user-defined regions in an individual cell or artificial membrane vesicle, which allows for comparison of protein or lipid dynamics in distinct membrane environments (15). FCS can also be used to measure kinetic rate constants, for example, the association and dissociation of fluorescently labeled molecules to the plasma membrane (48).

The fluorescence fluctuations measured by FCS can be analyzed in several ways. The most common technique, autocorrelation analysis, involves measuring the time dependence of fluorescence fluctuations over a wide range of timescales and then performing a Fourier transform of the autocorrelation. A notable advantage of using a laser-scanning microscope is that it provides a way to visualize the diffusional properties of individual molecules. In addition, FCS can also be used to estimate the rate constants of, for example, the association and dissociation of fluorescently labeled molecules to the plasma membrane (48).

Whereas FCS measures fluorescence fluctuations over time, a related technique, ICS, measures fluorescence fluctuations over space, in particular from images collected using a laser-scanning microscope. ICS analysis of pixels within a single image provides information about protein clustering and density. A variation of ICS known as imagewise correlation spectroscopy evaluates the interactions of molecules labeled with different fluorescent probes. ICS can also be performed on stacks of images collected as a function of time. This spatial-temporal version of ICS can be used to monitor slow protein dynamics, and even be used to generate vector maps of directed protein movements in living cells (60).

Acknowledgments

I apologize to those whose original research could not be cited because of space limitations. I thank Maria Kiskowski, Lynne Lapierre, Erik Snapp, and Michael Edidin for their comments on the manuscript and Erik Snapp for providing Figure 2. Supported by RO1GM-073846.

References


Further Reading


See Also

Imaging Techniques: Overview of Applications in Chemical Biology
Lipid Bilayers, Properties of
Membrane Assembly in Living Systems
Membranes, Fluidity of
Membrane Trafficking
Membrane Compartments Related to Signaling and Trafficking in Immune Cells

Stephanie Hammond, David Holowka and Barbara Baird, Cornell University, Ithaca, New York
doi: 10.1002/9780470048672.wecb307

Cells of the immune system detect and respond to the presence of foreign antigens by signaling through an array of receptors, which include the multichain immune recognition receptor (MIRR) family. Stimulated changes in membrane architecture, which includes protein and lipid segregation within membranes as well as membrane trafficking, regulate these cellular responses. We provide a brief overview of MIRR signaling, which focuses on signaling through the IgE receptor in mast cells, and we highlight specific examples in which membrane compartmentalization plays a role in this signaling pathway. We summarize biochemical methods used to isolate and characterize membrane subdomains. Finally, we discuss cross-correlation microscopy methods, application of antigen-patterned surfaces, and fluorescence-based trafficking assays to study dynamics of stimulated proteins interactions and membrane trafficking.

Membrane compartmentalization is required for rapid and efficient signaling responses in immune cells, such as mast cells, T cells, and B cells. Changes in the organization of membrane proteins and lipids are necessary for a spatially regulated response, and trafficking of intracellular compartments contribute to this spatial reorganization. Mast cells that express the high-affinity receptor for IgE, FcεRI, are a well-characterized system for investigating the role of membrane domains known as lipid rafts in the signaling pathway. However, changes in membrane compartmentalization necessary for signaling are not limited to plasma membrane lipid rafts. Subdomains of intracellular organelles, specifically the endoplasmic reticulum and the endosomal system, function to regulate membrane interactions and directional trafficking in this system. In this article, we discuss molecular biology tools that are being applied to the study of protein trafficking and membrane compartmentalization. We outline standard methods to isolate membrane subdomains, which include isolation of detergent-resistant membranes, cholesterol depletion, and subcellular fractionation. We then conclude this article by discussing spectrofluorometric methods used to study protein colocalization and membrane trafficking in living cells.

Signaling Through Antigen-Binding Receptors

Immune cells express a variety of receptors to detect foreign antigens and respond in a selective manner to clear pathogens and infected cells from the body. Both soluble and membrane-bound antigens are recognized by the multichain immune recognition receptor (MIRR) family of antigen-binding receptors. This family includes three well-characterized members: the high-affinity receptor for IgE (FcεRI) expressed on mast cells and basophils, the T cell receptor (TCR), and the B cell receptor (BCR). These receptors consist of transmembrane ligand-binding subunits that are noncovalently associated with signaling subunits that contain immunoreceptor tyrosine-based activation motifs (ITAMs) in their cytoplasmic domains (Fig. 1a). The binding of specific ligands cross-links these receptors and initiates signaling. The FcεRI consists of the IgE-binding α subunit, which associates with the signal-transducing γ-homodimer and the signal-amplifying β subunit, and it serves as an archetype for studying immunoreceptor signaling. The TCR consists of an
antigen-binding \( \alpha \beta \) heterodimer, the signal-transducing CD3 \((\kappa \epsilon \gamma)\) complex, and the \( \epsilon \) chain homodimer. TCRs recognize short peptide sequences bound to major histocompatibility complex (MHC) proteins on the surface of antigen-presenting cells. The BCR consists of an antigen-binding membrane immunoglobulin subunit that associates with the signal-transducing Ig-\( \alpha \)/\( \gamma \) complex. B cell signaling is enhanced through interactions with the B cell coreceptor complex that contributes additional phosphorylation sites.

**Mast cell signaling through FcRI and lipid rafts**

Mast cells are the primary cellular mediators of allergic reactions and also function in innate immunity and defense against helminths (1). These tissue-resident cells reside in strategic locations throughout the body, most prominently in the mucosal membranes of the respiratory and digestive tracts. The rat basophilic leukemia (RBL-2H3) cell line, which is derived from mucosal mast cells, is often used as a model system to study FcRI signal transduction. These cells abundantly express FcRI \((-200,000\) receptors per cell) at the plasma membrane, where they bind allergen-specific IgE with high affinity \((K_D \sim 10^{-10} \text{M})\) (2), sensitizing the cell and priming it to respond rapidly during subsequent exposure. Mast cells contain numerous secretory granules that store preformed allergic mediators, which include histamine, and stimulated release of these contents is responsible for the immediate symptoms of allergic reactions.

Signaling through FcRI is initiated by binding of multivalent antigen by receptor-bound IgE and consequent aggregation of individual receptor complexes. FcRI clustering promotes stable association with specialized regions of the plasma membrane, which are termed lipid rafts (3). Lipid rafts are microdomains of ordered lipids that are enriched in cholesterol and phospholipids with saturated acyl chains, including sphingolipids, and they coexist with regions of more fluid, disordered lipids (4). Several types of measurements indicate that as much as 40% of membrane lipids exist in ordered regions, although most transmembrane proteins prefer disordered regions (3). In unstimulated cells, lipid rafts seem to be small, \( \sim 20-100 \) nm, and highly dynamic structures (5, 6). Coalescence and stabilization of lipid rafts by antigen-mediated cross-linking of IgE-FcRI is believed to initiate this signaling cascade. This step is a ligand-sensing event that depends on antigen valency and receptor affinity for the antigen; receptor oligomerization increases receptor residency time in lipid rafts and allows for signal initiation (4).

**Phosphorylation cascade**

Lipid rafts are one mechanism by which cells can segregate plasma membrane components and compartmentalize signaling events. Lyn, which is a Src family kinase, is dually acylated at its N-terminus by myristoylation and palmitoylation that target Lyn to the plasma membrane and confer its association with lipid rafts. Monomeric receptors are located largely outside of lipid rafts and are effectively segregated from the activating kinase in unstimulated cells. After receptor cross-linking and association with lipid rafts, Lyn phosphorylates Fc RI ITAMs (Fig. 1b). The lipid raft environment enhances signaling by protecting Lyn from inactivation through the exclusion of transmembrane phosphatases (7). Phosphorylation of the \( \gamma \) subunits is the first biochemically detectable step in FcRI signaling. The actin cytoskeleton plays a negative regulatory role in mast cell signaling (8), in part by limiting the functional interactions of FcRI and Lyn through segregation of Lyn from cross-linked FcRI complexes (9). The lipid composition of ordered microdomains in the plasma membrane is also regulated by the actin cytoskeleton (3).

In the primary signaling pathway, Syk kinase is recruited to the phosphorylated y subunit via its tandem Src homology 2 (SH2) domains (10). Once activated through phosphorylation and conformational changes, Syk phosphorylates several downstream targets, which include the linker for the activation of T cells (LAT), an adaptor protein with multiple ITAMs that, when phosphorylated, serve as binding sites for SH2 domain-containing proteins (11). LAT forms complexes with several signaling proteins downstream of Syk including phospholipase C-\( \gamma \) (PLC-\( \gamma \)). Activated PLC-\( \gamma \) hydrolyzes phosphatidylinositol-4,5-bisphosphate (PIP2) to produce soluble inositol-1,4,5-trisphosphate (IP3) and membrane-bound diacylglycerol (DAG) (10). These products lead to calcium mobilization and protein kinase C activation, which are both required for mast cell degranulation. Additionally, scaffolding proteins recruit several guanine nucleotide exchange factors (GEFs), which in turn bind to low-molecular-weight GTPases, leading to activation of the MAPK signaling cascade and cytokine production, as well as cell morphological changes.

A complementary pathway activated during FcRI stimulation is initiated by the Src family kinase Fyn, as revealed in bone-marrow-derived mast cells. Fyn mediates phosphatidylinositol-3-kinase (PI-3-K) activation that phosphorylates PI(4,5)P2 to PI(3,4,5)P3 by recruiting pleckstrin homology (PH) domain-containing proteins, such as the Tec family kinase Btk, PLC-\( \gamma \), and GEFs that regulate changes in the actin cytoskeleton through the Rho family of GTPases, PI-3-K activation is necessary for the maintenance, but not the initiation, of calcium signaling required for degranulation and may influence cytokine production because of its ability to affect intracellular calcium concentrations (10). Studies in RBL mast cells with trivalent ligands of defined lengths are consistent with a PI-3-K pathway that operates in parallel to the Lyn-Syk pathway (12).

**Elevation of intracellular calcium levels**

Cytoplasmic calcium concentrations in unstimulated cells are \( \sim 100 \) nM but can rapidly increase to \( \geq 1 \mu \text{M} \) during stimulation. In lymphocytes, this increase is a biphasic process. The initial increase is caused by IP3 binding to the IP3 receptor (IP3R) in the endoplasmic reticulum (ER) and rapid release from intracellular stores. The second, sustained phase results from extracellular calcium influx through calcium-release activated calcium (CRAC) channels in a process termed store-operated calcium entry (SOCE) and also through other calcium channels including the transient receptor potential channels and plasma membrane.
Membrane Compartments Related to Signaling and Trafficking in Immune Cells

Figure 1  Signal transduction pathways from antigen receptors. (a) The MIRR family of receptors consists of the ligand-binding units associated with signal transducing units that contain ITAM sequences on transmembrane receptor subunits. (b) In unstimulated mast cells, monomeric IgE receptors largely reside outside of lipid rafts. After cross-linking by a multivalent antigen, IgE receptors cluster and stably associate with lipid rafts where Lyn phosphorylates the receptors and initiates a signaling cascade that culminates in the release of allergic mediators. Molecules are not shown to scale.

membrane $p_{70,75}$ (13). SOCE is the main mechanism by which cytoplasmic calcium levels are increased in lymphocytes (14). Rapid increases in intracellular calcium are necessary to trigger secretory vesicle fusion and extracellular secretion, whereas sustained calcium elevation is required to activate gene transcription and cytokine production.

ER heterogeneity and organization of ER subdomains in relation to calcium signaling received increased attention recently after several groups independently identified an ER calcium sensor protein, STIM1, and a component of the CRAC channel, Orai1/CRACM1, using genetic approaches (15). After depletion of ER calcium, STIM1 oligomerizes and accumulates in ER subregions near the plasma membrane. Orai1 in the plasma membrane co-clusters with STIM1 in the ER, and the simultaneous accumulation of both proteins allows for activation of CRAC channels (16). Calcium levels are returned to basal levels after stimulation through the actions of sarcoendoplasmic reticulum $\text{Ca}^{2+}$-ATPase channels in the ER, which refill the calcium store, and plasma membrane calcium pumps and exchangers, which export calcium into the extracellular environment.

Degranulation and stimulated membrane trafficking

Mast cell stimulation results in spatially regulated trafficking of several intracellular compartments that include secretory lysosomes, which release allergic mediators during degranulation, and recycling endosomes, a heterogeneous perinuclear compartment of mildly acidic membranes (17). Stimulation increases the outward trafficking of a lipid-raft component of recycling endosomes, as monitored by cholera toxin subunit B (CTxB) binding to GM$_1$ (18), which are targeted toward sites of cross-linked IgE receptors (19), whereas secretory lysosomal fusion occurs at sites distinct from receptor-signaling complexes (19). Furthermore, secretory lysosomal fusion is differentially regulated from cytokine trafficking; the latter may occur through recycling endosomal trafficking (20) as recently demonstrated in macrophages (21).
Members of the Rab and Araf GTase families organize membrane compartmentalization within the endosomal system that is important for regulating transport through these organelles. Several Rab proteins localize to distinct subdomains on early and recycling endosomes and contribute to the sorting and trafficking functions of these compartments (22). These proteins also regulate the outward trafficking of lipid rafts from recycling endosomes (23).

**T cell signaling through the TCR**

T cell stimulation begins with binding of the antigenic peptide-MHC complex on antigen-presenting cells to the TCR, and the Src family kinase Lck initiates a phosphorylation cascade (24). The Syk family kinase Zap-70 is recruited to phosphorylated TCR ITAMs, where it is phosphorylated by Lck. Activated Zap-70 then phosphorylates the adaptor protein LAT, which coordinates the recruitment and activation several downstream targets including PLCγ, PI-3-K, and GEFs that activate low molecular weight GTases. These proteins couple receptor activation to calcium mobilization and changes in the actin cytoskeleton. Sustained calcium signaling is required for the formation of the immunological synapse at the interface between the T cell and the antigen-presenting-cell (14). In this structure, the TCR is localized to a central supramolecular activation cluster and is surrounded by a peripheral ring of adhesion molecules (25). The immunological synapse may facilitate sustained signaling necessary to induce cytokine production and T cell proliferation (14).

Initial signaling begins in small microclusters that form quickly after T cell stimulation and precedes formation of the immunological synapse (26). Coalescence of small structures into a larger signaling complex is dependent on stimulated actin polymerization and requires signaling through additional costimulatory molecules, which lead to actin cytoskeleton-regulated reorganization of lipid rafts (26). TCR signaling induces a transient dephosphorylation of the ezrin, radixin, and moesin (ERM) protein family, which temporarily uncouples lipid rafts from the actin cytoskeleton and allows for their coalescence into larger signaling structures (27). Trafficking of raft-enriched intracellular compartments increases the raft content of the plasma membrane and may contribute to targeting of rafts to the immunological synapse (26).

Concurrent with polarization of actin polymerization, the microtubule organizing center and secretory granules, which contain cytoplasmic agents in cytotoxic T cells and cytokines in helper T cells, are also polarized toward the antigen-presenting-cell (27). Helper T cells also spatially regulate cytokine secretion, releasing IL-2 and IFN-γ into the immunological synapse but repressing TNF multi-directionality (28). Targeting granule release to the target cell produces locally high concentrations while minimizing unwanted effects on surrounding cells, whereas multi-directional release promotes recruitment of additional immune cells.

**B cell signaling through the BCR**

The BCR mediates antigen uptake for processing and presentation of the antigenic peptide-MHC class II complex to T cells, which promotes B cell differentiation into antibody-producing plasma cells and memory cells. A ligand binding to the BCR initiates a signaling cascade that involves Src, Syk, and Tec kinases that results in phosphorylation of the adaptor protein BLNK (B cell linker protein), which couples BCR signaling to calcium mobilization similarly to LAT (29). The role lipid rafts play in BCR signaling is still under investigation as some BCR signaling can occur outside of rafts (30), and the outcome of BCR raft association changes during B cell development (31). BCR signaling also results in transient dephosphorylation of ERM proteins and rapid depolymerization of the actin cytoskeleton, which facilitates coalescence of lipid rafts into stable signaling domains (32). In this system, the actin cytoskeleton may segregate lipid rafts in unstimulated cells and, after a strong stimulus, extensive actin depolymerization leads to stable lipid raft clusters together with dense signaling sites at the synapse (32). BCR signaling coordinates receptor internalization and reorganization of the endomembrane system essential for B cell antigen presentation. Lipid raft association triggers endocytosis of the antigen-BCR complex and delivery to MHC class II-containing intracellular compartments for peptide loading onto MHC class II molecules (30), which is necessary for antigen presentation to helper T cells. Consequent interactions with helper T cells sustain B cell activation, which promotes immunoglobulin class switching and antibody production. The strength of the activation signaling (32) and the B cell microenvironment determines the ultimate outcome of B cell activation.

**Techniques to Study Membrane Compartmentalization**

Changes in protein and membrane interactions are most effectively studied with integrated biological and physical techniques. Recent advances in molecular biology enable mutation, overexpression, or depletion of a specific protein to assess its function within the signaling pathways. These tools to isolate functional membrane domains are used to assess stimulated changes in protein and lipid content. Fluorescence techniques are used to visualize and quantify protein interactions and directional membrane trafficking in live cells under physiological conditions.

**Molecular biology methods**

A protein’s role in a signaling pathway can be assessed by introducing mutations that disrupt interactions, localization, or activity or by silencing gene expression. The dynamics of engineered proteins can be monitored via selective placement of fluorescent tags within the protein sequence.

**Site-directed mutagenesis**

DNA engineering has become a standard laboratory procedure by which engineered proteins are used to study the role of protein localization and function within signaling pathways. Site-directed mutagenesis, in which only one or a few specific amino acids are altered in a protein, is a valuable tool for elucidating the structural properties of proteins that mediate
protein dynamics. This approach was used, for example, to examine the role of MIRR transmembrane domains in mediating lipid raft association in mast cells (33) and B cells (34). The addition or deletion of a lipid modification site within the target protein can also be used as a tool to either enhance or decrease membrane or lipid raft association. A membrane targeted form of the C-terminal Src kinase (Csk), mCsk, was generated via the addition of a myristoylation site and was found to suppress basal Lyn activity in mast cells by phosphorylating the C-terminus of Lyn (35). Site-directed mutation of the myristoylation and palmitoylation sites in Lyn demonstrated that both modifications are required for membrane targeting and that palmitoylation enhances lipid raft association (36).

Differential epitope tagging

One way to distinguish the genetically engineered protein from the cellular endogenous protein is to tag the engineered protein with either a fluorescent label or a nonfluorescent epitope tag. Fusion proteins tagged with green fluorescent protein (GFP) or any flavor of the many spectral variants of fluorescent proteins (FPs) (37) can be used to visualize the spatio-temporal distribution of several different proteins simultaneously within a single cell. Protein dynamics and interactions can be monitored using real-time fluorescence microscopy and quantitative microscopy techniques described below. Protein localization is often mediated by modular protein domains, such as PH domain binding to lipid species within the target membrane or SH2 and SH3 domains that bind to specific protein sequences of the target protein, which results in protein redistribution. The fusion of FPs to these modular domains generates fluorescent probes that can be used to monitor changes in lipid composition (38) and protein phosphorylation (39).

Alternatively, epitope tagging of proteins, in which a small epitope is inserted into the target protein sequence, is used to visualize protein localization by immunofluorescence microscopy of fixed and permeabilized cells. A variety of well-characterized epitope tags are available (40), and commercially available antibodies are used to detect the epitope tagged protein. These antibodies are used to determine the level of endogenous protein by either transient or stable knock down provides an environment for directly assessing the tagged-protein function without the endogenous protein (41). Stable knock down can be achieved with short hairpin RNAs (shRNA), which are processed inside the cell to generate active siRNA structures (42). Viral-mediated RNAi can reduce endogenous mRNA levels by ~75–90%, and co-transfection with multiple siRNA sequences can be used to produce double knockdowns (43). These double knockdowns can be used to generate cell lines or transgenic animals with stably silenced gene expression (43). A further approach to produce stably silenced cell lines is to use antisense cDNA within a plasmid vector, and this method can reduce the levels of endogenous protein by ~90% (46). Either transient or stable knock down provides an environment for directly assessing the tagged-protein function without the endogenous protein background.

Biophysical and biochemical methods

Two common methods to study the role of lipid rafts in signaling pathways are detergent insolubility followed by isolation on a density gradient and cholesterol depletion. Other methodologies exist for isolating plasma membrane samples and intracellular organelles for additional characterization (47).

Detergent-resistant membranes

Detergent-resistant membranes (DRMs) were initially defined in terms of their resistance to solubilization by nonionic detergents, such as Triton X-100 (Pierce, Rockland, IL), at low temperatures, followed by flotation at low densities in equilibrium sucrose gradients (Fig. 2a) (1, 48). Lipids that are tightly packed in ordered regions of membrane are less susceptible to solubilization by small amounts of nonionic detergents than more disordered lipid regions (49). DRMs are also sensitive to the level of cholesterol, which is an abundant lipid in mammalian plasma membranes that comprises ~30–40 molar percent of total plasma membrane lipids (49). Cyclodextrins, which rapidly remove cholesterol from cellular membranes, are used to extract cholesterol and perturb raft association and, thereby, MIRR signaling (50). DRMs do not correspond directly to pre-existing lipid rafts in cell membranes. Because detergents can cause reorganization of membrane lipids, isolated DRMs do not strictly measure
Membrane Compartments Related to Signaling and Trafficking in Immune Cells

(lipid and protein associations that occur prior to detergent treatment (5)). However, protein association with DRMs provides a correlative indication of association with lipid rafts in cell membranes (48, 51). Cholesterol plays multiple roles in cells, and cholesterol depletion can alter the lateral mobility of both raft and non-raft proteins as well as affecting association of the actin cytoskeleton with the plasma membrane (3, 6, 24).

Centrifugation in a dense medium is a common method for separating organelles and membranes from the postnuclear supernatant. Differential centrifugation separates organelles based on size and weight; larger, heavier membranes pellet at slower speeds than smaller, lighter membranes. Successive centrifugation at faster speeds fractionates the sample into cellular components. Equilibrium, or isopycnic, centrifugation separates organelles based on their buoyant density independent of their size and shape. Cellular components are separated on a discontinuous gradient, and fractions float in the medium density that matches their own buoyant density. Sucrose is often used to create discontinuous gradients, but the relative separation of organelles and their distribution within the gradient will vary depending on the gradient medium and the method of sample loading.

Subcellular fractionation

Subcellular fractionation separates organelles based on their biophysical and biochemical properties. Several methods exist for cellular disruption, which include homogenization, sonication, and nitrogen cavitation (47). Homogenization shears a sample by forcing suspended cells through a narrow space and can result in membrane fragments of various size caused by variability in shear forces. When smaller cells, such as lymphocytes, are homogenized, the nuclei may be fragmented along with the plasma membrane. Sonication shears samples by using high-frequency sound waves to disrupt membranes, which nonselectively disrupts intracellular membranes as well as the plasma membrane and can damage membranes because of production of local heating. Nitrogen cavitation ruptures suspended cells, which are first equilibrated with high pressures of an inert gas, during rapid return to atmospheric pressure. This method fragments the ER and plasma membrane into uniform vesicles under inert conditions, which minimizes sample damage. After disruption, the nuclei and large cellular debris (nuclear pellet) are separated from the other organelles and cytosolic components (postnuclear supernatant) by a low-speed centrifugation step. Each of these components can be purified for analysis.

Fluorescence spectroscopic methods

Temporal and spatial changes of fluorescently tagged proteins can be monitored in live cells under physiological conditions using a variety of spectroscopic methods that allow for the quantification of dynamic protein and membrane trafficking responses. Increasing capabilities to monitor protein translocation, protein–protein interactions, lipid turnover, and membrane trafficking will provide a more complete picture of the large-scale reorganization in cells that occurs during antigen receptor stimulation.

Fluorescence microscopy

Direct visualization of fluorescently tagged proteins continues to be a powerful method for examining subcellular localization and dynamics within a single cell. Confocal microscopy, in which only a thin, optical section (~0.3 μm) of the sample is imaged, is
commonly used to study protein localization in both fixed and living samples. The development of advanced imaging tech-
niques at the micron- and submicron-scale resolution have been essential to monitoring small domains of the plasma membrane and intracellular organelles. A more-detailed discussion of ad-
vanced imaging techniques such as fluorescence recovery after photobleaching, single particle tracking, and fluorescence corre-
lution spectroscopy used to study protein dynamics is provided elsewhere (52). A additionally, ultra high-resolution techniques, such as photoactivated localization microscopy and stimulated emission depletion microscopy are being developed to investi-
gate plasma membrane dynamics (53, 54, 55).

Protein colocalization is qualitatively assessed by overlaying individual confocal images and looking for overlapping protein distributions, which is an approximate measurement of protein colocalization on the scale of optical resolution (~300 nm). Alternatively, a structure can be manually identified, and the fluorescence intensity profiles of two proteins along the trace can be correlated. This approach has been used to quantify receptor-protein interactions in mast cells (39, 56).

Correlation methods

Several image correlation techniques, termed image correlation microscopy or image correlation spectroscopy (ICS), have been developed to analyze the density, diffusion, velocity, and inter-
actions of fluorescently labeled membrane proteins. Spatial ICS measures the number and density of protein aggregates as well as changes in aggregation state, whereas temporal ICS measures diffusion coefficients and flow rates (57, 58). A recent exten-
sion of these techniques, which is termed spatio-temporal image correlation spectroscopy, can be used to monitor interactions be-
tween two fluorescently labeled proteins and also enables vector mapping of directed protein movement (58) even when a large percentage of the protein population is immobile (59). Raster image correlation spectroscopy can provide diffusion rates for both fast moving cytosolic species and slow moving membrane species (60).

Another recently developed cross-correlation microscopy methodology analyzes protein redistribution that occurs at the plasma membrane by imaging equatorial sections of the cell (61). This automated methodology allows for rapid analysis of a time-lapse image series in which a plasma membrane trace, defined with a fluorescent probe, is generated for each time point and, therefore, it can accommodate small changes in cell morphology (Fig. 3a). The stimulated recruitment of cyto-
lolic proteins to the membrane (Fig. 3b), and their interactions with aggregated receptors (Fig. 3c) are quantified as a function of time. This methodology was used to analyze early signal-
ing events in FcεRI signaling (61). Because events that occur at the plasma membrane are selectively analyzed, stimulated membrane trafficking events to and from the plasma membrane can also be studied with this approach.

Antigen-patterned surfaces

Mast cells, in addition to responding to soluble antigen, can also polarize responses to the cell-pathogen interface when stimulated by helminths or large pathogens. Patterned silicon surfaces, in which the antigen is deposited on a silicon chip in micron-sized features (Fig. 4a), are used to study protein traf-
ficking and membrane reorganization events that occur at the cell-stimulus interface (Fig. 4b). These surfaces have been used to investigate protein recruitment to aggregated FcεRI com-
plexes in mast cells and demonstrate differential redistribution of inner-leaflet and outer-leaflet markers (62). The differential trafficking of recycling endosomal membranes, which are traf-
ficked to sites of aggregated receptors, and histamine-containing secretory granules, polarized to the interface but not toward ag-
ggregated receptors, was examined using antigen-patterned sur-
faces (19).

A antigen-patterned surfaces have also been used to study the formation and regulation of the immunological synapse
in T cells (63). Patterned surfaces with micron-sized corrals have been used to investigate the dynamics of immunological synapse formation by preventing central clustering of TCR ligands (64). Patterns that prevented TCR microclusters from reaching the central synapse enhanced T cell signaling, which suggests that spatial translocation of TCR microclusters to the central synapse is a method to downregulate TCR signaling (64). The role of peripheral TCR microclusters in T cell signaling can be further investigated with new technologies to produce sub-micron scaled patterns (63).

Spectroscopic methods to study membrane trafficking

The organelles of the endocytic pathway are characterized by progressively more acidic structures as endocytosed material moves through early endosomes, late endosomes, and sent to either acidic lysosomes for degradation or to recycling endo-
somes for return to the plasma membrane. Fluorescent labels that are sensitive to pH, such as pH-sensitive GFP derivatives (pHluorins) or fluorescein isothiocyanate (FITC), can be used to monitor changes in the trafficking along the endocytic and exocytic pathways (65). A "targeted fluorescence" strategy to characterize pH gradients along the endocytic pathway takes advantage of pH-sensitive changes in FITC fluorescence (66). Fluorophore pH-sensitivity can be used to quantify the traf-
ficking of internal membranes to and fusion with the plasma membrane. FITC fluorescence is quenched in endocytic or-
ganelles, but fluorescence increases after exposure of the FITC-tagged protein to more pH-neutral environments. A traf-
ficking assay has been developed to characterize the stimulated outward trafficking of recycling endosomes labeled with either FITC-CTxB or FITC-anti-transferrin receptor in mast cells (18). This assay can be used to screen for inhibitors of recycling endo-
somal trafficking and to elucidate the mechanisms that regulate trafficking of raft, CTxB-labeled endosomes, versus trafficking of nonraft, transferrin receptor-containing endosomes.

Summary

The redistribution of plasma membrane lipid rafts from small, dynamic entities to stable signaling complexes represents one example of stimulated membrane reorganization central to MIRK signaling. However, changes in membrane compartmen-
talization necessary for sustained immune cell signaling are not

---

Membrane Compartments Related to Signaling and Trafficking in Immune Cells

MIRR signaling. However, changes in membrane compartmentalization necessary for sustained immune cell signaling are not.
restricted to the plasma membrane; the membrane organization of intracellular organelles such as the ER and vesicles along the endocytic and exocytic pathways contribute to the functionality of these organelles and also to changes that occur at the plasma membrane. Biochemical methods to characterize membrane subdomains, such as detergent resistance, cholesterol depletion, and subcellular fractionation, used in combination with imaging techniques to study protein and membrane dynamics in living cells, are essential to understanding the role and regulation of membrane compartmentalization in immune cell signaling and trafficking.

Acknowledgements

Our research described in this review was supported by the Nanobiotechnology Center (NSF: ECS-050623) and by NIH grants: R01-AI2244, R01-AI18306, and T32-GM 08210.

References

Membrane Compartments Related to Signaling and Trafficking in Immune Cells

(a) Parylene layer on silicon chips is patterned with micron-sized features
(b) Alexa488-IgE

Figure 4 Antigen-patterned surfaces to study cellular activation. (a) A lipid mixture that contains antigen is patterned onto silicon surfaces in micron-sized features (62). (b) Mast cells that were labeled with Alexa488-IgE (top) show clustering over the antigen clusters (bottom).


Membrane Compartments Related to Signaling and Trafficking in Immune Cells


Further Reading


See Also

Cell Membranes, Dynamics of Imaging Techniques for Proteins Lipid Domains, Chemistry of Membrane Trafficking


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
In the popular fluid mosaic model for biomembranes, membrane proteins and other membrane-embedded molecules are in a two-dimensional fluid formed by the phospholipids. Such a fluid state allows free motion of constituents within the membrane bilayer and is extremely important for membrane function. The term “membrane fluidity” is a general concept, which refers to the ease of motion for molecules in the highly anisotropic membrane environment. We give a brief description of physical parameters associated with membrane fluidity, such as rotational and translational diffusion rates, order parameters etc., and review physical methods used for their determination. We also show limitations of the fluid mosaic model and discuss recent developments in membrane science that pertain to fluidity, such as evidence for compartmentalization of the biomembrane by the cell cytoskeleton.

Introduction

In 1972, Singer and Nicolson (1) suggested the so-called fluid mosaic model of the biological membrane (Fig. 1) (2). This useful hypothesis explained many phenomena that occur in model and biological membranes. According to this model, membrane proteins and other membrane-embedded compounds are suspended in a two-dimensional (2-D) fluid formed by phospholipids. The phospholipids are assumed to be in a liquid state, so they are capable of rapid diffusion within their layer and are in constant motion. This fluid state of membrane lipids is critical for membrane function. It allows, for example, free diffusion and equal distribution of new cell-synthesized lipids and proteins, lateral diffusion of proteins and other molecules in signaling events and other membrane reactions, membrane fusion (i.e., fusion of vesicles with organelles), separation of membranes during cell division, and so on.

Generally speaking, two principal mechanisms operate in the biology of membrane processes, such as membrane transport and permeation. One involves a network of active sites and operates by metabolic energy, and it is referred to as active; another is directed by passive diffusion, and it is called passive. This passive mechanism is determined by various aspects of lipid dynamics and lipid-protein interactions, and it can be described in quantitative terms of chemical and phase equilibrium and molecular physics. However, in the highly anisotropic membrane environment, some aspects must be generalized and redefined compared with the simple isotropic case. One term, which received a new interpretation in the context of membrane science, is “fluidity.” Membrane fluidity, which describes the ease of movement for molecules in the membrane environment, is a general concept that lacks a precise definition. It is much broader than the strict physical definition of fluidity as the reciprocal of viscosity in the case of isotropic liquids. In general, “membrane fluidity” implies various anisotropic motions, which contribute to the mobility of components of a membrane. It includes lateral diffusion of molecules in the plane of the membrane, flexibility of acyl chains, “flip-flop” diffusion of molecules from one monolayer to the other, and so on. The most important parameters to quantify the notion of membrane fluidity are translational and rotational diffusion constants, order parameters (or tensors), packing, and permeability. In general, greater membrane fluidity is associated with higher diffusion rates, high permeability, lower ordering, and looser packing. The relation between the parameters, however, is purely empirical, and in most cases speculative. Many membranes have large diffusion constants and large order parameters and vice versa.

The lipid membrane, as a whole, shows a unique combination of fluidity and rigidity. In terms of the solubility and the diffusion of small nonpolar molecules, the membrane behaves very much like an oil drop. In contrast, the translational diffusion...
constants of lipids and proteins in membranes are characteristic of media with the viscosity over two orders of magnitude greater than that of oil, such as hexadecane. Also, in most cases the membrane represents an impermeable barrier for ions and other hydrophilic compounds.

### Physical Parameters Associated with Membrane Fluidity

#### Diffusion constants

Diffusion is the random movement of a particle because of an exchange of thermal energy with its environment. Membrane lipids and proteins participate in highly anisotropic translational and rotational diffusion motion. Translational diffusion in the plane of the membrane is described by the mean square lateral displacement after a time $\Delta t$: \( \langle r^2 \rangle = 4 D_{ll} \Delta t \). Lateral diffusion coefficients in fluid phase bilayers are typically in the range $D_{ll} \sim 10^{-8}$ to $10^{-7}$ cm$^2$/s.

Rotational diffusion is characterized by the mean square angular deviation during the time interval $\Delta t$: \( \langle \theta^2 \rangle = 60_s \Delta t \). Highly anisotropic motion, which is typical for lipid molecules in the membrane, is usually described by two rotational diffusion coefficients $D_{\parallel}$ and $D_{\perp}$, which correspond to diffusion about the long diffusion axis and perpendicular to it, respectively. The diffusion coefficients are related to corresponding rotational correlation times measured by nuclear magnetic resonance (NMR), electron spin resonance (ESR), fluorescent depolarization, and so on, as:

\[
\tau_{\parallel} = \frac{1}{6} D_{\parallel}, \quad \tau_{\perp} = \frac{1}{6} D_{\perp}.
\]

For fluid phase bilayers, the typical rotational diffusion coefficients are of the order of $D_{\parallel} \sim (1-4) \times 10^{-8}$ s$^{-1}$ and $D_{\perp} \sim (1-4) \times 10^{-7}$ s$^{-1}$.

#### Flip-flop diffusion

Lipid molecules, in principle, can exchange between the two monolayers of the bilayer. For polar lipids, it is an extremely slow process with characteristic times of hours or even days. For membrane proteins, no appreciable flip-flop mobility has yet been observed, which is in good accord with the fact that inner and outer leaflets of natural membranes are usually asymmetric with respect to their protein and lipid composition. On the other hand, cholesterol has a relatively high rate of spontaneous flipping between two membrane leaflets ($t_{\text{flip}} \sim 1$ s).

#### Order parameters

The membrane lipid layer is a lamellar phase with the preferred orientation of the lipid molecules perpendicular to the membrane plane. By definition, if $\hat{n}$ is the angle of the long molecular axis with respect to the bilayer normal, then the order parameter $S$, which is a measure of the orientation distribution, is given as the average of $P_2(\cos \theta)$, the second Legendre polynomial: $S = \frac{1}{2} (3 \cos^2 \theta - 1)$.

One can see that $S$ varies between $-1/2$ and 1. These limiting cases have the following meaning: when $S = 1$, all molecules are exactly perpendicular to the membrane plane. When $S = -1/2$, $\theta = 90^\circ$, and all molecules have their long axis parallel to the membrane surface. The case of $S = 0$ usually corresponds to a random distribution of molecular axes relative to the membrane normal. In energy terms, the rotation of the molecular long axis
in the liquid crystal is restricted within an orienting potential that is simply approximated as: $U(\theta) = \lambda \cos^2 \theta$, where $\lambda$ is the strength of the potential. The ordering of the lipid chain relative to the bilayer normal can then be expressed \((7)\) as:

$$S_\theta = \frac{1}{2} \left[ 3 \cos^2 \theta - 1 \right] \exp \left[ \frac{U(\theta)}{kT} \right] \sin \theta d\theta$$

Permeability

The passive permeability of lipid membranes is another fluidity related parameter. In general, two mechanisms of membrane permeability can operate in the membrane (8). For many nonpolar molecules, the predominant permeation pathway is solubility-diffusion, which is a combination of partitioning and diffusion across the bilayer, both of which depend on lipid fluidity. In a few cases, such as permeation of positively charged ions through thin bilayers, an alternative pathway prevails (9, 10). It is permeation through transient pores produced in the bilayer by thermal fluctuations. This mechanism, in general, correlates with membrane fluidity. However, for model membranes undergoing the main phase transition, permeation caused by this mechanism exhibits a clear maximum near the phase transition point (11).

---

**Phase State and Membrane Fluidity**

A remarkable property of lipid bilayers is their structural phase transitions (thermotropic polymorphism). For example, fully hydrated pure diacyl-phosphatidyl cholines exist in one fluid phase, $L_\alpha$, and three crystalline phases: P$_{α}′$, L$_{α}′$, and L$_β′$ (12). Because of the high degree of disorder caused by defects, the P$_{α}$ and L$_{α}$ phases usually are called gel phases. The P$_{α}$ phase is sometimes called a "ripple phase," because the surface of the bilayer is rippled (13) and presents a wave-like appearance in electron micrographs (Fig. 2). Depending on the nature of the lipid and the presence of additional components (cholesterol etc.), the P$_{α}$ phase may be present or absent in the phase diagram, and a tilted gel $L_{α′}$ could be replaced by the L$_{α}$ phase, which has similar physical properties but no tilt of the hydrocarbon chains.

In the gel phase, lipid chains are usually well aligned with little rotation around the C-C bonds that are predominantly in the trans position. The lipid chains are tightly packed, the chain ordering is high, the bilayer thickness is maximal, and the surface area per lipid headgroup is relatively small. Lipids in the gel phase in most cases can be handled as solids. For example, aligned thick multibilayers in the gel phase can be sliced in thin pieces in an arbitrary direction so that each piece retains macroscopic alignment (14). The physiological importance of gel-like phases is limited.

At the "main transition" temperature, $T_{m}$, the gel phase undergoes a transition to the $L_{β′}$ (liquid crystal) phase. At the transition point, the surface area increases (15), and the bilayer thickness and chain order decrease (16). In the fluid phase, hydrocarbon chains tend to contain more gauche isomers (17). At this transition, the DSC (differential scanning calorimetry) shows a sharp peak in the heat capacity that occurs over a very narrow temperature range. The transition between $L_{α}$ and P$_{α}$ phases also can be detected by DSC and is called the pretransition (18). The transition between $L_{α}$ and L$_{α′}$ phases is in most cases hard to observe because of typical supercooling of the L$_{α}$ phase. Depending on the membrane composition, hydration, and temperature, several 2-D and three-dimensional nonlamellar lipid phases are possible, including the well-studied hexagonal (H$_{β}$ and H$_{α}$) and cubic phases (19).

Natural biomembranes contain a complex mixture of various phospholipids with cholesterol and sphingomyelins. In general, they exist in the fluid phase. Maintaining membrane fluidity seems to be extremely important for the survival of the cell and the whole organism. It is well known for model membranes that a decrease in the chain length or the introduction of unsaturation into the hydrocarbon chain causes a decrease in the main transition temperature. Consistent with this observation, microorganisms, plants, and animals (pikoklothers or hibernating mammals) are acclimated to low temperatures by altering their membrane lipid composition, increasing the degree of lipid unsaturation, or decreasing the average chain length (20–23).

In an attempt to relate natural and model membranes, many lipid mixtures have been examined experimentally. It has been shown that additional components broaden the main phase transition, with a wider temperature range of coexisting gel and liquid phases (24). Another important feature of lipid mixtures is the formation of nonideal solutions with nonzero enthalpy.
Membranes, fluidity of and/or entropy of mixing. It often makes the components completely or partially immiscible in one or both phases and manifests itself in complex phase diagrams (25).

Cholesterol, which is an important constituent of cell membranes, plays an important role in maintaining membrane fluidity. It effectively inhibits the transition to the gel phase (26, 27). Even though some plasma membranes, such as nerve myelin membranes, contain a high concentration of lipids that form gel phase bilayers, the presence of cholesterol keeps these membranes in a fluid phase. However, interaction with the rigid cholesterol ring affects hydrocarbon chains of lipids in the liquid crystal phase (Lα) and leads to formation of a new phase, the liquid ordered (Lo) phase (27). This phase is well characterized by a variety of physical methods and does not exist in pure lipids or their mixtures. In the liquid ordered phase, the long axis rotation and lateral diffusion rates are similar to the Lα phase, but the acyl chains are predominantly in an all-trans conformation and, hence, the order parameters are similar to the Lα phase (see Table 1). Recently, the cholesterol-rich Lo phase has been strongly associated with microdomains in live cells—the so-called "lipid rafts."

### Physical Methods for Measuring Fluidity Parameters

In the membrane environment, a wide range of motions has been observed and studied by a variety of physical methods. The characteristic times of the motions span 20 orders of magnitude, from about $10^{-14}$ s for molecular vibration to days for transbilayer flip-flops. Figure 3 shows characteristic frequencies (reciprocal of characteristic times) of different kinds of molecular motions in the membrane in comparison to frequency ranges in which various spectroscopic techniques are sensitive to molecular motion (28).

A spectroscopic technique that probes membrane fluidity can either directly measure mobility and order parameters for membrane constituents (NMR) or use probes (ESR, fluorescence). Some fluorescent and ESR probes are shown in Fig. 4. The connection between the rotational correlation time of a membrane embedded probe and the membrane fluidity can be illustrated using the example of a simple isotropic liquid, in which fluidity is merely a reciprocal viscosity $\eta$ and the rotational correlation time $\tau$ for a molecule with a hydrodynamic volume $V$ is given by the well-known Debye–Stokes–Einstein relation:

$$\tau = \frac{V}{6\pi \eta} \frac{1}{kT},$$

where $k$ is the Boltzmann constant and $T$ is the temperature.

### Table 1: Translational diffusion coefficients of lipids and order parameters in some membrane phases

<table>
<thead>
<tr>
<th>Phase</th>
<th>$D_T$ (cm$^2$/s)</th>
<th>$S^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liquid-crystalline (Lα)</td>
<td>$10^{-9}$</td>
<td>0–0.2</td>
</tr>
<tr>
<td>Gel (Lα)</td>
<td>$10^{-16}$</td>
<td>0.2–0.9</td>
</tr>
<tr>
<td>Gel (Pβ)</td>
<td>Similar to Lα, but the bilayer is rippled</td>
<td></td>
</tr>
<tr>
<td>Gel (Lβ)</td>
<td>The same as Lα, but the chains are tilted 32°</td>
<td></td>
</tr>
<tr>
<td>Liquid-ordered (Lo)</td>
<td>$10^{-16}$</td>
<td>0.2–0.9</td>
</tr>
</tbody>
</table>

*Measured by NMR

**Figure 3** The characteristic frequencies of molecular motions of membrane proteins and lipids compared with the frequency ranges in which various spectroscopic techniques are sensitive to molecular motion. (Modified from Reference 28.)
Membranes, Fluidity of

Fluorescent probes and labels

- Diphenylhexatriene (DPH)
- (m, n)-anthroylstearic acid
- Dansyl chloride
- NBD-phosphatidylcholine

A triplet label for membrane proteins
- Eosine-malemide

Nitroxide spin labels for ESR
- TEMPO
- 3d-DOXYL-Sw-cholestanec (CSL)

PC spin labels
- MTSL, cysteine-specific spin label

Figure 4 Probes for fluorescence studies and ESR spectroscopy.
absolute temperature, the rotational diffusion coefficient \( D_\theta \) is given by \( D_\theta = \frac{1}{6} \nu_\theta \).

In the highly anisotropic membrane environment, one can expect several different correlation times that correspond to the anisotropic membrane environment and/or the nonspherical molecular probe as well as variations that exist along the membrane normal within the bilayer. Furthermore, the molecular motion is often limited by constraints imposed by the ordered surroundings of the probe. Properly designed spectroscopic experiments can, in many cases, extract both mobility and order parameters and can give a comprehensive picture of membrane fluidity.

**Fluorescent experiments for measuring molecular mobility and ordering**

A absorption of a photon instantaneously brings the probe molecule to the first excited singlet state \( S_1 \). Usually it takes \( 10^{-8} \) seconds to return to the ground state. It can occur via collision with neighbors and loss of the energy as heat or through the emission of a photon (fluorescence). The fluorescence lifetime \( \tau_\parallel \) is the characteristic time for the population of excited molecules to return to the ground state after a flash of excitation light.

The fluorescence depolarization technique for mobility and ordering is based on the fact that the probability of absorption and emission is directional. Light polarized along a certain axis will preferably excite molecules oriented with their transition dipole moment in the same direction. The probability varies with \( \cos^2 \theta \), where \( \theta \) is the angle between the transition dipole moment and the electric field vector of the light. Emission of a photon obeys the same \( \cos^2 \theta \) \((28)\) rule. That means that a molecule oriented with its transition dipole moment along the \( Z \)-axis will be likely to emit a photon with the same polarization. In the depolarization technique, polarizers are used to quantify the intensity of the parallel (\( I_\parallel \)) and perpendicular (\( I_\perp \)) components to the original direction of polarization.

The values of \( I_\parallel \) and \( I_\perp \) are obtained after exciting the anisotropy with light of parallel polarization in a random direction, \( I_\parallel > I_\perp \), and \( \tau \) has a maximal value equal \( 0.4 \). In the opposite limit \( \tau \gg \tau_\parallel \), and isotropic rotation) the molecules will emit in a random direction, \( I_\parallel = I_\perp \), and \( r \) has a minimal value equal \( 0.0 \). In the intermediate range, for \( \tau_\parallel < \tau < \tau \), the value of \( r \) is sensitive to molecular motions.

An important advantage of the depolarization technique is that it allows one to measure the molecular ordering, as well as the motional parameters. For this purpose, it is necessary to detect the time dependence of the anisotropy. In the presence of ordering constraints, the value does not decay to zero, but to some limiting value \( r_{lim} = r_{lim} = \frac{1 - 10^{-4} \nu_\parallel}{1 - 10^{-4} \nu_\perp} \). The rate of decay defines a rotational correlation time, and \( r_{lim} \) is a direct measure of the order parameter through the following relation: \( r^2 = r_{lim} \tau_\parallel \). The fluorescence depolarization method works well as long as fluorescence lifetimes, which are typically \( 10^{-8} \) s, are not too different from the rotation relaxation times to be measured. When the rotational correlation time is slower than about \( 10^{-8} \) s, however, the method fails because the fluorescent emission decays before any detectable rotation can occur. For studying rotational diffusion of membrane proteins, it is almost always the case that correlation times are in the microsecond time range or longer, because membranes are at least 100 times more viscous than water. Several phosphorescent probes (e.g., derivatives of eosin) were developed to measure such slow rotation correlation times. In such molecules, after initial excitation, transition into the lowest triplet state \( T_1 \) (intersystem crossing) effectively competes with \( S_1 \rightarrow S_0 \) transition. Because the \( T_1 \rightarrow S_0 \) transition is spin forbidden, the lifetime of the lowest triplet state (typically \( 10^{-3} \) s) is much longer than that of the \( S_1 \) (typically \( 10^{-5} \) to \( 10^{-6} \) s).

**Solid state NMR**

Solid state NMR is characterized by relatively broad lines. It can be described as magnetic resonance on molecules frozen in solids or with very slow rotational correlation times. It is appropriate not only for solids, but also for viscous solvents, lipids in membranes, or large macromolecules in solution. High resolution (solution) NMR deals with relatively small molecules with molecular weight less than about 50,000 Da in aqueous solution. The rotation of such molecules is so fast on the NMR time scale that it effectively averages out all orientational anisotropy yielding an isotropic spin Hamiltonian including the Zeeman term and the \( J \)-\( J \) coupling with another spin \( S \): \( H = \gamma B_0 \hat{I}_0 \). Here, \( I \) and \( S \) are spin operators for two nuclear spins. \( B_0 \) is the external magnetic field strength, \( \gamma \) is the magnetogyric ratio for the studied nucleus, \( \sigma \) is the chemical shielding constant, which is directly related to the chemical shift \( \delta \); \( J \) is the spin-spin coupling constant for spins \( I \) and \( S \). For no or very slow rotational reorientation the spin Hamiltonian requires a tensorial expression: \( H = \gamma I \cdot B_0 + (1/2) \sum_{s=1}^{3} J_{s} S_{s} \) where

\[
\delta = \begin{pmatrix}
\sigma_{XX} & \sigma_{XY} & \sigma_{XZ} \\
\sigma_{YX} & \sigma_{YY} & \sigma_{YZ} \\
\sigma_{ZX} & \sigma_{ZY} & \sigma_{ZZ}
\end{pmatrix}
\]

and \( J = \begin{pmatrix}
J_{11} & J_{12} & J_{13} \\
J_{21} & J_{22} & J_{23} \\
J_{31} & J_{32} & J_{33}
\end{pmatrix} \)

are chemical shielding and spin-spin coupling tensors.

Tensor values should also be assigned to quadrupole splittings (see below).

The tensor \( \delta \) is usually symmetrical, and in the appropriate molecular frame it is represented in a diagonal form:

\[
\delta = \begin{pmatrix}
\sigma_{XX} & 0 & 0 \\
0 & \sigma_{YY} & 0 \\
0 & 0 & \sigma_{ZZ}
\end{pmatrix}
\]

The components \( \sigma_{XX}, \sigma_{YY}, \) and \( \sigma_{ZZ} \) give chemical shifts along main molecular axis in \( X-, Y-, \) and \( Z- \) direction, respectively. When the molecule is oriented along the main axis, for example the \( Z \)-direction, the chemical shift is just \( \sigma_{ZZ} \). Other orientations give intermediate values. In the limiting case of absence of molecular motion (rigid limit) and a statistical distribution of molecular orientations, for example polycrystalline or glass, the spectrum shows a superposition of spectra for all orientations.
Membranes, Fluidity of

Phosphorus 31 has spin $I = \frac{1}{2}$. The chemical shift tensor is substantially anisotropic. For static, statistically disordered molecules, the spectrum is a superposition of spectra that correspond to different orientations. If the molecule undergoes fast rotation about one axis, then the motion averages the components perpendicular to that axis. In the case of fast isotropic rotation (limiting case of high resolution NMR), a single sharp line is observed in the spectrum (Fig. 5). For example, the rotation of the vesicle depends on its size and the viscosity of the environment. The smaller the vesicle, the faster is the rotation. Very fast isotropic tumbling of the vesicle gives a single isotropic peak. For large vesicles, however, the spectrum corresponds to anisotropic rotation of the lipid molecule itself about the main axis.

Diffusion along the membrane surface affects the averaging of tensor components, especially if the surface is curved. Lipids can form other aggregates than bilayer membranes. For instance, a common form is the inverse hexagonal phase (HII--phase). It consists of cylindrical lipid tubes with the lipid hydrocarbon chains on the outside and directed radially and the water phase in the center. Diffusion of lipids leads to a different type of anisotropic averaging of the tensor components than in the case of large vesicles. Figure 6 (33) shows $^{31}$P-NMR spectra from lipids in the bilayer membrane, the HII phase and the isotropic case. If a lipid system undergoes the L$_\alpha$–HII transition, then the $^{31}$P-NMR lineshape abruptly changes at the transition temperature as shown in Fig. 6.

Use of $^2$H-NMR for membrane studies is based on the fact that deuterium nuclei, with spin $I = 1$, have an electric quadrupole moment. It originates from the asymmetrical charge distribution in the nucleus. In the presence of an external field gradient, which is always almost the case for (deuterium) atoms in molecules, the different orientations of the nuclear spin experience different interaction energies with the quadrupolar field of the environment.

For a system with spin 1 and the absence of any quadrupolar interaction, one has three spin energy levels in the external magnetic field for $m = -1, 0, +1$. They have the same energy splitting between adjacent levels, which yield only one resonance line. Through the quadrupole interaction, the energy levels are no longer equidistant, and one observes two resonance lines. If the quadrupolar splitting is much less than the Zeeman splitting, then the approximate expression for a fixed orientation of the molecule is (34):

$$
\Delta \nu = \frac{3}{8} \left( \frac{\epsilon^2 Q}{h} \right) (3 \cos^2 \phi - 1),
$$

Where $\epsilon$ is the quadrupole coupling constant, $Q$ is the quadrupole moment of the nucleus, and $h$ is Planck’s constant.

**Figure 5** $^{31}$P-NMR spectra of phospholipids with different modes of molecular motion. a: rigid limit; b: fast axial rotation about X-axis averages Y- and Z-components of the chemical shift tensor. In the spectrum one can observe two principal values of $\sigma_\perp = 1/2(\sigma_{yy} + \sigma_{zz})$ and $\sigma_\parallel = \sigma_{xx}$; c: atypical case of fluid membrane. Along with fast X-axial rotation molecular motion also partially averages $\sigma_\parallel$ and $\sigma_\perp$. In the spectrum one observes effective values $\sigma'_\parallel < \sigma_\parallel$ and $\sigma'_\perp < \sigma_\perp$. d: Isotropic case (high resolution NMR): $\sigma_{iso} = 1/3(\sigma_{xx} + \sigma_{yy} + \sigma_{zz})$.
Membrane fluidity of

\[ \text{Figure 7} \] 2H-NMR spectra at 23.3 MHz of 1-[16,16,16-2H] palmitol,2-palmitoleoyl-PC at different temperatures. (From Reference 30.)

where \( q \) is the quadrupole moment, \( e_q \) is the magnitude of the principal component of the gradient of the electric field, and \( \phi \) is the angle between the direction of that component and the external magnetic field. The two resonance lines are split by a frequency of \( \Delta \nu \). For the special case of \( \phi = 54.7^\circ \) (magic angle), \( \Delta \nu = 0 \) and no quadrupole splitting occurs. A superposition of all orientations gives a sum of two broad spectra that are symmetrical about the central resonant frequency, which is a lineshape similar to a Packe doublet for two interacting protons in a single water molecule.

For fast rotation (the case of high resolution NMR), the quadrupolar splitting is averaged out yielding a single resonance line. For intermediate cases, the splitting and the line shape of the quadrupolar NMR signal is indicative of the rotational correlation time and/or ordering effects (Fig. 7) (30).

ESR spectroscopy

Although theoretically NMR can obtain both molecular dynamics and ordering, a close look at the literature shows that in membranes, NMR is used mostly for extracting order parameters. Much information on rotational mobility of membrane constituents is traditionally obtained using another magnetic resonance technique, ESR. ESR is extremely useful in the study of membrane fluidity, because of its unique time scale, which spans almost all motional range in membranes.

For studies of membrane fluidity, lipids or membrane proteins are usually spin-labeled with cyclic nitroxide radicals. The theoretical analysis of the spin Hamiltonian of ESR is analogous to that for NMR, although the interaction terms are much larger. The position of ESR lines for nitroxides is determined by the \( g \)-factor and the hyperfine splitting, which roughly correspond to the chemical shift and \( J \)-coupling in NMR, and the sensitivity of ESR spectroscopy to molecular motion emerges from the dependence of these parameters on the orientation of the nitroxide moiety in the magnetic field.

The ESR-Hamiltonian is given as

\[ H = \beta S \cdot gB_0 + \gamma I \cdot A, \]

where \( S \) and \( I \) are spin operators for electrons and nuclei, \( \beta \) and \( \gamma \) are principal values of the \( g \)-tensors and the hyperfine splitting tensors \( A \), respectively. At the fast motion limit, one can observe a narrow triplet centered around the average \( g \) value \( (g_{xx} + g_{yy} + g_{zz})/3 \) with a distance between lines of \( \alpha_{iso} = (A_{xx} + A_{yy} + A_{zz})/3 \), where \( g_{ii} \) and \( A_{ii} \) are principal values of the \( g \)-tensor and the hyperfine splitting tensor \( A \), respectively. At the slow motion limit, which is also referred to as the rigid limit, the spectrum (shown in Fig. 8) is a simple superposition of spectra for all possible spatial orientations of the nitroxide with no evidence of any motional effects. Between these limits, the analysis of the ESR lineshape and spectral simulations, which are based on the Stochastic Liouville Equation, provide ample information on lipid/protein dynamics and ordering in the membrane (36).

\[ \text{Figure 8} \] ESR spectra of TEMPO nitroxide radical in glycerol at different temperatures. (From Reference 35.)
For the most common ESR frequency of 9 GHz, the limits are approximately from $10^{-7} < \tau_c < 10^{-11}$ seconds. However, using a range of ESR frequencies and pulse and continuous wave (CW) ESR techniques (such as ELDOR or saturation transfer ESR) expands the range from $10^{-8}$ to $\sim 10^{-12}$ seconds, which covers virtually all modes of molecular motion in the membrane that can be associated with membrane fluidity (37). In the past, to study the range of $10^{-7} - 10^{-4}$ seconds, an empirical technique known as saturation transfer ESR was used. It is based on detection of an out-of-phase ESR signal with large modulation amplitude (typically 5G) of the external magnetic field (38). The most direct way to observe the “saturation transfer” is ELDOR or electron-electron double resonance, when two different microwave frequencies are applied to the sample. One is called the “pump” frequency and used to modify the populations by saturating certain portions of the ESR spectrum. Another one (the “probe” frequency) is used to observe the effect of the “pump” frequency on other parts of the spectrum. The response to an intense (saturating) microwave field is affected by diffusion of saturation (“saturation transfer”) between different portions of the resonance spectrum. In the case of nitroxide labels, this diffusion is dominated by rotational modulation of the anisotropic magnetic interactions. The method is most sensitive when the rotational correlation time is comparable with or greater than the spin lattice relaxation time, which for nitroxides near the rigid limit is $\sim 10^{-5}$ seconds. In the CW mode, usually the intensity of the ESR signal at the “probe” frequency is recorded as a function of the “pump” frequency. A huge improvement has been achieved by using the Pulsed Two-Dimensional Fourier Transform technique, which is known as 2-D ELDOR. This experiment very efficiently provides all combinations of pump and probe frequencies in a single 2-D experiment, and in addition it dramatically improves the spectral resolution, which enables accurate assessment of mobility and ordering parameters. As observed in Fig. 9 (39), 2-D ELDOR spectra show dramatic changes as membrane properties are varied.

Figure 9 2D-ELDOR spectra for DPPC membranes containing spin-label 16PC (cf. Fig. 4) with and without gramicidin A at different temperatures. (From Reference 39)

Membranes, Fluidity of
Membranes, Fluidity of

Lateral Diffusion in Membranes

As we will see later, studying lateral diffusion in biomembranes gives important insights into their structure and function. Methods for measuring diffusion coefficients for lipids in membranes can be classified into two distinct categories. In the first category, which corresponds to short-range diffusion measurements, \( D \) is obtained from determination of frequencies of bimolecular collisions within the membrane, through fluorescence, ESR, and so on. In the second category, which corresponds to long-range diffusion measurements, \( D \) is usually determined from the time required to fill a defined region of the membrane.

Short-range lateral diffusion

ESR is very useful in the study of dynamic properties of membrane components because of its high sensitivity and favorable time scale. Early ESR studies of short-range lateral diffusion in membranes were based on Heisenberg exchange (HE) effects of nitroxide spin-label linewidth. The HE contribution to the ESR linewidth is given for \(^{14}\text{N} \) nitroxides by

\[
\tau^{-1} = 8\pi^2 d^6 N_A C f
\]

where \( d \) is the encounter distance for two spin-bearing molecules, \( D \) is the microscopic self-diffusion coefficient, \( N_A \) is the Avogadro number, and \( C \) is the molar concentration of spins (41). The factor \( f \) can be deduced from the assumed model of intermolecular potential energy for the interaction between two spins. For a simple, hard-sphere potential \( f \equiv 1 \). The technique based on direct determination of changes in the ESR linewidth because of HE (42) was later improved by using ELDOR and/or saturation recovery. These techniques are based on the effects of the HE on saturation transfer between hyperfine lines of the spin labels, and they allow one to detect substantially lower HE rates. Additional improvement of this technique based on use of \(^{14}\text{N}-^{15}\text{N} \) spin-label pairs gave \( \sim 20 \)-fold in the sensitivity and allowed experiments on cells (43).

For the measurements of short-range diffusion, the collision rate can also be obtained by optical spectroscopy from fluorescence quenching. In the isotropic case of diffusion-controlled dynamic quenching, the change in fluorescence intensity obeys the equation:

\[
\frac{\Phi_1}{\Phi_0} = 1 + K_{SV}(Q)
\]

where \( \Phi_0 \) and \( \Phi \) are the quantum yield in the absence and presence of quencher, respectively; \( Q \) is the concentration of quencher, and \( K_{SV} \) is referred as the Stern-Volmer constant, which is related to the diffusion coefficient by the Smoluchowski equation. However, the isotropic and three-dimensional theory of fluorescence quenching does not apply in the membrane, and a Stern-Volmer modified treatment has been developed for this case (44).

Long-range diffusion measurements

For studying long-range diffusion, many studies have been conducted using the technique of fluorescence recovery after photo-bleaching (FRAP) (45). The translational diffusion coefficients that can be measured using this technique range from about \( 10^{-7} \) to \( 10^{-12} \) cm/s. For a small area of the membrane, that contains fluorescent-labeled molecules of interest, the initial level of fluorescence is determined. Then, a flash of intense laser beam is applied irreversibly to bleach a substantial (100% in the ideal case) fraction of fluorophores in this area, so it appears black after the flash. The intensity of the laser light for the bleaching is typically \( \sim 10^3 \) times greater than the light used to monitor the fluorescence. For the bleached area, the fluorescence intensity is recorded as a function of time. The brightness will gradually increase as fluorescent molecules diffuse into this area. Two parameters are determined from a FRAP experiment (Fig. 10): 1. The lateral mobility, which gives the diffusion coefficient directly and is determined by the slope of the fluorescent recovery curve. The steeper the curve, the more mobile the molecules.

![Figure 10 FRAP experiment. (From Reference 28.)](image-url)
2. The percent recovery, which is determined as $(Y/X) \times 100 = \%$ recovery. It gives the mobile fraction of the probe. If the radius of the illuminated area is small compared with the diffusion area (cell, vesicle, etc.) and the molecules are free to diffuse, then the percent recovery must be 100. On the other hand, if the fluorescence fails to recover to the same intensity observed before the bleaching pulse, then it indicates that a fraction of fluorophores exist, which are immobile on the time scale of the experiment ($D < 10^{-11}$ cm$^2$/s).

In recent years, FRAP for diffusion measurements in membranes has been superseded by fluorescence correlation spectroscopy (FCS). FCS is very similar to FRAP in both theoretical and experimental approaches to the observation of diffusion. The difference between these two closely related techniques is that FRAP measures relaxation from an initial nonequilibrium state after photobleaching, whereas FCS detects stochastic fluctuations that occur even in a system remaining in equilibrium (46).

ESR techniques for studying long-range diffusion usually observe the spreading of a small region of concentrated spins over time. For diffusion rates typical of a membrane environment, the most appropriate method is dynamic imaging of diffusion by ESR (DID-ESR) (47), which has been used to measure $D > 10^{11}$ cm$^2$/s. The measurement of the diffusion coefficient, $D$, by DID-ESR involves two stages. After preparing the sample with an inhomogeneous distribution of spin probes along a given direction, the investigator uses the ESR imaging method to obtain the (one-dimensional) concentration profiles at several different times. Spatial resolution results from a magnetic field gradient, because spin probes at each spatial point experience a different resonance frequency. With time, the inhomogeneous distribution will evolve toward a homogeneous distribution via translational diffusion. The second stage is to fit the time-dependent concentration profiles to the diffusion equation to obtain $D$. The ESR spectrum recorded in the presence of a magnetic field gradient $B$ ($G/cm$) is a convolution of the usual ESR spectra (gradient-off spectrum), $I_0(\xi)$, with the concentration of spins $C(x, t)$, which initially varies along the direction $x$. $I(\xi, t) = \int_{-\infty}^{\infty} C(t) I_0(\xi x - \xi x') dx'$

Here, $\xi = (B - B_0)$ measures the spectral position as the deviation of the magnetic field $B$ from the field $B_0$ at the center of the spectrum, which corresponds to the position $x = 0$, because the field gradient $B'$ maps $x$ onto $\xi$, as given by $\xi = B' x$. The determination of $C(x, t)$ from the two spectra $I_0(\xi, t)$ and $I(\xi, t)$ is a straightforward calculation through Fourier transformation. Fitting the $I(\xi, t)$ profile to the diffusion equation gives the diffusion rate.

Pulse field gradient (PFG) NMR spectroscopy is now generally regarded as the method of choice for measuring the translational diffusion coefficients of molecules of virtually any type under many conditions (48). $^1$H, $^13$C, $^19$F, and $^{31}$P variants of this method have been used successfully to study lateral diffusion of cholesterol, phospholipids, and water in model membranes (49, 50). This technique introduces two identical gradient pulses of the external magnetic field into the standard spin-echo NMR rf pulse sequence, one is between the $n/2$ and $n$ pulses and another is after the $n$ pulse. If spins do not undergo any translational diffusion, then the effect of the two applied gradient pulses cancels out and after the rf $n$-pulse the echo refocuses to the same value as in the absence of the gradient. However, the diffusion movement of spins between the gradient pulses causes additional dephasing, which cannot be refocused by the $n$-pulse and manifests itself in a decrease in the resulting echo intensity. The degree of dephasing is proportional to the displacement of spins in the direction of the gradient and, hence, the translational diffusion rate.

In the last decade, motion of lipid and protein single molecules in biomembranes was studied extensively by either single-particle tracking or by ultra-sensitive single-molecule fluorescent microscopy or fluorescence correlation spectroscopy. In single-particle tracking, a particle of typical diameter of 30–50 nm is attached to the lipid or protein molecule as a label. Colloidal gold and fluorescent particles have been used as labels. The particle motion is then followed by computer-enhanced video microscopy. Single-molecule fluorescence measurements are based on the repetitive excitation of a single fluorophore, which generates repeated cycles of absorption and fluorescence and count rates of up to tens of thousands of counts per second. The fluorescence is detected by single-photon counting modules.

Lipids diffuse freely in fluid model membranes. FRAP measurements show full recovery and diffusion coefficients on the order of magnitude of $10^{-4}$ cm$^2$/s. Free diffusion with a similar rate is often observed for lipids in the biomembrane. However, many cell membrane proteins show lower diffusion rates and incomplete recovery after photobleaching. For membrane proteins, dramatically different behavior in model and biological membranes is a common case. In model membranes, membrane proteins also diffuse freely and their diffusion coefficients are often similar to the diffusion coefficients of lipids. On the contrary, in biomembranes, the diffusion of proteins is 2–3 orders of magnitude slower and the fluorescence recovery is often incomplete. This observation points to limitations of the fluid mosaic model as will be discussed below.

Fluidity Versus Mosaicity—New Concepts in Membrane Science

The Singer-Nicolson model of the membrane played a very important role in understanding membrane structure and function. However, many properties of biomembranes are not consistent with this model. In recent years, a growing consensus points to limitations of the fluid mosaic model as will be discussed below.
membranes and biomembranes deprived of their cytoskeleton (blebs). A notable observation is the oligomerization-induced slowing of diffusion (51). It manifests itself in much greater effect of diffusant size on the translational diffusion rate than predicted by the theory of Saffman-DeBrück based on the Singer-Nicolson model (52). If a transmembrane protein is approximated as a rigid cylinder of radius $r$ and height $h$, floating in a two-dimensional liquid of viscosity $\eta_1$ with matching thickness $h$, then the theory gives the following expression for its translational diffusion coefficient $D$:

$$D_T = \frac{k_B T}{4 \pi \eta_1 h} \left( \ln \frac{h}{r} - \frac{1}{2} \right)$$

where $\gamma$ is the Euler constant ($\approx 0.577216$). This formula predicts very weak dependence of translational diffusion on the size of diffusant. For example, for a 10-fold increase in the protein radius, from 5 to 50 Å, the theory predicts for a 50-Å thick bilayer a decrease in the diffusion rate by a factor of 1.57. A rather factor of 10-increase in the radius, which corresponds to a 75-fold increase in the aggregation number, slows the diffusion rate by a factor of 2.6. Although this insensitivity looks somehow counterintuitive, the Saffman-Debrück formula gives good results for proteins incorporated into model membranes.

On the contrary, for the plasma membrane, the effect of oligomerization on the diffusion rate is much stronger than predicted by the formula. For example, for linked couples of green fluorescent protein-E-cadherin oligomerization with an aggregation number between 2 and 10 slows down the diffusion up to 40 times (53). Also, spectacular single-particle tracking experiments carried out by Kusumi et al. (51) showed that in natural biomembranes, the diffusion does not follow usual Brownian patterns but consists of a series of random Brownian walks within confined areas (compartments) followed by longer-distance hops between compartment. (see trajectory in Fig 11) (54). The compartment size varies depending on the cell nature but is relatively insensitive to the diffusant, which shows simple Brownian diffusion with a single diffusion coefficient (55).

Because the assumption of simple Brownian diffusion breaks down, the diffusion in biomembranes cannot be described by a single diffusion coefficient. For instance, FRAP experiments in the plasma membrane showed that the observed translational diffusion rates depend on the size of the initial photobleached spot, which is also inconsistent with a simple Singer-Nicolson model.

A cumulating evidence clearly points at involvement of the cell cytoskeleton in the compartmentalization of the membrane, in particular, the fine cytoskeleton filaments formed by actin in most eukaryotic cells or spectrin in mammalian red blood cells. However, single-particle tracking experiments show the same patterns of hop-diffusion for lipid molecules located in the extracellular leaflet of the plasma membrane. How can the membrane skeleton, which is located only on the cytoplasmic surface of the membrane, suppress the motion of lipids on the extracellular side?

To reconcile this apparent contradiction the membrane skeleton fence and anchored transmembrane picket model was proposed (54). According to this model, transmembrane proteins anchored to and lined up along the membrane skeleton (fence) effectively act as a row of posts for the fence against the free diffusion of lipids (Fig. 31). This model is consistent with the observation that the hop rate of transmembrane proteins increases after the partial removal of the cytoplasmic domain of transmembrane proteins, but it is not affected by the removal of the major fraction of the extracellular domains of transmembrane proteins or extracellular matrix. Within the compartment borders, membrane molecules undergo simple Brownian diffusion. In a sense, the Singer-Nicolson model is adequate for dimensions of about 10 × 10 nm, the special scale of the original cartoon depicted by the authors in 1972. However, at these distances simple extensions of the fluid mosaic model fail and a substantial paradigm shift is required from a two-dimensional continuum fluid to the compartmentalized fluid.

**Acknowledgment**

This work was supported by grants from NIH NCRR and NIH NIBIB.

**References**


Figure 11  Membrane skeleton fence and anchored transmembrane picket model. (From Reference 51.)


Further Reading


Microtubule Dynamics
Manu Lopus, Mythili Yenjerla and Leslie Wilson, Department of Molecular, Cellular, and Developmental Biology and The Neuroscience Research Institute, University of California, Santa Barbara, California
doi: 10.1002/9780470048672.wecb338

Microtubules are cylindrical, cytoskeletal protein polymers found in all eukaryotic cells. They perform a plethora of functions associated with cellular structure, organization, and movement. They polymerize and depolymerize by the reversible addition and loss of \( \alpha: \beta \) tubulin heterodimers, which are their building block subunits, at the microtubule ends. Importantly, microtubules are not simple equilibrium polymers. The hydrolysis of GTP that occurs when tubulin dimers add to the ends of microtubules creates two distinct and often coexisting dynamic behaviors: dynamic instability and treadmilling. These dynamic properties are crucial in guaranteeing the faithful segregation of chromosomes during cell division, in many kinds of intracellular transport, in cell signaling, and even during programmed cell death. Microtubule dynamics play critical roles in terminally differentiated cells as well. In neurons, microtubule dynamics regulate neuronal plasticity, growth-cone motility, and maintenance of mature neurons. Suppression of mitotic spindle microtubule dynamics by small drug molecules is an important therapeutic strategy for treatment of many types of cancer. In addition, agents that stabilize microtubule dynamics offer potential for treating neurodegenerative diseases in which the dynamics are misregulated. Agents that modulate microtubule dynamics can be used as experimental tools or could be used therapeutically to perturb the functions of any cellular disease processes dependent on the dynamics.

This review focuses on the dynamic behaviors of microtubules. Because of space limitations, we concentrate on the molecular, mechanistic, and kinetic aspects of microtubule dynamics, their significance and regulation in dividing cells, and how microtubule-targeted drugs that modulate microtubule dynamics act to inhibit mitosis and kill tumor cells. It is an extensive and rich area of research; thus we must direct the reader to the cited literature including the Further Reading list for a more thorough coverage of the field.

General Features of Microtubule Structure and Polymerization

Microtubules are physically robust, dynamic, cylindrical, cytoskeletal polymers (Fig. 1) composed of the protein tubulin, which is a heterodimer that consists of one alpha and one beta subunit. The structural, mechanical, and polymerization properties of microtubules are essential for normal cell division, for development and maintenance of cell structure, for many kinds of intracellular transport, for positioning of intracellular organelles, and for programmed cell death (reviewed in References 1 and 2). Microtubules polymerize and depolymerize by the reversible noncovalent addition and loss of \( \alpha: \beta \) tubulin dimers at their ends. The orientation of \( \alpha: \beta \) tubulin dimers in the microtubule provides the polymer with structural and kinetic polarity, with the end designated as the plus end, where \( \beta - \)tubulin is exposed, being relatively more dynamic than the opposite or minus end, where \( \alpha - \)tubulin is exposed. The opposite ends of microtubules exhibit remarkably different behaviors. For example, during assembly of microtubules, tubulin polymerizes more rapidly at plus ends than at minus ends. Also, as described in more detail below, the plus ends alternate between phases of growth and shortening more frequently and fluctuate in length to a greater extent than the minus ends (3).

The self-assembly of tubulin to form microtubules was described initially in a classic polymerization model of nucleated helical polymerization by Masayama and Osawa (4). A assembly involves two phases: a nucleation phase followed by an elongation phase. With purified systems in vitro, nucleation can...
be achieved in several ways, such as by inclusion in polymerization reactions of various stabilizing microtubule-associated proteins (MAPs), by using preformed microtubule seeds prepared either by shearing preformed microtubules through a 25 gauge needle or by using sea urchin axonemes (5). The nucleation of tubulin assembly in vitro also can be achieved with microtubule-stabilizing chemical substances such as glutamate, dimethyl sulfoxide, and glycerol (6).

Microtubule assembly in cells differs in some ways from assembly in vitro. In cells, nucleation of microtubules requires a third type of tubulin, which is called γ-tubulin, that functions in concert with other proteins in the form of a γ-tubulin ring complex. In most animal cells, the γ–tubulin ring complex is located at the pericentriolar region of the microtubule organizing center (or centrosome) where it nucleates microtubule assembly at the minus ends (7). The γ–tubulin does not become incorporated into the microtubule, but rather it only localizes to the minus ends. Assembly of tubulin to form microtubules during the early stages of polymerization in vitro can be considered a pseudo first-order reaction. A steady state is eventually attained in which both the soluble tubulin concentration and the microtubule polymer mass attain stable plateaus (8). The critical concentration at apparent equilibrium (actually a steady state, see below) is the concentration of soluble tubulin in apparent equilibrium with the microtubule polymers.

The assembly of tubulin to form functional microtubules is a complex process (see References 2, 8, and 9). Two GTP binding sites are on the tubulin dimer: One is on β-tubulin, which is readily exchangeable when the tubulin is in solution, and the other on α-tubulin, which is not exchangeable, located at the interface between α- and β-tubulin (1). The GTP bound to the exchangeable site (the E site) undergoes hydrolysis when soluble tubulin adds to the microtubule ends, which creates the nonequilibrium dynamic properties of the microtubules. In its simplest form the assembly may be written as:

$$P = (1 - f)T - Sc$$

where the concentration of polymers, $P$, depends on the critical concentration of the tubulin subunits for assembly ($Sc$), the fraction of the proteins that are not participating in the process ($f$) (for example, because of sequestration by regulatory proteins or denaturation of the tubulin), and the total protein concentration ($T$) (9). $Sc$ can be determined from the total protein concentration by measuring the polymer content as a function of the total protein concentration at apparent equilibrium. Extrapolation of the total protein concentration to zero polymer content allows determination of the critical concentration. Studies that involve video-enhanced differential interference contrast microscopy and cryo-electron microscopy have provided considerable insight into the nature of growth and shortening of microtubules (10). Accordingly, microtubules, when growing rapidly, display sheet-like extensions at the growing ends, with the eventual closure of the sheets to form the cylindrical microtubule structure. These growing tips are thought to be stabilized by a so-called stabilizing GTP cap (see below). The protofilaments at the ends of rapidly shortening microtubules form tightly curled oligomeric rings, which indicates that GTP hydrolysis creates a strain in the microtubule lattice, and when the cap is lost, the strained and destabilized protofilaments can dissociate rapidly from the microtubule end because of their intrinsic curvature.

**Microtubule Dynamics—Mechanistic Aspects**

Although the classic model of nucleated helical polymerization accurately describes many aspects of microtubule polymerization, it is now well established that microtubules are not simple equilibrium polymers. Rather, the hydrolysis of GTP to GDP that occurs as tubulin adds to growing microtubule ends creates two unusual nonequilibrium dynamic behaviors, which are known as treadmilling (11, 12) and dynamic instability (13). Treadmilling is the phenomenon of net growth of individual microtubules by the addition of tubulin at one end and net shortening by the loss of tubulin at the other end. Dynamic instability is the stochastic switching between growth and shortening (often called shrinking) phases at the microtubule ends. Although both behaviors are intrinsic properties of microtubules composed solely of tubulin, microtubule dynamics in cells are modulated by a wide variety of microtubule-associated proteins (MAPs) and in an enormous variety of ways (e.g., see Reference 2).
Treadmilling

Treadmilling, which is the unidirectional flow or flux of tubulin from plus ends to minus ends (Fig. 2a), was discovered in vitro using single- and double-radiolabeled GTP (3H–GTP and 14C–GTP) incorporation and loss experiments at the microtubule ends by Margolis and Wilson in 1978, while investigating the mechanism of substoichiometric poisoning of microtubule assembly by colchicine (11, 12). Use of double-label and pulse-chase strategies demonstrated that the net uptake and loss of tubulin occurred at opposite ends of the microtubules, whereas the total polymer mass and the lengths of the microtubules remained constant as determined by electron microscopy. These first treadmilling studies were conducted with MAP-rich brain microtubules. Because of the high MAP contact (see below), the treadmilling rates were slow (∼0.7 μm/h) and dynamic instability was almost completely suppressed. Later, Hotani and Horio provided the first visual proof for treadmilling (14). They decorated the center block of a three-block microtubule with Tetrahymena dynein and found that the length of microtubules at one end of the decorated block decreased and that at the other end increased, which left the length of the center block unchanged. The authors also discovered that in the presence of neural MAPs, dynamic instability was strongly suppressed, which left treadmilling the prevailing dynamic behavior (see Reference 14). Most recently, experiments with microtubules made in the absence of MAPs, but stabilized sufficiently with low concentrations of glycerol to inhibit dynamic instability, demonstrated that the treadmilling rate of MAP-free microtubules in vitro is quite rapid and can approach the rates observed in living cells (15) (see below). This work indicated even more that the treadmilling rate can be increased greatly by increasing the dissociation rate constant for tubulin loss at the minus ends, which is an action that can be accomplished readily by certain MAPs. Treadmilling is believed to be caused by the differences in the individual critical concentrations for tubulin addition at the opposite microtubule ends (15). The idea is that
the critical concentration for tubulin assembly at the growing end is lower than at the shortening end, and that at steady state, the overall critical subunit concentration is maintained between the two. For treadmilling to occur, both microtubule ends must be free for tubulin exchange (that is, neither end can be blocked such as occurs at minus ends of microtubules attached to centrosomes), and the microtubules must be at or near steady state so that the soluble tubulin level is not so high or so low that both ends grow or shorten. Because treadmilling can occur in the absence of the rapid shortening typical of that observed during dynamic instability, it is reasonable to think that both ends of a treadmilling microtubule retain their stabilizing cap (15).

Dynamic instability

In 1984, Mitchison and Kirschner (13) discovered dynamic instability in vitro when they observed the coexistence of growing and shrinking populations of microtubules that interconverted infrequently. The coexistence of the two populations of microtubules in a dynamically unstable manner was substantiated by Horio and Hotani by analyzing individual microtubules using dark-field microscopy (3). Dynamic instability thus came to be defined as the stochastic switching between the growing and shortening phases at microtubule ends (Fig. 2A (16, 17)). Dynamic instability in vitro occurs at both microtubule ends, with the dynamics at the plus ends considerably more robust than those at the minus ends (18). In cells, however, dynamic instability has only been observed at plus ends. Minus ends in cells have not been observed to grow; they either remain the same length or they shorten (19).

A minimum of four parameters are now used to define the various features of dynamic instability. These parameters are the growth rate, the shortening rate, and the switching frequencies from growth to shortening and from shortening to growth. The abrupt switching of an end from growth to shortening is referred to as a “catastrophe,” and the switching from shortening to growth is referred to as a “rescue” (18). Additional parameters are used to describe dynamic instability behavior. Specifically, microtubules both in cells and in vitro often do not change length for periods of time. This parameter is called “pause” in cells and “attenuation” in vitro. In both situations, tubulin may be exchanging at the microtubule ends, but the extent of addition or loss may be too low to be detected by video microscopy.

An especially useful parameter is termed the “dynamics” (see Reference 16), which is a measure of the total tubulin exchange per unit time for a microtubule, including periods of attenuation or pause. As examples of the various parameters, Table 1 shows the major dynamic instability parameters for a set of control microtubules in vitro and for microtubules in the presence of the microtubule-targeted drug tasidotin (20).

Hydrolysis of E-site GTP during or shortly after addition of tubulin to the microtubule ends and the gain and loss of a short region of GTP-(or GDP–Pi)-liganded tubulin at the extreme ends of the microtubules that stabilize the microtubule tips are believed responsible for dynamic instability. The tubulin dimer has intrinsic GTPase activity, with a rate that is relatively slow when tubulin is in solution (21). However, hydrolysis is triggered when the $\beta$-subunit of an incoming tubulin dimer with bound GTP docks at the end of an exposed $\alpha$-subunit at the end of the microtubule (22). Evidence in support of the existence of a very short stabilizing GTP cap, perhaps no larger than a single layer of tubulin-GTP (or GDP–Pi), has been obtained in many studies (too numerous to be described here). Studies show that tubulin addition to the microtubule ends and GTP hydrolysis are very closely coupled events (23, 24), and in experiments that use the slowly hydrolysable GTP analog guanylyl-(a, b)-methylene-diphosphonate (GMPCPP), a stable cap is formed from tubulin-GMPCPP subunits (25). The current thinking is that only the tip of the microtubule contains GTP-tubulin (or GDP–Pi) and that the “GTP cap” is required for continued growth to occur. Most of the microtubule core consists of GDP-tubulin, which is believed to be in a strained conformation (1, 10, 26). Thus, when the cap is lost (a catastrophe), the strained GDP-tubulin core is exposed at the ends, which rapidly depolymerize in the form of curved protofilaments (26). The re-establishment of the cap (a rescue) would result in regrowth. The interconversion between these two phases with microtubules composed of pure tubulin (no accessory MAPs) is explained by stochastic loss or by reacquisition of a stabilizing cap by this two-state model, but it is speculated that a closed-tube state, which presumably exists as a structural intermediate between polymerizing ends with sheets and depolymerizing ends with peeling GDP–tubulin oligomers, may represent a structural correlate of a kinetic intermediate in a three-state model (2). A similar model with a growing open state, shrinking state, and a third intermediate closed state has been proposed by Tran et al. (27).

Microtubule Dynamics and its Modulation by Microtubule-Targeted Drugs and Regulatory Proteins

Both dynamic instability and treadmilling occur extensively in all eukaryotic cells. Except for extremely stable microtubules such as those found in organelles such as cilia and flagella, most microtubules are dynamic and display dynamic behaviors that vary enormously in their robustness and type from one cell type to another and even within the cytoplasm of individual cells (2, 18, 28). Dynamic instability and treadmilling (or flux) are extremely rapid during mitosis, and the rapid dynamics are critically important because they are required for distribution of the duplicated chromosomes to the daughter cells in an exquisitely time-sensitive and accurate fashion (29). At the onset of mitosis in animal cells, the dynamics of the microtubules increase many fold and change qualitatively from the dynamics of interphase microtubules. For example, the minus ends of the microtubules tethered at the centrosome during interphase do not seem to grow or shorten. But when the centrosome develops into a mitotic spindle pole, the minus ends become unblocked and dynamically active (29). Also, dynamic instability at the plus ends of the microtubules that grow out from the spindle poles, characterized by frequent switching between growth and shortening states, facilitates microtubule attachment to the
kinetochores of the chromosomes (called chromosome capture) and facilitates chromosome alignment at the metaphase plate and their accurate segregation to the daughter cells (29). Reduced tubulin flux of kinetochore microtubules from their attachment point at the kinetochore toward their attachment region at the poles, which seems to be treadmilling facilitated by various associated motors and other regulators, creates necessary tension in the spindle and facilitates the accurate and timely segregation of the chromosomes to the daughter cells at anaphase (29). Clearly, whereas spindle microtubules remain tethered at both their plus and minus ends, their ends remain free for rapid flux (see Reference 2). Thus, they must be transiently anchored to other structures near their ends in a fashion that leaves the ends free for substrate loss or gain.

Dynamic microtubules are important not only in dividing cells but also in terminally differentiated cells. For example, they play crucial roles both in the post mitotic development of neurons and in mature neurons such as in the formation of functional neuronal networks and the correct arborization (branching) of dendrites (30). Misregulation of microtubule dynamics, as for example caused by mutations in the neuronal MAP-lau, can lead to microtubules whose dynamics fall outside the normally permissible range thus possibly contributing to neurodegeneration in tauopathies such as Alzheimer’s disease and FTDP-17 (Fronto-Temporal Dementia with Parkinsonism associated with Chromosome 17) (see Reference 31) (see below).

Modulation of microtubule dynamics by microtubule-targeted drugs

Here we will focus on how the functions of dynamic microtubules can be modulated powerfully by chemical agents. The fact that microtubule dynamics are indispensable for normal cell function and survival leads naturally to the idea that targeting their dynamics with small molecules is a highly attractive strategy for drug development and chemical biology. This has been especially fruitful in cancer chemotherapy. Several classes of microtubule-targeted drugs are vitally important in the treatment of cancer, which includes the vinca alkaloids and the taxanes. Although high concentrations of these drugs can increase or decrease the mass of assembled microtubules in vitro and in cells, the most sensitive actions of these drugs on microtubules, which occurs at low drug concentrations in the absence of changes in polymer mass, is to suppress their dynamics (16). Despite their opposite effects on microtubule polymerization and their different specific mechanisms of action, most successful chemotherapeutic drugs that act on microtubules share the common property of suppressing spindle microtubule dynamics, which leads to inhibition or slowing of cell cycle progression at prometaphase of mitosis and at the transition from metaphase to anaphase (32). In sensitive tumor cells, mitotic arrest is followed by apoptosis. The mechanism that underlies the relationship between disruption of spindle dynamics and induction of apoptosis is yet to be fully explored, although evidence that diffusible factors maybe released as the result of changes in microtubule polymerization is consistent with the hypothesized role of microtubules in sequestering signals (33). In fact, several successful microtubule-targeting drugs are known to promote apoptosis in tumor cells (34).

Microtubule interfering drugs act by binding to various sites on the tubulin dimer and at different positions within the microtubule. Although other categories clearly exist, currently most drugs are classified into three major categories based on their respective tubulin binding domains: which include the Vinca alkaloid domain, the colchicine domain, and the paclitaxel domain. These may be some of the same regions of microtubule surfaces used by natural regulators of dynamics in cells and, thus, the drugs can be thought of as possible mimics of microtubule regulatory proteins (35). Vincristine and vinblastine have long been used for the treatment of hematological cancers. Vinca alkaloids currently used for treatment of cancer include two natural products: vincristine, and vinblastine, and several novel semi-synthetic drugs, vindesine, vinorelbine and vinflunine. At low concentrations, vinca alkaloids suppress both dynamic instability and treadmilling, apparently by binding to microtubule ends. The superior anti-tumor efficacy and reduced toxicity of vinflunine (36) over vinblastine may be attributed to its less powerful inhibitory effects on microtubule dynamics. Defective checkpoints in certain cancer cells may make them more susceptible to the less powerful inhibitory effects of vinflunine than normal cells (36). Drugs such as vinflunine, various derivatives of the dolastatins including tasidotin (20) (Table 1), cryptophycin analogs (37), and halichondrin B analogs such as eribulin (E7389) (38) comprise a series of chemically distinct compounds that bind in the vicinity of the Vinca binding domain and are under various stages of development for cancer treatment. The microtubule seems to be exquisitely sensitive to the action of these drugs as the binding of only a few drug molecules along the microtubule surface or at its ends is sufficient to suppress microtubule dynamics (see Reference 16). Colchicine and compounds that bind in the vicinity of the colchicine-binding domain of tubulin comprise another class of drugs with potential for treatment of cancer. Colchicine binds to tubulin at αβ dimer interface, and acts by being incorporated with low stoichiometry at the ends of the microtubules as a tubulin-colchicine complex (39). Although colchicine has...
antitumor properties, its therapeutic use is hampered by its high toxicity. Drugs that bind to tubulin at or near the colchicine site that are in clinical trials include 2-methoxyestradiol and combretastatin A-4 phosphate. 2-methoxyestradiol induces mitotic arrest by suppression of microtubule dynamics (40), whereas the vascular targeting agent combretastatin A-4 phosphate acts by destabilizing microtubules at plus ends (41).

Pacilaxel and docetaxel are among the most successful microtubule-targeted drugs currently used to treat solid tumors, which include ovarian, breast, head and neck, lung, and prostate (42). The binding site for the taxanes is on the β-tubulin and is located on the inside surface of the microtubule (43). Binding of only a few molecules of paclitaxel to tubulin in microtubules strongly suppresses dynamic instability at microtubule plus ends with only a marginal increase in microtubule polymer mass (44). Other promising anticancer compounds believed to act at the taxane binding sites include the epothilones (45), discomelaid (46), eleutherobin, and several novel taxanes (47). Pacilaxel may find a role in the treatment of neurodegenerative diseases. Specifically, it improved axonal function and ameliorated neurological problems in mice that carry a defective human tau gene (48). Interestingly, some evidence indicates that pacilaxel and tau may share the same binding site on the microtubule surface (49), and the effects of tau on microtubule dynamics are rather similar to those of pacilaxel (31).

Regulation of microtubule dynamics by cellular MAPs

A plethora of proteins regulate microtubule dynamics in cells. All these regulators are potential targets for chemical biology. As indicated on the left, presence indicates that the drug molecules may mimic the effects of natural regulators of dynamics, perhaps by binding to the microtubules at similar sites or regions of tubulin (50). Identifying proteins that regulate microtubule dynamics have traditionally been classified into two groups: microtubule-stabilizing proteins and microtubule-destabilizing proteins. A third recently described group of proteins known as +TIPs, which track the plus ends of growing microtubules in cells, also are thought to regulate microtubule dynamics (28).

Regulation by stabilizing proteins

Several MAPs have been known to promote tubulin assembly and to stabilize microtubules. Members of this family include the neuronal proteins tau and MAP2, which are present in axons and dendrites, respectively, and MAP4, which is found in all non-neuronal vertebrate cells (50). This group of structural MAPs also includes MAP1A and MAP1B, which are found mainly in axons and dendrites (51, 52). Regulation by these MAPs is complex because their ability to regulate polymerization and dynamics is in turn regulated by phosphorylation (53). Currently, tau is attracting considerable attention because of its involvement in various neurodegenerative diseases and because it may be a target possible treatment of Alzheimer’s disease (31,54–56). Tau is a complex family of MAPs found specifically in neurons that strongly promotes microtubule polymerization and stabilizes microtubules. Although only a single tau gene exists, six developmentally regulated tau isoforms are produced in human brain because of alternative splicing of the single tau gene. Tau is also the target of multiple kinases that can phosphorylate it at a great many sites, which produces many more isoforms. The various tau isoforms differentially modulate dynamic instability (reviewed in Reference 31). For example, an adult form of tau with 4 microtubule binding repeats (called 4R tau) strongly suppresses the shortening rate at plus ends, whereas a fetal form of tau with only 3 repeats (3R tau) does not (55). In the neurodegenerative disease FTDP-17, mutations that result in altered mRNA splicing change the crucial expression ratio of the 3R and 4R tau isoforms (56). Altered expression of normal tau isoforms could result in misregulation of microtubule dynamic in axons and thus contribute to neurodegenerative disease.

Regulation by destabilizing proteins

Stathmin/Osp1B (oncoprotein1B) and certain members of the kinesin-related motor proteins are among the major classes of proteins that cause microtubule destabilization. Stathmin is a ubiquitous microtubule-destabilizing protein that is believed to play an important role in linking cell signaling to the regulation of microtubule dynamics. It is known to sequester free tubulin and therefore to impede microtubule formation (57). Clearly, such an action can affect microtubule dynamics. But like the action on microtubules of many drugs, the effects of stathmin on tubulin and microtubules are complex. Specifically, relatively low concentrations of stathmin increase the steady-state catastrophe frequency by a direct action on microtubules, with the catastrophe-promoting activity being considerably stronger at minus ends than at the plus ends (58). Stathmin also greatly increases the microtubule treadmilling rate. These data indicate that stathmin may be an important regulator of minus-end dynamics, as for example, by increasing depolymerization at microtubule minus ends at spindle poles during mitosis thereby increasing the plus to minus end flux rate (58). Other destabilizing proteins include microtubule severing proteins such as katanin (59) and the kinesin-related motor protein, XKC M1 (60).

Plus end tracking proteins

Finally, we want to mention briefly a group of proteins known as +TIPs or plus-end tracking proteins (28). This group of proteins is composed of microtubule motor complex components, signal transduction molecules, and molecular adaptors. These proteins can recognize and associate with growing microtubule plus ends and include CLIP-170 (cytoplasmic linker protein), which is an endosome-microtubule linker protein that was the first +TIP identified; members of the family EB1, EB2, and EB3 (end-binding proteins) that bind to APC (adenomatous polyposis coli, a tumor suppressor protein); the dynactin/dynactin microtubule motor complex, in particular the 150kDa subunit of dynactin; and several other proteins such as CLASP1 (CLIP-associated protein) and LIS1 (28, 61). EB1 enhances microtubule polymerization by increasing rescues and preventing catastrophes (62). Using dominant negative constructs, it was shown that in the absence of CLIPs, the microtubule rescue frequency was reduced, which suggests a mechanism that
involve CLIPs by which microtubule plus ends may be concentrated near the cell margin (63). Although many functions of the +TIPs are not yet understood, targeting their association with the growing tips of microtubules could be an area for future development through chemical biology.

Conclusion

Treadmilling and dynamic instability—two intrinsic dynamic properties of microtubules—are critical for mitosis and many other cellular functions that involve dynamic microtubules. Because the dynamics of microtubules and their tight regulation by a host of MAPs play crucial roles in so many cell functions, it is reasonable to think that modulating their dynamics in specific cell functions by targeting the microtubules themselves or by targeting the regulatory proteins through chemical biology will remain an attractive area of research for many years to come.

Acknowledgments

We acknowledge the support of USPHS grant NS13560, which supported the authors.

References


Specific receptors on the surface of mammalian cells actively internalize cell-impermeable ligands by the mechanism of receptor-mediated endocytosis (RME). This process is critical for the acquisition of nutrients, signal transduction, development, neurotransmission, and cellular homeostasis. Binding of ligands to internalizing receptors on the plasma membrane results in clustering of the complex in clathrin-coated pits or other dynamic membrane regions. Invagination of these regions yields intracellular vesicles that fuse to form membrane-sealed endosomes. Receptors typically dissociate from ligands in these acidic compartments, which allows the free receptor to cycle back to the cell surface, whereas ligands are often degraded on delivery to lysosomes, which liberates amino acids and other nutrients. By mimicking endogenous ligands, certain protein toxins, viruses, and other pathogens exploit RME to enter the cytoplasm or reach other intracellular destinations. Similarly, artificial delivery systems that mimic ligands or receptors can enhance efficiently the cellular uptake of impermeable molecules, including drugs, proteins, and nucleic acids. Advances in small-molecule probes, structural biology, and genetic methods are beginning to illuminate the complex mechanisms of this process at the molecular level.

The plasma membrane of eukaryotic cells encapsulates the inner cellular machinery, thereby protecting fragile biologic structures from potentially toxic or opportunistic extracellular materials. Only small hydrophobic molecules can penetrate rapidly this lipid bilayer through passive diffusion. More polar essential amino acids, sugars, and ions access the cell interior by interacting with membrane proteins that function as selective pumps or channels. For many other cell-impermeable small molecules, macromolecules, and particles to access the cell interior, cells must facilitate uptake actively, with regions of the plasma membrane functioning to capture solutes by invagination and pinching off to form intracellular vesicles. This process is termed endocytosis, which represents multiple related mechanisms for the internalization of extracellular molecules (1). Endocytosis is divided into two primary categories: phagocytosis (cell eating) and pinocytosis (cell drinking). Phagocytosis enables the uptake of large particles, including intact bacteria and yeast, through an actin-mediated mechanism that is generally restricted to specific cell types, such as macrophages, monocytes, and neutrophils. Pinocytosis, by contrast, occurs in all nucleated mammalian cells and involves the active invagination of small regions of cellular plasma membrane to capture solutes within vesicles of less than 200 nm in diameter. These vesicles fuse in the cytoplasm to form membrane-sealed compartments termed endosomes, and their contents are sorted to allow trafficking to specific destinations. Pinocytosis can involve the nonspecific uptake of extracellular fluid, as well as the uptake of specific molecules in the extracellular environment, mediated by receptors on the plasma membrane. Some mechanisms of endocytosis operate rapidly and continuously. In cultured fibroblasts, under physiologic conditions, membrane equivalent to the entire cell surface is perpetually internalized with a half-life of 15 to 30 minutes (2). Most pinocytic pathways involve specific interactions of receptors with ligands. In receptor-mediated endocytosis (RME), internalizing receptors on the cell surface bind cell-impermeable ligands to concentrate ligands in the cell. This mechanism is thousands of times more efficient than nonspecific pinocytosis for the cellular acquisition of nutrients and other impermeable molecules. The receptors involved in RME comprise a structurally diverse group of biomolecules that project ligand-binding motifs into the extracellular environment. Cell-impermeable small molecules, lipids, peptides, proteins, nucleic acids, and carbohydrates are internalized by RME, which enables the consumption of nutrients, elimination of pathogens, and termination of signals initiated by extracellular stimuli. RME followed by subsequent exocytosis...
of the ligand from one side of the cell to another is termed transcytosis, and this mechanism allows the delivery of nutrients across membrane barriers, such as the blood-brain barrier (3). Receptors involved in RME are listed in the following sections. Other representative examples of the interactions with the receptor include the structure of LDL determined by cryoelectron microscopy (shown reduced in scale compared to our model above or as a complex with receptor extracellular fragments) (15). Transferrin, human growth hormone, the bovine rhodopsin (12), and the FcγRIIB (CD16) (13) are shown as part of a composite image that illustrates the nature of attachment of the receptor to the plasma membrane. The small glycolipid receptor ganglioside GM1 is shown to the right, rendered as a molecular model (14). Structures of cognate ligands are positioned above or as a complex with receptor extracellular fragments. These ligands include the structure of LDL determined by cryoelectron microscopy (shown reduced in scale compared to the receptor) (15), transferrin, human growth hormone, the bovine rhodopsin (12), and the bovine rhodopsin (16). Brief descriptions of these and related receptors and ligands are provided in the following sections. Other representative examples of receptors and ligands involved in RME are listed in Table 1.

### The LDL receptor: a macromolecular membrane-spanning protein critical for cellular uptake of cholesterol

Uptake of cholesterol-laden LDL particles by the LDL receptor (LDLR, Fig. 1) is one of the best-characterized examples of RME (9, 17–18–19). The mature LDL receptor is a single-pass transmembrane glycoprotein of 839 amino acids (∼115 KDa, Fig. 1). LDL ligands are characterized as particles of ∼22 nm in diameter (∼2500 KDa) that comprise a core of ∼1500 molecules of cholesterol esters, esterified primarily by linoleic acid, encapsulated by a monolayer of free cholesterol, phospholipids, triglycerides, and a single large protein termed apolipoprotein B-100 (apo-B, ∼550 KDa). By recognizing the protein component of LDL, the LDLR enables cells in all tissues of vertebrate animals to internalize exogenous cholesterol, which is a key building block required for the biosynthesis of steroid hormones, bile acids, and cellular membranes. By interacting with the protein apolipoprotein B, which forms coated pits on the cytosolic face of the plasma membrane, the LDLR constitutively delivers LDL into endosomes, followed by cycling of the receptor back to the cell surface. Inherited mutations in the LDLR that disrupt endocytosis, and thereby increase serum LDL, have been shown to accelerate atherosclerosis in patients with familial hypercholesterolemia (17). Rapidly proliferating cells have a particularly high demand for cholesterol because mammalian plasma membranes are composed of one-third protein and two-thirds lipid plus ∼30% of the cellular plasma membrane lipids are cholesterol (20). For this reason, the LDL receptor is often overexpressed on cancer cells, and LDL receptors provide a target for the selective delivery of anticancer and tumor imaging agents (21, 22). The LDLR is also a portal exploited by Hepatitis C virus and other Flaviviridae viruses to penetrate into cells (23).

The transferrin receptor: a homodimeric transmembrane protein that enables cellular uptake of iron

Iron is an essential nutrient that functions as an enzyme cofactor in redox reactions and plays a structural role through ligand coordination. Under physiologic conditions, iron can be converted readily between the ferrous (Fe$^{2+}$) and the ferric (Fe$^{3+}$) oxidation states. However, ferrous iron is dangerous to living cells because it can generate hydroxyl radicals that oxidatively damage proteins, nucleic acids, and lipids. Additionally, iron in the ferric (Fe$^{3+}$) oxidation state forms a highly insoluble hydroxide complex that is not readily available to cells. In vertebrate animals, ferric iron is transported in serum bound to the protein transferrin (TF), which is a bilobed glycoprotein of 85 KDa in humans (Fig. 1; 10). This protein binds Fe$^{3+}$ using a synergistic anion, typically carbonate, two Tyr, one His, and one Asp residue. Cellular uptake of TF is mediated by the transferrin receptor (TFR, Fig. 1), which is a homodimeric transmembrane protein of BKDa in humans that binds two dfferent transferrin ligands. Internalization of TF by RME results in the release of Fe$^{3+}$ in acidic endosomal compartments. However, the apo-TF remains bound to TFR, the receptor-ligand complex cycles back to the plasma membrane, and apo-TF is released from the TFR at neutral pH. Ferric iron is reduced to the ferrous state within endosomes, and the iron transporter DMT1 delivers the Fe$^{2+}$ ion into the cytoplasm. Like the LDLR, the TFR is upregulated on certain cancer cell lines, and drugs conjugated to transferrin have been used as targeted delivery systems (5). In mice, the mouse mammary tumor virus exploits the TFR to enter cells (24).

### Receptors for Growth Factors and Hormones

Human growth hormone (GH1, Fig. 1), epidermal growth factor (EGF), insulin (INS), platelet-derived growth factor (PDGF), and many cytokines bind receptors that activate intracellular tyrosine kinase activity. The major isoform of GH1 is a protein of 191 amino acids (22 KDa) that functions in part to stimulate the growth of bone and internal organs in children. As shown in Fig. 1, the human growth hormone receptor (GHR), which is a member of the cytokine-hematopoietin receptor superfamily,
Receptor-Mediated Endocytosis

Figure 1  Representative structures of receptors and ligands involved in receptor-mediated endocytosis. The gray bar at the bottom of the figure represents the cellular plasma membrane. From left to right, X-ray crystal structures of the extracellular domains of the human LDL receptor, the human transferrin receptor, the human growth hormone receptor, bovine rhodopsin, and FcγRIIIB are shown illustrating the nature of attachment to the plasma membrane. A molecular model of the glycolipid ganglioside GM1 is on the far right. A structure of LDL determined by electron cryomicroscopy (27 Å resolution) is shown on the upper left (not drawn to scale; image courtesy of Dr. Wah Chiu, Baylor College of Medicine). Other ligands shown from left to right include receptor-bound transferrin, receptor-bound human growth hormone, receptor-bound Fc region of human IgG, and the B-subunit of cholera toxin.

comprises a transmembrane glycoprotein of 620 amino acids (130 KDa) (25). Binding of GH1 results in dimerization and conformational changes in the GHR that initiate cellular signaling via recruitment and activation of tyrosine kinases. The GHR is internalized constitutively via clathrin-coated pits, and both this receptor and its ligand are degraded by proteolysis in lysosomes, which provides a mechanism to terminate the extracellular signal. The receptors for EGF and insulin are receptor tyrosine kinases (RTKs) that become internalized only upon binding of ligands. The EGF family of RTKs includes EGFR (HER1, erbB-1), HER2 (erbB-2), HER3 (erbB-3), and HER4 (erbB-4). Upregulation of expression or the production of activating mutants of this family is known to cause several cancers (26). By binding its extracellular domain, the FDA-approved monoclonal antibody drug Herceptin downmodulates HER2, thereby inhibiting the proliferation of the subset of breast cancers that overexpress this receptor.

G-protein-coupled receptors

G-protein-coupled receptors (GPCRs), also known as seven transmembrane receptors (7TMs), are the largest known superfamily of proteins. They are involved in all types of responses to stimuli, from intercellular communication to the senses of vision, taste, and smell. They respond to diverse ligands ranging from photons (e.g., rhodopsin, Fig. 3) to small molecules (e.g., binding of epinephrine to the β-adrenergic receptor) and proteins (e.g., chemokine receptors). Binding of ligands to the extracellular or transmembrane domains of these proteins causes conformational changes that relay a signal to intracellular G proteins that trigger additional cellular responses. Many GPCRs undergo RME by binding to intracellular arrestin proteins that associate with clathrin. The importance of GPCRs in normal biologic processes and disease has made this family of proteins the target of up to 50% of all modern drugs.

Receptors anchored to the plasma membrane by lipids: Glycosylphosphatidylinositol (GPI)-anchored proteins and glycolipids

Some cell-surface receptors are attached to the plasma membrane by lipids that penetrate only into the outer leaflet of the bilayer. Posttranslational modification of proteins with GPI-lipids allows proteins such as folate receptor-2 (FOLR2) to attach to the cell surface and promote RME of the vitamin 5-methyltetrahydrofolate. Folate receptors are upregulated in certain cancers, and folate derivatives have been linked to drugs and molecular probes to treat and image certain tumors. The related GPI-linked receptor FcγRIIB (CD16, 26.2 KDa, Fig. 3) is involved in the immune response. This receptor binds the invariant Fc region of immunoglobulin-G to promote RME of this
Receptor-Mediated Endocytosis

Table 1

<table>
<thead>
<tr>
<th>Receptor</th>
<th>Ligand</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asialoglycoprotein receptors (ASGR1)</td>
<td>Asialoglycoproteins</td>
<td>TM</td>
</tr>
<tr>
<td>β-Adrenergic receptor (ADRB1)</td>
<td>Epinephrine</td>
<td>TM</td>
</tr>
<tr>
<td>CD16b (FCGR3B)</td>
<td>IgG</td>
<td>GPI</td>
</tr>
<tr>
<td>CD99 (FCAR)</td>
<td>Polymeric IgA</td>
<td>TM</td>
</tr>
<tr>
<td>Coxsackievirus &amp; adenovirus receptor (CXADR)</td>
<td>Coxsackievirus / adenovirus</td>
<td>TM</td>
</tr>
<tr>
<td>EGF receptor (EGFR)</td>
<td>Epidermal growth factor (EGF)</td>
<td>TM</td>
</tr>
<tr>
<td>Folate receptor 2 (FOLR2)</td>
<td>5-Methyltetrahydrofolate</td>
<td>GPI</td>
</tr>
<tr>
<td>Ganglioside GM1</td>
<td>Cholera toxin / SV40</td>
<td>Lipid</td>
</tr>
<tr>
<td>Globoside GB3</td>
<td>Shiga toxin</td>
<td>Lipid</td>
</tr>
<tr>
<td>Glucagon receptor (GCCR)</td>
<td>Glucagon (GCG)</td>
<td>TM</td>
</tr>
<tr>
<td>Growth hormone receptor (GHR)</td>
<td>Growth hormone (GH1)</td>
<td>TM</td>
</tr>
<tr>
<td>Heparin-binding EGF-like growth factor (HBEGF)</td>
<td>Diphtheria toxin</td>
<td>TM</td>
</tr>
<tr>
<td>IgE Fc receptor (FCER1A)</td>
<td>IgE</td>
<td>TM</td>
</tr>
<tr>
<td>Insulin receptor (INSR)</td>
<td>Insulin (INS)</td>
<td>TM</td>
</tr>
<tr>
<td>Interferon alpha receptor (IFNAR1)</td>
<td>Interferon alpha (IFNA2)</td>
<td>TM</td>
</tr>
<tr>
<td>Low density lipoprotein receptor (LDLR)</td>
<td>LDL, Flaviviridae viruses</td>
<td>TM</td>
</tr>
<tr>
<td>Neotinal Fc receptor (FCGR1)</td>
<td>Maternal IgG</td>
<td>TM</td>
</tr>
<tr>
<td>Opioid receptor (OPRD1)</td>
<td>Enkephalinins</td>
<td>7TM</td>
</tr>
<tr>
<td>PDGF receptor alpha (PDGFRα)</td>
<td>Platelet derived growth factor (PDGFB)</td>
<td>TM</td>
</tr>
<tr>
<td>Prolactin receptor (PRLR)</td>
<td>Prolactin (PRL)</td>
<td>TM</td>
</tr>
<tr>
<td>Terminal galactose</td>
<td>Ricin</td>
<td>Variable</td>
</tr>
<tr>
<td>Terminal sialic acid</td>
<td>Influenza virus</td>
<td>Variable</td>
</tr>
<tr>
<td>Thyroid stimulating hormone receptor (TSHR)</td>
<td>Thyroid stimulating hormone (TSHB)</td>
<td>TM</td>
</tr>
<tr>
<td>Transferin receptor (TFRC)</td>
<td>Transferin (TF)</td>
<td>TM</td>
</tr>
<tr>
<td>Tyrosine kinase receptor A (NTRK1)</td>
<td>Nerve growth factor (NGFB)</td>
<td>TM</td>
</tr>
</tbody>
</table>

Examples of receptors and ligands involved in RME. Specific gene symbols of representative human receptor and ligand proteins are listed in parentheses. TM: transmembrane protein; GPI: glycosylphosphatidylinositol anchored protein; 7TM: Seven-transmembrane protein.

ligands. Much smaller glycolipids also participate in RME. Ganglioside GM1 (Fig. 1), a 1.6 KDa glycolipid, enables the protein cholera toxin (16) and the nonenveloped virus SV40 to penetrate into cells upon binding to its pentasaccharide headgroup (27).

Because of the lack of a direct connection to clathrin via a cytoplasmic region, the endocytosis of GPI-linked proteins and other lipid-linked receptors is slower than the uptake of most transmembrane proteins. Instead of clathrin-mediated endocytosis, the internalization of many lipid-linked receptors has been proposed to involve distinct membrane subdomains termed lipid rafts (28). These domains are enriched in cholesterol and sphingolipids and in some cell types include flask-shaped invaginations termed caveolae (29, 30). Many proteins covalently or noncovalently associated with cholesterol, sphingolipids, and in some cell types include lipid rafts with segregation and concentration membrane proteins, regulate signal transduction pathways (31), and control the endocytosis of specific receptors (32). Protein toxins and viruses often exploit receptor-mediated endocytosis involving lipid rafts or clathrin to penetrate into the cell interior (4).

Receptor-Mediated Endocytosis Visualized by Confocal Laser Scanning Microscopy

Microscopy has been used extensively to investigate mechanisms of RME. Electron microscopy was instrumental in the identification of clathrin-coated pits, endosomes, and other cellular features involved in this process (33). More recently, confocal laser scanning microscopy of living cells has been employed to investigate the uptake of fluorescent ligands, the influence of molecular probes, and the localization of fluorescent receptors and other proteins during endocytosis. An example of RME as imaged by confocal microscopy is shown in Fig. 2. In this figure, Jurkat lymphocytes, a human helper-T cell line, were treated with a mixture of transferrin conjugated to the bright green fluorescent Alexa Fluor 488 and cholera toxin B-subunit conjugated to the red fluorescent Alexa Fluor 594. After treatment for 5 minutes, transferrin is internalized by its receptor substantially more rapidly than cholera toxin, which remains partially localized at the cellular plasma membrane. Uptake of these proteins results in delivery in part into distinct early endosomal compartments, which is consistent with significant differences in the mechanisms of endocytic uptake.
Receptor-Mediated Endocytosis

After 1 hour of treatment, fluorescent transferrin can be localized in early endosomes and the endosomal-recycling compartment, whereas the cholera toxin B-subunit distinctly traffics to the trans-golgi network of living cells. In cells treated with holo-cholera toxin comprising the B-subunit and the catalytic A-subunit, the toxin would traffic further to the endoplasmic reticulum, which would enable the release of the toxic A-subunit into the cytoplasm.

Mechanisms of Receptor-Mediated Endocytosis

Cell-surface receptors are involved in both phagocytosis and pinocytosis. At least four distinct mechanisms of pinocytosis have been characterized: macropinocytosis, clathrin-mediated endocytosis, raft/caveola-mediated endocytosis, and clathrin- and caveola-independent endocytosis (1). Selected receptor-mediated aspects of these mechanisms are outlined below.

Phagocytosis

Phagocytosis is an actin-mediated mechanism of RME predominantly employed by specialized cells such as macrophages, neutrophils, and monocytes (34). This process allows these cells to clear large pathogens, such as bacteria and yeast, or debris, such as the remnants of dead cells, or deposits of cholesterol and other lipids in arteries. Binding of specific cell-surface receptors to their ligands triggers phagocytosis. For example, Fc receptors on macrophages bind the Fc region of antibodies that coat surface antigens on pathogens or particles. These recognition events activate signaling cascades involving Rho-family GTPases, which triggers the actin-driven formation of membrane protrusions. The formation and collapse of these plasma membrane ruffles generate large endocytic vesicles, termed macropinosomes, with diameters of 0.5 to 2.5 \( \mu \)m. Constitutive macropinocytosis by dendritic cells allows the efficient capture of exogenous antigens; presentation on the cell surface bound to MHC molecules provides a mechanism for stimulation of immune responses (36). Macropinocytosis also may be involved in the downregulation of activated signaling molecules.

Clathrin-mediated endocytosis

Clathrin-mediated endocytosis (CME) is responsible for the uptake of about 50% of all ligands internalized by cell-surface receptors. This process is critical for the continuous uptake of nutrients, intercellular communication during development, and regulation of signal transduction throughout life. CME modulates cellular signaling by controlling the levels of cell-surface receptors and downregulating activated signaling receptors. CME affects cell and serum homeostasis by controlling the internalization of membrane pumps involved in the transport of ions and small molecules across the plasma membrane. In...
Receptor-Mediated Endocytosis

Figure 3  Pathways of endocytic trafficking of receptors and ligands. The model illustrates uptake of LDL mediated by the LDLR, diferric transferrin (TF-Fe3+) internalized by the transferrin receptor (TFR), and entry of cholera toxin (CTX) and simian virus-40 (SV40) after binding to ganglioside GM1. Related trafficking of glycosylphosphatidylinositol-anchored proteins (GPI-AP) is also shown. The LDLR and TFR concentrate in clathrin-coated pits and initially deliver ligands into endocytic vesicles that fuse with sorting endosomes. The acidic environment of sorting endosomes dissociates most receptor–ligand complexes, and membrane proteins typically exit these compartments rapidly and return directly to the plasma membrane or are shuttled to the endocytic recycling compartment (ERC or recycling endosome). LDLRs and TFs are recycled from the ERC back to the cell surface. LDL is sorted to late endosomes and lysosomes, where it is degraded and releases cholesterol and amino acids into the cell. TF-Fe3+ releases iron in the acidic sorting endosome, but under acidic conditions, iron-free TF remains bound to the TFR. Upon return to the cell surface, at neutral pH, iron-free TF dissociates from the receptor. Binding of CTX and SV40 to GM1 primarily results in endocytic uptake via uncoated pits and caveolae, respectively. Internalized GPI-APs, CTX, and SV40 traffic through either the GPI-anchored protein-enriched early endosomal compartment (GEEC) or a related compartment termed the caveosome. SV40 moves from the caveosome directly into the endoplasmic reticulum (ER). In contrast, CTX exits the GEEC and passes through the trans-Golgi network to the ER. From the ER, CTX and SV40 penetrate into the cytosol, resulting in toxicity or infection, respectively. The t1/2 values shown are approximate and are cell-type dependent.

neurons, CME promotes the uptake of voltage-gated ion channels, which affects the strength of synaptic transmission, and it is involved in the recycling of membrane proteins of synaptic vesicles after neurotransmission. The protein clathrin comprises 190-KDa and 25-KDa subunits that form a basket-like structure on the cytoplasmic face of the plasma membrane. These subunits assemble as complexes with adapter proteins into highly ordered polygonal arrays that define pits on the cell surface. When clathrin-coated pits invaginate and pinch off, they form clathrin-coated vesicles (CCVs) with a diameter of ~120 nm. Assembly of a CCV in cultured cells takes ~1 minute, and hundreds to a thousand or more can form every minute. Clathrin-mediated endocytosis is the best-characterized mechanism of ligand uptake, and interactions of clathrin with receptors such as LDLR and TFR result in clustering of receptor–ligand complexes in clathrin-coated pits (Fig. 3). The LDLR and TFR interact with clathrin via adapter proteins such as the autosomal recessive hypercholesterolemia (ARH) protein and AP-2, respectively, but the LDLR also interacts directly with clathrin (37). GPI-APs are linked to clathrin via an adapter-like protein, β-arrestin, which interacts with AP-2. Specific peptide sequences bind adapter proteins and couple receptors to the clathrin-controlled endocytic machinery. The best-defined coated pit internalization signals are the tyrosine-based FxNPxY (F = phenylalanine, x = any amino acid, N = asparagine, P = proline, and Y = tyrosine) motif found in the LDLR, the YxΨ1ø motif of the TFR, and a dileucine motif of the insulin and β2-adrenergic receptors (1). Internalization of LDL and transferrin is a constitutive and rapid clathrin-mediated process. However, binding of other ligands, such as EGF to EGFR or epinephrine to the β2-adrenergic receptor, induces internalization via clathrin (32). Expression of dominant negative mutants of dynamin, the AP-2 binding partner Eps15, and its binding partner epsin have been used to inhibit CME and to identify mechanisms controlling receptor-mediated endocytosis.

Receptor–ligand complexes clustered in clathrin-coated pits are internalized when the plasma membrane invaginates, GTP-driven conformational changes of dynamin trigger membrane...
EGFR and GPCRs, ubiquitination of receptor lysine residues is for use by the cell (18, 38). For some receptors, such as the other components of LDL in lysosomes liberates these nutrients drolytic enzymes. Hydrolysis of cholesteryl esters, protein, and lysosomes, more acidic organelles (pH the cell surface, they can become internalized on binding of Although caveolae are considered relatively static structures on caveolins and lack these morphological membrane features. the pit. Lymphocytes and many neuronal cells do not express proteins of the caveolin family on the cytoplasmic face of cell types, including adipocytes, endothelia, and muscle cells. These lipid rafts can be observed by electron microscopy as distinctive flask-shaped pits of ~60 nm in diameter and include proteins of the caveolin family on the cytoplasmic face of the pit. Lymphocytes and many neuronal cells do not express caveolins and lack these morphological membrane features. Although caveolae are considered relatively static structures on the cell surface, they can become internalized on binding of ligands to receptors that associate with these raft subdomains. After activation, caveolae are internalized with relatively slow kinetics (half-life ~ 20 min) compared with CME (1).

In cells bearing caveolae, binding of simian virus-40 (SV40) to ganglioside GM1 in these subdomains triggers internaliza-
tion into compartments termed caveosomes. This DNA virus subsequently undergoes trafficking to the endoplasmic reticu-
lim, which is a destination that allows escape into the cytoplasm through an unknown mechanism, before the virus enters the nu-
cleus. In line cells lacking caveolae, other lipid raft domains are thought to promote delivery into distinct compartments termed GPI-anchored protein early endosomal compartments (GEECs) (37). For example, binding of chlorella toxin to ganglioside GM1 results, at least in part, in uptake through uncoated pits and in trafficking to the GEEC through the trans-golg network even-
tually to the ER, where a catalytically active fragment can escape into the cytosol and exert toxic effects. Chlorella toxin, however, is not a specific marker for raft-mediated endocyto-
sis; this protein is known to be internalized by three distinct mechanisms: clathrin coated pits, caveolae, and a clathrin- and caveolin-independent pathway (40). The mechanism that cou-
ples recognition of the glycolipid on the outer leaflet of the membrane to the clathrin machinery on the inner leaflet is un-
known. The relationship between lipid rafts and clathrin is also not yet well defined. For example, the EGFR is internalized through a mechanism that seems to simultaneously involve both lipid rafts and clathrin (41). Mechanisms of endocytosis that are independent of both clathrin and caveolin are not well under-
stood.

Receptor-mediated endocytosis via caveolae and other mechanisms

Lipid raft microdomains of mammalian plasma membranes are thought to regulate the endocytosis of specific ligands (32, 39). These domains are envisaged to comprise islands of or-
dered cholesterol, sphingolipids, and saturated lipids that move within the plane of disordered unsaturated lipids. The forma-
tion of lipid rafts depends on the availability of cholesterol in the membrane, and agents that sequester cholesterol, such as niacinamide, can selectively disrupt these microdomains. Dominant negative mutants of dynamin also block endocy-
tosis via this mechanism. GPI-anchored receptors, such as FcγRIIIB and folate receptors, and glycolipids, such as gan-
glioside GM1, are thought to reside in lipid rafts and become internalized at least partially by caveole mediated and clathrin-
independent/caveolin-dependent endocytic pathways. Caveolae represent a subset of lipid rafts found on specific cell types, including adipocytes, endothelia, and muscle cells. These lipid rafts can be observed by electron microscopy as distinctive flask-shaped pits of ~60 nm in diameter and include proteins of the caveolin family on the cytoplasmic face of the pit. Lymphocytes and many neuronal cells do not express caveolins and lack these morphological membrane features. Although caveolae are considered relatively static structures on the cell surface, they can become internalized on binding of molecular and cellular probes of receptor-mediated endocytosis

Molecular and Cellular Probes of Receptor-Mediated Endocytosis

Small molecules, proteins, and genetic constructs that activate, block, or label specific endocytic components or pathways represent key tools for studies of RME. Representative examples of probes of RME and related cellular processes are provided in Table 2. An overview of these approaches is provided in the following subsections.

Small-molecule regulators of phagocytosis and macrophagy

Many compounds that perturb the cellular cytoskeleton affect phagocytosis and macrophagy. Binding to actin filaments by the natural product cytochalasin D blocks both of these up-
take mechanisms. Disruption of microtubules by the antimitotic agent colchicine and nocodazole inhibits macrophagy and affects some mechanisms of phagocytosis. The diuretic drug amiloride, which is an inhibitor of Na+/H+ antiporters, selec-
tively blocks macrophagy. By activating protein kinase C, phorbolesters represent a class of small molecules that promote macrophagy.
Table 2

<table>
<thead>
<tr>
<th>Molecular probe or inhibitor</th>
<th>Target or mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetic acid</td>
<td>Acidifies the cytoplasm and freezes clathrin networks</td>
</tr>
<tr>
<td>Amantadine</td>
<td>Blocks budding of clathrin-coated vesicles</td>
</tr>
<tr>
<td>Amiloride</td>
<td>Inhibits macroinocytosis</td>
</tr>
<tr>
<td>Baflomycin A1</td>
<td>Inhibits v-ATPases; raises endosomal pH</td>
</tr>
<tr>
<td>BODIPY TR ceramide</td>
<td>Fluorescent marker of the golgi complex</td>
</tr>
<tr>
<td>Brefeldin A</td>
<td>Inhibitor of protein transport in the golgi complex</td>
</tr>
<tr>
<td>Chloroquine</td>
<td>Weak base; raises endosomal pH; disrupts endosomes</td>
</tr>
<tr>
<td>Chlorpromazine</td>
<td>Inhibits clathrin lattice formation</td>
</tr>
<tr>
<td>CID of clathrin fusion protein</td>
<td>Disrupts clathrin lattices</td>
</tr>
<tr>
<td>Colchicine</td>
<td>Inhibits microtubule polymerization and macroinocytosis</td>
</tr>
<tr>
<td>Cytochalasin D</td>
<td>Disrupts actin and inhibits phagocytosis / macroinocytosis</td>
</tr>
<tr>
<td>Dominant negative proteins</td>
<td>Targeted inhibition of clathrin, AP2, Eps15, dynamin, others</td>
</tr>
<tr>
<td>Dynasore</td>
<td>Small molecule inhibitor of dynamin</td>
</tr>
<tr>
<td>ER tracker Blue-White</td>
<td>Fluorescent marker of the endoplasm reticulum</td>
</tr>
<tr>
<td>Filipin</td>
<td>Sequencers cholesterol and disrupts lipid rafts</td>
</tr>
<tr>
<td>Fluorescent cholera toxin</td>
<td>Marker for lipid rafts and raft-mediated endocytosis</td>
</tr>
<tr>
<td>Fluorescent dextran conjugates</td>
<td>Marker for fluid phase endocytosis</td>
</tr>
<tr>
<td>Fluorescent Dil-DLD</td>
<td>Marker for late endosomes and lysosomes</td>
</tr>
<tr>
<td>Fluorescent fusion proteins (GFP)</td>
<td>Markers for clathrin, caveolin, other proteins, and organelles</td>
</tr>
<tr>
<td>Fluorescent / neutralizing IgG</td>
<td>Immunofluorescence labeling; microinjection of inhibitory IgG</td>
</tr>
<tr>
<td>Fluorescent transferrin</td>
<td>Marker for early endosomes / endocytic recycling complex</td>
</tr>
<tr>
<td>Hypertonic sucrose</td>
<td>Conditions that disrupt clathrin coated pits</td>
</tr>
<tr>
<td>Ikarugamycin</td>
<td>Inhibitor of clathrin-mediated endocytosis</td>
</tr>
<tr>
<td>Intracellular potassium depletion</td>
<td>Disrupts actin polymerization and phagocytosis</td>
</tr>
<tr>
<td>Latrunculin A</td>
<td>Conditions that disrupt clathrin coated pits</td>
</tr>
<tr>
<td>Lucifer yellow</td>
<td>Fluorescent marker for fluid phase endocytosis</td>
</tr>
<tr>
<td>Lysotracker and lyssensor</td>
<td>Fluorescent markers for lysosomes</td>
</tr>
<tr>
<td>Media temperature ≤ 10 °C</td>
<td>Metabolic inhibitor</td>
</tr>
<tr>
<td>Methylene</td>
<td>Weak base; raises endosomal pH</td>
</tr>
<tr>
<td>Methyl-β-cyclodextrin</td>
<td>Sequencers cholesterol and disrupts lipid rafts</td>
</tr>
<tr>
<td>Monensin</td>
<td>Ionophore; raises endosomal pH; blocks recycling</td>
</tr>
<tr>
<td>Monodansylcadaverine</td>
<td>Inhibitor of transglutaminase</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>Depolymerizes microtubules</td>
</tr>
<tr>
<td>Nystatin</td>
<td>Sequencers cholesterol and disrupts lipid rafts</td>
</tr>
<tr>
<td>Phorbol esters</td>
<td>Blocks receptor recycling; promotes macroinocytosis</td>
</tr>
<tr>
<td>Primeneurine</td>
<td>Weak base; raises endosomal pH</td>
</tr>
<tr>
<td>Phenytoxolamine</td>
<td>Metabolic inhibitor</td>
</tr>
<tr>
<td>RNAi</td>
<td>Targeted inhibition of clathrin, AP2, epsin, others</td>
</tr>
<tr>
<td>Sodium acid</td>
<td>Metabolic inhibitor</td>
</tr>
<tr>
<td>Wortmannin</td>
<td>Inhibitor of PI3 kinases</td>
</tr>
</tbody>
</table>

Examples of small molecules, altered cell culture conditions, proteins, and genetic constructs used to probe RME. CID: chemical inducer of dimerization.

Small molecules and modified cell culture conditions that block CME

Clathrin-mediated endocytosis can be blocked by several pharmacologic inhibitors, including the antipsychotic drug chlorpromazine (Thorazine), the natural product ikarugamycin, and the antiviral drug amantadine. The metabolic poisons phenylarsine oxide and sodium azide also block CME but additionally inhibit protein synthesis. Culture of cells under conditions that deplete potassium or calcium, treatment of cells with hypertonic sucrose, or acidification of the cytoplasm by addition of acetic acid to media, have also been used to block this mechanism of cellular uptake. However, many of these treatments tend to be relatively nonspecific and inhibit multiple cellular uptake processes. To block selectively dynamin-dependent endocytic mechanisms, a small molecule termed dynasore has been identified as a specific inhibitor of dynamin (42). Another strategy for inhibiting CME with high specificity uses a chemical inducer of dimerization (CID) combined with a genetic approach to promote aberrant oligomerization of clathrin in cells transfected with a clathrin fusion protein. This system can rapidly and reversibly inhibit 70% of the endocytosis of TFR (43).
Inhibitors of raft/caveolin-mediated endocytosis

Lipid raft domains of plasma membranes are enriched in cholesterol and sphingolipids. As a consequence, compounds that extract or sequester cholesterol, such as β-cyclodextrins, nystatin, and filipin, can block selectively endocytosis of cholera toxin, GPI-linked proteins, and other receptors that associate with lipid rafts and caveolae. However, cholesterol is also critical for CME, secretion of proteins, and the actin network. Therefore, conditions designed to affect selectively raft-mediated endocytosis by perturbing cholesterol levels must be carefully controlled to avoid disrupting other mechanisms of endocytosis (40).

Small-molecule probes of other endocytic trafficking pathways

Phosphatidylinositol-3-0H kinase [PI(3)k] plays an important role in the fusion of endosomes. Phosphatidylinositol-3-phosphate [PI(3)P], a product of this enzyme, is enriched in early endosomes, and blocking PI(3)k activity with the small molecule wortmannin prevents endosome fusion. This fungal natural product has been shown to inhibit the endocytosis of transferrin, horseradish peroxidase, and albumin (44, 45).

Fluorescent probes of endocytosis

Fluorescent small molecules and proteins represent powerful tools for labeling ligands, receptors, and other targets involved in endocytosis. By conjugating ligands of receptors to fluorophores, the uptake of these molecules can be analyzed by fluorescence microscopy, flow cytometry, and other methods. Small-molecule fluorophores have been installed through site-specific protein labeling on modified cell-surface receptors, such as TFR (46), NK1 (47), and EGFR (48), which are expressed in transfected cell lines. Studies of trafficking of the TFR/TFR complex that combine site-specific protein labeling with analysis by fluorescence resonance energy transfer (FRET) have demonstrated that this method is a powerful tool for studying endocytosis and exocytosis (46).

Green fluorescent protein (GFP) and related fluorescent proteins can be used to label practically any protein or subcellular compartment of living cells (49). Transfection of cells with plasmids that encode appropriately targeted fluorescent fusion proteins has been used to define the plasma membrane, early endosomes, late endosomes, caveolae, the golgi complex, the ER, and other subcellular locations. Several fluorescent small molecules are also available for labeling specific cellular organelles, including endosomes and lysosomes, for analysis by fluorescence microscopy.

Antibody probes and regulators of endocytosis

Immunofluorescence techniques are often used to identify specific cellular targets, including proteins, involved in endocytosis. However, because antibody reagents are not cell permeable, cells typically must be fixed or microinjected to allow binding to intracellular proteins. For example, microinjection of antibodies against clathrin (50) and dynamin (51) has been used to block CME. However, because of their cell impermeability, antibody reagents are limited in studies of dynamic cellular processes (52).

Genetic approaches for targeting specific components of endocytic pathways

Genetic methods, such as the expression of dominant negative proteins and RNA interference (RNAi), represent some of the most highly specific approaches for studies of endocytosis. Dominant negative mutants of clathrin, dynamin, Eps15, and other components of the endocytic machinery have been used widely to probe endocytic pathways. More recently RNAi has emerged as an important new tool for downregulating specific targets, including clathrin, AP-2, and epsin (53). Although these genetic methods have the potential to inactivate proteins in cells with a high degree of specificity, they also have associated limitations. For example, RNAi against clathrin heavy chain and AP-2 can suffer from cross-reactivity with other endocytic pathways (54). Another disadvantage of RNAi is that it can require several days to eliminate the targeted protein, which allows the activation of alternative compensatory mechanisms. The identification of temperature-sensitive alleles that result in inactivation of a specific protein at a defined temperature can also provide powerful tools for studies of dynamic cellular processes such as RME. However, these systems can be difficult to implement in mammalian cells.

Synthetic Mimics of Ligands and Receptors

Mimics of ligands and receptors have been used to promote the endocytic uptake of drugs, proteins, DNA, and other cell-impermeable molecules. Ligands linked to cargo of interest can often access efficiently the cell interior via binding to internalizing cell-surface receptors. Modified natural cell-surface receptors and small synthetic mimics of receptors have been shown to promote the uptake of specific ligands via RME. Examples of mimics of ligands and receptors are shown in Fig. A.

Cellular uptake of cargo conjugated to ligands of internalizing receptors

Numerous ligands of cell-surface receptors have been linked to cell-impermeable macromolecules, drugs, and other cargo for delivery into mammalian cells. Macromolecular ligands modified in this way include transferrin (5), LDL (22), growth factors, and antibodies that bind the extracellular domains of cell-surface receptors. An example of an FDA-approved drug that functions in this way is mylotarg, which is a monoclonal antibody drug that comprises the anticancer agent calicheamicin linked to a humanized antibody that binds the CD33 antigen on myeloid leukemia cells (55). Binding of mylotarg to CD33 results in endocytosis of the antibody drug conjugate, release
Receptor-Mediated Endocytosis

Figure 4 Examples of mimics of ligands and receptors used to deliver cargo into cells. (a, b) Ligands such as folate can be linked to macromolecules (e.g., nucleic acids and proteins) and small molecules (e.g., drugs, radiochemicals, and fluorophores) to promote cellular uptake. (c) Peptides and small molecules linked to $N$-alkyl derivatives of $3\beta$-cholesterylamine or $3\beta$-dihydrocholesterylamine mimic cell-surface receptors by cycling rapidly between the plasma membrane and the intracellular endosomes. The synthetic Fc receptor shown in (c) enables treated mammalian cells to internalize human IgG.

of calicheamicin in endosomes, escape of the drug into the cell nucleus, and antiproliferative effects against targeted cells.

The vitamins folate (vitamin B9) (56, 57) and cobalmin (vitamin B12) (58) have been investigated as vehicles for delivery of impermeable molecules into cells. By binding with high affinity ($K_d \sim 10^{-10} M$) to folate receptors, folate-linked drugs, radio-pharmaceuticals, nucleic acids, and nanoparticles can become internalized by RME. Similarly, transcobalamin receptors bind and promote the endocytosis of cobalamin and cobalamin conjugates complexed to soluble transcobalamin (59). This approach has been used to target selectively tumor cells that overexpress these receptors.

Artificial cell-surface receptors

Synthetic mimics of cell-surface receptors have been constructed using plasma membrane anchors derived from $N$-alkyl-3$\beta$-cholesterylamine and related compounds (14). Synthetic receptors comprising protein and other binding motifs linked to this membrane anchor become incorporated rapidly into plasma membranes of mammalian cells. By constitutively cycling between the plasma membrane and the endosomes, cells treated with these compounds gain the capacity to endocytose cell-impermeable ligands. For example, the synthetic Fc receptor shown in Fig. 4 enables human cells that lack Fc receptors to internalize human IgG (60). This strategy, termed synthetic receptor targeting, seems to mimic the initial steps of uptake of cholera toxin mediated by ganglioside GM1. This approach may have applications as a method for drug delivery (61).

A related strategy, termed cellular painting, has been used to incorporate proteins linked to GPI lipid anchors into cellular plasma membranes (62, 63). This approach has been used to study cellular signaling, plasma membrane organization, and immunologic responses to modified cell surfaces. Because GPI-linked proteins undergo raft-mediated endocytosis, Jurkat lymphocytes treated with a GPI-linked variant of the immunoglobulin Fc receptor FcγRIII will endocytose ligands that bind this receptor (64). Single-chain antibodies covalently linked to lipids have also been used to construct related cell-surface receptors (65).

Metabolic cell-surface engineering to promote the endocytic uptake of ligands

Metabolic cell-surface engineering has been used to modify carbohydrate components of cell-surface receptors and to enable endocytic uptake of impermeable ligands (66). By feeding cells unnatural sugars, cellular metabolism can be harnessed to display bioorthogonal functional groups, such as ketones and azides from glycoproteins and glycolipids, on cell surfaces. The reaction of these ketones with hydrazine derivatives to yield
Future Research Directions

Molecular and cellular probes of RME

The molecular mechanisms controlling CME are beginning to be characterized. However, other mechanisms of RME, particularly those involving lipid rafts, caveolae, and other pathways, are less well understood. Although genetic methods such as RNAi and expression of dominant negative proteins represent important tools for inactivating key players in endocytic pathways, the delayed cellular response associated with many of these approaches can allow compensatory gene expression or other mechanisms that can complicate the analysis. Because small molecules can inactivate rapidly defined protein targets, the identification of highly specific inhibitors of proteins involved in RME may provide the best tools for studying the molecular mechanisms of this process. Screens based on “composite synthetic lethality” are particularly promising for the identification of small molecules that target specific membrane trafficking pathways (69).

Mimicry and modifications of ligands

Mimics of ligands that bind cell-surface receptors and undergo RME are becoming an increasingly important class of targeted therapeutics. For example, folic acid derivatives show substantial advantages of green fluorescent and related fusion proteins, including their high molecular weight and relatively low intrinsic brightness as well as the reaction of azides with modified phosphines in the Staudinger ligation (68), can immobilize molecules on the cell surface and promote delivery of proteins such as the toxin ricin (67).

References


Further Reading

See Also
Biomacromolecule-Directed (Target-Specific) Drug Delivery: Underlying Factors, Principles and Design; Chemical Events in Neurotransmission; Endocytosis and Exocytosis, Membrane Dynamics of; Lipid Domains, Chemistry of; Receptor Tyrosine Kinases; Receptor-Ligand Interactions; Receptors, Chemistry of; Signal Transduction Across Membranes; Virus-Based Drug Delivery
Inorganic Chemistry in Biology
Ivano Bertini and Paola Turano, University of Florence, Florence, Italy
doi: 10.1002/9780470048672.wecb250

Several inorganic chemical elements play fundamental roles in biological processes. The contribution of inorganic chemistry to the understanding of biological processes is presented here from a historical perspective: from the first discoveries of metal ions in living organisms to the modern approaches of inorganic structural biology and bioinformatics, through the characterization of metal binding sites in proteins and in biomimetic model compounds. Definitions are provided for the fundamental concepts of metal cofactor, metalloprotein, and metalloenzyme.

The importance of inorganic chemistry for chemical biology is based on the distribution of the elements of the periodic table in living organisms, which is summarized in Fig. 1. Amino acids, nucleic acids, carbohydrates, lipids, organic cofactors (e.g., ATP, ADP, and NADH), and metabolites are composed of the six bulk biological elements: carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. The chemistry of these elements in living systems is the subject of classic biological chemistry. However, it has been established that at least 20 elements other than C, H, N, O, P, and S are essential for life, even though they are considered generally “inorganic” (Fig. 1). The first metal-containing species identified in living systems were pigments such as chlorophylls and hemoglobin, whose discovery dates back to the nineteenth century. To complement the work aimed at the chemical characterization of these molecules, several studies showed that some metal ions, such as zinc and iron, are essential for life. Hemoglobin, myoglobin, and cytochromes were among the first biological macromolecules to be investigated at the molecular level because of their abundance in living organisms and their intense color, which eased their detection. Nevertheless, it was only in the second half of the twentieth century that the contribution of inorganic chemistry to the characterization of biological systems translated into a well-defined discipline, named bioinorganic chemistry (and later also called biological inorganic chemistry). The establishment of this new branch of science paralleled advancements in spectroscopic tools, in structural methods, and in the development of inorganic chemistry as the chemistry of coordination compounds. This ensemble of approaches has provided momentum and has given great impetus to the characterization of the binding mode of metal ions in biological macromolecules and to the understanding of their reactivity. Two primary approaches have guided the study of metal binding sites in proteins: 1) the synthesis and characterization of model compounds and 2) the direct characterization of metalloenzymes and metalloproteins.

Metal Cofactors

Typically, metal ions are cofactors that function as catalytic centers in several fundamental biological reactions, play a role in electron transfer reactions, or impart structural stabilization to the macromolecular fold. Proteins offer such a large variety of metal binding sites associated with widely disparate functions that biological inorganic chemistry focuses primarily on the study of metal ions in proteins. Other areas of research in the field include the metal–RNA and metal–DNA interactions. However, these aspects are not addressed here, as the chemistry largely electrostatic in nature. On the contrary, the ability of proteins to bind metal ions is related essentially to the presence of amino acid side chains that can act as metal ligands. Figure 2 represents the possible binding modes of these side chains. In some cases, the coordination geometry of the metal ion is completed by exogenous ligands such as H₂O or OH⁻, or by protein backbone amides or carbonyls.

Traditionally, when the association between the metal ion and the protein is relatively strong (i.e., binding constant higher than 10⁸ M⁻¹) the complex is called a metalloprotein. When the protein is performing catalytic activity at the metal center, it is called a metalloenzyme.

Although the different sequences and folds of proteins provide the most disparate metal binding sites (some examples of which are provided in Fig. 3), Nature has evolved to select other organic or inorganic ligands for metal ions in proteins, which we call "special metal cofactors." These cofactors can be grouped into two broad classes: tetrapyroles and metalloclusters.

Tetrapyroles are macrocyclic ligands that provide a common skeleton to hemes (that contains iron), chlorophylls (that contains magnesium), corrinoids (that contains cobalt), siroheme (that contains iron), and methanogenes factor F430 (that contains nickel).
Iron-porphyrins are a widespread group of tetrapyrroles present in heme proteins (1). They are all derived from protoporphyrin IX, but they have different substituents (Fig. 4). The so-called heme b coincides with protoporphyrin IX. Heme o differs from heme b by the presence of a farnesylhydroxyethyl group at position 2, and it differs from heme a by the presence of a farnesylhydroxyethyl group at position 2 and a formyl group at position 8. In heme c, the two vinyl groups are substituted with thioether bridges that involve sulfhydryl groups of protein cysteine residues, which results in two covalent linkages between the porphyrin and the protein matrix. Heme P460 can be viewed as a c-type heme with an additional covalent bond between the \(\alpha\)-meso position and a \(C_\varepsilon\) of a nearby tyrosine ring. Heme d, heme d1, and siroheme have unconjugated pyrrole rings. The common oxidation states for iron in heme proteins are +2 and +3, although the iron(IV) = 0 moiety
Metalloclusters consist of at least two metal ions associated with inorganic and/or otherwise nonprotein ligands. Binuclear centers in which the metal ion ligands are composed of H₂O/OH⁻ ligands in addition to amino acid side chains are excluded from the definition of a metallocluster. Several metalloclusters have been characterized structurally in metalloproteins. Among the most abundant metalloclusters are those of the iron-sulfur family (2, 3). Iron-sulfur clusters are characterized by iron ions that exhibit almost exclusively tetrahedral coordination to donor sulfur atoms, provided by either Sγ of cysteines or by bridging sulfides (sometimes called inorganic sulfurs). Metalloproteins contain basic cluster types with two, three, or four iron ions, as depicted in Fig. 5. In this figure, the mononuclear iron center of rubredoxin is shown: This iron center is not a cluster but can be considered as the “prototype” of the tetrahedral iron units that constitute the Fe₅S₅ clusters. The simplest type of iron-sulfur cluster is represented by the diamond structure of the Fe₂S₂ center, in which the two iron ions are coordinated to two bridging sulfides. Each iron is then bound to two Cys or two His ligands (the latter in the case of Rieske proteins, Fig. 5). The Fe₅S₅ unit is constituted by four iron ions and four sulfide ions arranged in a cubane structure. A gain, the tetra-coordination of each iron ion is accomplished on binding of the sulfur of a Cys protein residue. The Fe₅S₅ cluster seems to be derived from the Fe₅S₅ cluster by the removal of one iron. Rubredoxins are electron transfer proteins in which the iron oxidation state cycles between +3 and +2. In the clusters, the number of potential oxidation states increases with the number of metal ions. Assuming that each iron can exist formally in the ferric and ferrous states, an iron-sulfur cluster with n iron can exhibit a maximum number of n + 1 oxidation states. Nevertheless, only a few oxidation states have been observed in proteins. For instance, in Fe₂S₂ systems, only two of the possible three oxidation states are found: the one that contains two ferric ions [i.e., the (Fe₂S₂)⁺ state] and the one that contains one ferric and one ferrous ion [(Fe₂S₂)⁺]. This type of cluster is typical of electron transfer proteins or of electron transfer centers in multidomain redox metalloenzymes. The Fe₅S₅ cluster in which the iron coordination is completed by two Cys per iron is present in ferredoxins.

In recent years, X-ray crystallography has led to the discovery of several novel metalloclusters of complex architecture that contain at least four metal ions (4, 5). They represent the active site of several redox enzymes that contain molybdenum, nickel, and manganese, as well as the most commonly encountered iron and copper (Fig. 6). These enzymes are extremely specialized in the oxidation or reduction reactions of the smallest molecules and anions (which include N₂, CO, and H₂). A common feature of such clusters is that they are present in enzymes as part of a more extensive electron transfer chain that involves a series of
metallocenters (often heme and iron-sulfur centers) that serve to carry electrons into and out the active site. For several of them, extensive spectroscopic and functional studies are required to unravel oxidation states, substrate binding sites, and reaction mechanisms at these sites.

The first and second coordination sphere modulates the reactivity of metal ions so that the functional roles of a given ion may be largely different, as summarized in Table I. A good example of the effects of the first and second coordination sphere on modulating the reactivity of metal cofactors in proteins is provided by heme proteins (1). Heme-iron in heme proteins can be five-coordinate or six-coordinate. Axial ligation seems to be related strongly to protein function. Six-coordination (His/His or His/Met) is typical of cytochromes and electron transfer proteins in which the heme iron has to cycle between the iron(III) and the iron(II) oxidation states for its function. Usually, penta-coordination is found in globins that bind oxygen and in redox heme-enzymes. In globins, the imidazole ring of the “proximal” His residue provides the fifth heme iron ligand; the other axial heme iron position remains essentially free for O₂ coordination. In heme-enzymes, the His has a sixth coordination position is available for substrate binding. The basicity of the fifth (or proximal) His ligand modulates the redox potential of the heme iron. In peroxidases, the His has a strong imidazolate character because of a strong H-bond between the Nδ₁ of the imidazole ring and a nearby Asp residue; this bond facilitates the higher oxidation states for the heme iron (6). The resting state of these enzymes contains iron(III), and higher oxidation states are reached during the catalytic cycle. In globins the His is essentially neutral, with a weak H-bond between the imidazole and a backbone carbonyl, and these proteins are commonly in the iron(II) state.

Model Compounds

The discovery of metal centers in metalloproteins has stimulated inorganic chemists to synthesize compounds capable of mimicking the spectroscopic and functional properties of the protein metal cofactors. Biomimetic model compound chemistry has flourished since the 1970s, when several advantages existed in the study of the model compound instead of the protein itself. The much smaller size of the model compound made it more suitable for biophysical studies, and advancements in chemical synthesis offered opportunities to play with ligands and metal geometries. Some notable examples of the synthetic analog approach deal with iron-porphyrin model compounds (and metalloporphyrins in general) (7), iron-sulfur clusters (3), and model compounds aimed at reproducing the unusual spectroscopic and electrochemical properties of blue copper proteins (8). In more recent times, synthetic models have been developed to reproduce the characteristic features of the binuclear Cu₄ electron-transfer center of cytochrome c oxidase (9) (Fig. 3e), of high-nuclearity MoFe₅S₇ clusters of the cofactors of nitrogenases (10) (Fig. 6a), and of the Cu₄ center of nitrous oxide reductase (11) (Fig. 6b).
Table 1. Main biologic functions of metal ions

<table>
<thead>
<tr>
<th>Element</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium</td>
<td>Structure stabilization—Signaling—Enzyme activator</td>
</tr>
<tr>
<td>Cobalt</td>
<td>Redox catalyst</td>
</tr>
<tr>
<td>Copper</td>
<td>ET—Redox catalyst—O₂ carrier</td>
</tr>
<tr>
<td>Heme iron</td>
<td>ET—Redox catalyst—O₂ carrier</td>
</tr>
<tr>
<td>Iron in Fe₅ clusters</td>
<td>ET—Redox catalyst—O₂ carrier</td>
</tr>
<tr>
<td>Nonheme nonFe₅ iron</td>
<td>Redox catalyst—O₂ carrier</td>
</tr>
<tr>
<td>Magnesium</td>
<td>Enzyme activator—DNA structure stabilization—Ribozymes</td>
</tr>
<tr>
<td>Manganese</td>
<td>Redox catalyst—Activator of hydrolases, ligases and transferases</td>
</tr>
<tr>
<td>Molybdenum</td>
<td>Redox catalyst (nitrogense, oxotransferases)</td>
</tr>
<tr>
<td>Nickel</td>
<td>Hydrolitic activity—Redox catalyst</td>
</tr>
<tr>
<td>Potassium</td>
<td>ATPase pumping—Structure stabilization—Electrostatic effects on the reduction potential of distant centers</td>
</tr>
<tr>
<td>Rubidium</td>
<td>Function similar to that of K in some plants</td>
</tr>
<tr>
<td>Selenium</td>
<td>Redox catalyst</td>
</tr>
<tr>
<td>Sodium</td>
<td>ATPase pumping—Structure stabilization—Electrostatic effects on the reduction potential of distant centers</td>
</tr>
<tr>
<td>Sulfur</td>
<td>Function similar to that of Ca in some plants</td>
</tr>
<tr>
<td>Tungsten</td>
<td>Replaces Mo in some hyperthermophilic archea</td>
</tr>
<tr>
<td>Vanadium</td>
<td>Peroxidase and nitrogenase activity in some algae and possibly fungi</td>
</tr>
<tr>
<td>Zinc</td>
<td>Hydrolitic activity—Structure stabilization—Transcription factor</td>
</tr>
</tbody>
</table>

Today, the advancements in spectroscopic, structural, and biological tools make it easier than ever before to study metal centers in their biological context. As a consequence, the impact of model compounds has been reduced; although from an inorganic point of view, biological metal centers may represent a synthetic challenge and are relevant for science beyond their importance as model compounds.

The synthetic approach is also suffering from the intrinsic difficulties encountered in the design of ligands capable of mimicking the secondary coordination sphere effects and...
Inorganic Chemistry in Biology

Figure 6. (a) The FeMo-cofactor of nitrogenases: The overall stoichiometry is MoFe$_7$S$_9$, and it can be viewed as a MoFe$_3$S$_3$ cluster and a Fe$_4$S$_3$ cluster, bridged by three sulfides. The molybdenum is also bound to a homocitrate molecule. (b) The C-cluster of carbon monoxide dehydrogenase: This can be viewed as a Fe$_4$S$_4$ cluster bridged to a binuclear Ni-Fe center. (c) The oxygen-evolving center of photosystem II: A cubane-like Mn$_3$CaO$_4$ cluster is linked to a fourth Mn ion (Mdh) by a µ-oxo bridge. (d) The tetranuclear Cu$_Z$ center of nitrous oxide reductase: Three of four copper ions bind two His, the fourth binds a single-His, and it has been suggested that it represents the substrate coordination site.

Biogeochemical Cycles

The second half of the twentieth century has observed the illustration of many of the central reaction steps and enzymatic catalysts of biogeochemical cycles, which led to the discovery of key complex cofactors that contain transition metal ions. Indeed, the key catalysts in the global cycles of oxygen (13, 14), nitrogen (15–18), carbon (19–24), sulfur (25–32), and hydrogen (33, 34) are redox metalloenzymes that contain unique metal cofactors. The cycles are represented in Figure 7 and describe the natural transformations of several "inorganic" molecules and ions. The biogeochemical cycles of C, O, N, S, and N are interlinked intimately. The O$_2$ produced in the oxygen cycle serves as an oxidant for the reduced compounds of C, N, and S by both biological and human activities. H$_2$ is a key carrier to reduce equivalents in the anaerobic world. The full understanding of these enzymatic mechanisms represents a challenge for biological inorganic chemistry at present.

The occurrence of metalloproteins

Between 1995, when the first genome (H. influenzae) was sequenced and, when the first draft of the human genome was published, a revolution in the approach to the study of gene products (i.e., proteins) occurred. This revolution has led to so-called post-genomic research. Genome sequencing projects provide researchers with lists of all the proteins that an investigated organism can produce and their amino acid sequences. From a bioinorganic point of view, however, the question is which proteins need a metal ion to perform their physiologic function. This question is fundamental, and the answer cannot be derived from genomic information alone. Bioinformatic tools have been developed to identify metalloproteins in genome databanks (35–37). This method relies on the exploitation of known metal binding-patterns (MBPs), which are available experimentally from the three-dimensional structures deposited in the protein data bank. MBPs are strings of the type A$.X$.B$.X$.C...$, where A, B, C... are the amino acids that act as metal ligands, and n, m... are the number of amino acid residues in between two subsequent ligands. Gene banks are then browsed to search for MBPs and primary structure.
Inorganic Chemistry in Biology

Figure 7  Simplified global biogeochemical cycles. (a) Combined C and O cycles. The biomass formed by photosynthesis can be transformed anaerobically by bacteria and fungi to produce small C-containing molecules such as CO, CO\textsubscript{2}, formate, and acetate. These products can be transformed (anaerobically) into methane by archaea. CH\textsubscript{4} can then be converted aerobically into CO\textsubscript{2}. (b) The hydrogen cycle, with the indication of the nature of the metal centers at the catalytic site of the involved enzymes. (c) The sulfur cycle. The top part of the external cycle (thick arrows) summarizes the key steps of aerobic sulfide oxidation; the bottom part summarizes the steps of sulfate respiration. Thiosulfate, trithionate, and tetrathionate may also represent possible intermediates for this cycle. The inner cycle describes the cycling between HS\textsuperscript{-} and S\textsuperscript{0} performed by different organisms with respect to those involved in the outer cycle. The thin arrows that connect SO\textsubscript{2}\textsuperscript{2-}, H\textsubscript{2}SO\textsubscript{3}\textsuperscript{-}, and H\textsubscript{2}S represent the essential steps of an assimilatory pathway. (d) The nitrogen cycle. Denitrification is the anaerobic use by certain bacteria of nitrogen oxide species (NO\textsubscript{3}\textsuperscript{-}, NO\textsubscript{2}\textsuperscript{-}, NO, and N\textsubscript{2}O) as terminal electron acceptors instead of O\textsubscript{2}. The final product N\textsubscript{2} is released into the atmosphere; therefore, the process is referred to as dissimilatory. The nitrogen fixation reaction is carried out by the enzyme nitrogenase. The lower part of the cycle represents the assimilatory nitrate reduction that leads to the incorporation of reduced nitrogen species into biomass. The upper part of the cycle describes the nitrification process (i.e., the oxidation of reduced nitrogen compounds in the presence of oxygen).

Such an approach has been used successfully to identify all zinc, nonheme iron, and copper proteins. The human genome encodes approximately 2800 zinc-proteins, 250 nonheme-iron proteins, and 100 copper-proteins. These figures correspond to 10%, 0.8% and 0.3% of the human proteome, respectively. Corresponding averages for five eukaryotes of known genome sequence are 8.8%, 1.1%, and 0.3%, respectively and for 40 selected bacterial organisms are 3.9%, 4.9%, and 0.3%, respectively.

Metal homeostasis

It is well established that for each metal ion a dose-dependent effect exists (Fig. 8). Still, the optimal concentration for healthy organisms should not be considered as the concentration of the free metal ions, which is generally extremely low, but rather as the concentration of the metal ions bound to proteins and other metal ligands. The way in which organisms uptake and control the trafficking of metal ions has attracted the attention of researchers in recent times (38–40). This research has led to the identification of several metallochaperones. No established molecules that serve this function were known before 1997, but today metal trafficking pathways have been identified for metal ions such as copper, manganese, and zinc, although at different degrees of understanding. The comprehension of such processes requires the identification of the proteins involved in metal uptake; transfer and incorporation into the final metallo-enzymes; and identification at the molecular level of the factors that control the specificity, selectivity and efficiency of the mechanisms. Even essential elements, at high concentrations, can become toxic. Organisms have developed mechanisms to detoxify from nonessential metal ions or from excess of essential metal ions.
(40): The study of these pathways is another aspect of the quest to understand metal homeostasis.

**Metal-based drugs**

A special application of inorganic chemistry to biological problems concerns the use of metal-containing compounds as therapeutic or diagnostic agents. The greatest success of metallotherapeutics dates back to 1978, when cisplatin was first approved to treat clinical testicular and urinary tumors after the discovery in 1965 that cisplatin inhibits cell division (41). The biological target of cisplatin is DNA, where its primary binding site is the major groove N7 of guanine. The bending and local unwinding of the DNA double helix induced by cisplatin binding causes the loss of important structural motifs to recognize and to process damaged DNA. The need for a new anticancer drug that can overcome the limitations of cisplatin (high toxicity, activity against a wider range of cancer types, resistance to cisplatin after repeated treatment) has prompted chemists to develop new generation platinum drugs or compounds that contain different metal ions (e.g., palladium, ruthenium, gallium). Metal complexes are also employed to treat other diseases, for example gold(I) compounds are used as antiarthritic drugs.

Metal complexes are used largely in imaging and diagnostics. Several radiolysis products of metal ions (e.g., Ga, In, Tl, Tc) are suitable as radiodiagnostic and radiomaging agents. Finally, gadolinium (III) compounds are largely used as contrast agents for magnetic resonance imaging because of the effectiveness of this metal ion to relax H2O protons.

**Perspectives**

It is clear from the discussion of metal homeostasis that the full characterization of a protein requires not only its study in the isolated form but also the investigation of its relevant interactions. Therefore, the interactions of metal-binding proteins with small molecules and other proteins are an obvious perspective in biological inorganic research. This implies the importance of the identification and structure characterization of weak protein–protein complexes that represent the key steps of the biochemical processes discussed here. Moreover, metal-mediated protein–protein interactions are just beginning to be identified and will become an important field of research.

**References**


sulfite oxidase deficiency from the structure of sulfite oxidase.


Nature: A Model System for Chemists

Ronald Breslow, Department of Chemistry, Columbia University, New York, New York
doi: 10.1002/9780470048672.wecb378

Chemists have learned to imitate the principles of biologic chemistry in the quest to produce artificial enzymes, artificial receptors, and, ultimately, artificial living cells. The field of natural products chemistry is concerned chiefly with discovering the structures of the molecules produced in living systems as well as the paths by which they are produced. Biomimetic chemistry takes nature as a model and extends what exists in biology into what is possible. This process includes using hydrophobic effects in water to achieve selective reactions and making new amino acids and new nucleotides to generalize the properties of proteins and nucleic acids. Potent new catalysts are based on the chemistry performed by coenzymes using bifunctional acid–base catalysis and the water exclusion achieved by natural enzymes. Exciting new areas include molecular machines, mimics of the chemical senses, and work on the origin on earth of the homochirality observed in biomolecules. The studies using nature as a model have enriched chemistry greatly and furnished interesting insights into biochemistry.

Since the beginning of time, people have learned about the world by observing nature and by imitating aspects of what we have learned. This observation is true particularly of what we see in living creatures. For example, people observed birds and insects in flight and dreamed of being able to fly. Our most important clue was wings. We adopted this idea in the successful invention of airplanes. However, birds and insects power their flight by flapping the wings; this concept proved less useful. Nature has limitations, and one of them is that a powered propeller—and even more so a jet engine—were not practical for living creatures. Thus, we did not mimic birds slavishly; we adopted the central idea of their ability to fly and added ways to power the flight that nature could not achieve. As Philip Ball said (1): “A jumbo jet is not just a scaled up pigeon.”

Chemists have also taken many lessons and inspirations from natural chemistry; but, again, they have adapted their ideas so as not to be limited by the special requirements of natural chemistry. For example, nature designs proteins in three dimensions by folding linear polypeptide chains. This design indicates that genetic information is one dimensional, as is the sequence of nucleotides in genes. Although it might be interesting to use such a scheme in a mimic of an enzyme, chemists generally plan and then synthesize their three-dimensional structures, an option not open to nature.

Natural Products

In this brief article, I will focus on the ways we have learned from the chemistry of the life process itself. However, I must first call attention to another part of natural chemistry that has an effect on chemical thinking: the field called the Chemistry of Natural Products. Chemists have explored nature and have discovered several special types of chemicals: sugars, fats, terpenes, alkaloids, acetogenins, amino acids, heterocyclic coloring matters, macrocyclic lactones, steroid hormones, and so on. These compounds have been produced by living organisms and they are secreted by or isolated from them. This field is concerned with the substances of nature and the biochemical processes by which they are formed.

Chemists have devoted much effort to exploring this natural world of chemistry as well as to determining structures; the natural world has stimulated the extension of the chemical world into models and analogs of the natural chemicals. The field of organic chemistry was influenced heavily by the types of chemical structures found in natural products; many medicinal compounds are still invented by using natural products as models for analogs. Chemists have also invented important polymers once nature showed us the natural polymeric carbohydrates, polypeptides, nucleic acids, and the polymers such as rubber that are produced from natural materials.
Biomimetic chemistry

The field concerned directly with imitating the chemistry of life processes has been called biomimetic chemistry or bio-inspired chemistry. The two terms differ in how close the mimic is to the natural process. As the author of this article, I will favor biomimetic, a word I coined (3).

Chemistry in water

A most important difference between human-made chemistry and living chemistry is that the latter occurs in water. Most chemical processes use organic solvents, which dissolve the components better; but water has special properties that are appreciated increasingly. Most significantly, the hydrophobic effect in water is the force that causes membranes, micelles, and liposomes to form, which pushes hydrocarbon chains together to diminish the high-energy hydrocarbon/water interface. The hydrophobic effect is also a major reason that hormones bind to receptors and substrates bind to enzymes, and it causes the folding of polypeptide chains into the three-dimensional structure of proteins by promoting the clustering of hydrocarbon side chains in the protein interior away from water.

Chemists now use water as a solvent for synthetic reactions, which takes advantage of the selectivities that the hydrophobic effect can induce (4). Many enzyme mimics (vide infra) also use water as the medium to promote substrate binding into the catalyst. Of course, water as a solvent also has important environmental advantages over volatile organic solvents. It remains to be seen how much impact the special properties of water as a solvent will have within chemical synthesis and manufacturing.

Molecular recognition and self-assembly

Small molecules such as hormones and allosteric effectors can bind to proteins selectively — including enzymes — and can modify their properties. Indeed, such binding is a common mechanism for the regulation of biologic effects; living cells do not simply run all processes as rapidly as possible. These natural processes have engendered a biomimetic field of molecular recognition and self-assembly, in which chemists develop such selective binding systems for analysis and as part of molecular devices. As mentioned above, medicinal chemists invent new molecules to bind to important receptors and enzymes, which imitate the natural ligands.

Proteins are assembled in living cells from amino acids, guided by the information in DNA that is passed on through RNA. Generally, chemists are inspired to generalize and extend such natural processes. A particularly interesting example is the synthesis of unnatural amino acids and the development of ways in which they can be incorporated into new kinds of proteins by modifying the biologic protein synthesis machinery (5). Some of these novel protein types promise to have new useful properties.

Nature appreciation

Another kind of generalization of natural chemistry involves viewing the natural chemistry in context by observing how well other related structures could perform the function of their natural models. In other words, why were these particular biologic molecules selected? A good example is the study of generalized forms of DNA.

In one example, some of the heterocyclic bases were modified drastically and were replaced by substituted benzene rings, and yet the base-pairing process of DNA was still observed (6). In other studies, the deoxyribose of DNA was replaced by other sugars, in some cases with an improvement in binding properties (7). As such, the 2-deoxyribose in DNA was replaced by the isomeric 3-deoxyribose to observe how the binding properties were affected (8). Interestingly, the resulting isoDNAs did not pair well with their isoDNA partners, or with the normal DNA partner that had the appropriate base sequences. This result clarified why 2-deoxyribose was selected for DNA even though the 3-deoxy isomer could be formed more easily under prebiotic conditions. However, the isoDNA base did pair well with its conjugate RNA, which furnishes insight into the difference between DNA–DNA pairing and DNA–RNA pairing (9).

The coenzyme thiamine pyrophosphate (1) plays a central role in many parts of metabolism (Fig. 1). Its mechanism of action involves the formation of a thiazolium zwitterion 2 that was stabilized by a carbene resonance form 3 (10). This discovery opened up studies of the chemistry that such “stabilized carbene” could catalyze, as chemists realized that the otherwise impossible chemistry that thiamine pyrophosphate catalyzes in nature could be generalized and adapted for useful synthetic processes.

However, another study was an example of nature appreciation—the structure of thiamine was varied to learn what was special about the particular thiazolium derivative that was natural thiamine (11). As a chemical catalyst—ignoring the question of what effect changes would have on the ability of the coenzyme to bind to the proteins that have evolved to use it—thiamine proved to be the optimal relative to other related structures because of a balance of catalytic ability and chemical stability. The anion 4 derived from an imidazolium ring instead of a thiazolium ring was a weaker catalyst but was more stable in water (10). The imidazolium anion and its dihydro derivative have proven to be very useful metal ion ligands, including...
Enzymatic selectivity involves geometric control

Enzymes operate by different rules than we normally use in synthetic chemistry. In nonenzymatic reactions, the intrinsic reactivity of the substrate dominates the chemistry. If we want to reduce a ketone, an aldehyde group in the same molecule will be more reactive. To obtain selectivity for the ketone group, we would need to block the aldehyde somehow. Moreover, if we want to oxidize a saturated carbon to form an alcohol, usually we would find the product alcohol oxidizes more readily, so the final product would be a ketone or aldehyde. Also, if we want to oxidize an isolated saturated carbon, a carbon–carbon double bond better not exist in the molecule or it would oxidize more readily. Generally, our synthetic procedures involve blocking such side reactions in some way. Enzymatic reactions have no such problems.

In the enzyme that oxidizes lanosterol to convert it ultimately to cholesterol (Fig. 2, three unactivated methyl groups are oxidized, whereas two double bonds and a secondary carbonyl are left untouched initially. This selectivity reflects geometric control in the enzyme, and it is perhaps one of the strongest lessons that chemists have learned from nature. The enzyme binds the substrate in such a position that the oxidizing group of the enzyme, an iron oxo species, can reach the hydrogens on the inactivated methyl groups but cannot reach the more reactive double bonds or secondary carbonyl (it is oxidized at some point later to permit decarboxylations, but then reduced again).

This general principle has inspired chemists to create enzyme mimics that perform biomimetic reactions that are also directed by geometric control (13). Selective attack on particular carbons of steroid substrates occurs as a function of the geometry imposed by binding, as in the enzyme. In one example, a carbon–carbon double bond was left untouched whereas a saturated inactivated carbon was oxidized, as in natural enzymes (14).

Of course, enzymes use geometry to control not only positions of reaction in a substrate but also selectivity in the formation of chiral centers and selectivity among substrates. It is fair to say that chemists were inspired to develop ways to produce optically active compounds as single enantiomers by observations of reaction in a substrate but also selectivity in the formation of chiral centers and selectivity among substrates. If we want to oxidize a saturated carbon to form an alcohol, usually we would find the product alcohol oxidizes more readily, so the final product would be a ketone or aldehyde. Also, if we want to oxidize an isolated saturated carbon, a carbon–carbon double bond better not exist in the molecule or it would oxidize more readily. Generally, our synthetic procedures involve blocking such side reactions in some way. Enzymatic reactions have no such problems.

In the enzyme that oxidizes lanosterol to convert it ultimately to cholesterol (Fig. 2, three unactivated methyl groups are oxidized, whereas two double bonds and a secondary carbonyl are left untouched initially. This selectivity reflects geometric control in the enzyme, and it is perhaps one of the strongest lessons that chemists have learned from nature. The enzyme binds the substrate in such a position that the oxidizing group of the enzyme, an iron oxo species, can reach the hydrogens on the inactivated methyl groups but cannot reach the more reactive double bonds or secondary carbonyl (it is oxidized at some point later to permit decarboxylations, but then reduced again).

This general principle has inspired chemists to create enzyme mimics that perform biomimetic reactions that are also directed by geometric control (13). Selective attack on particular carbons of steroid substrates occurs as a function of the geometry imposed by binding, as in the enzyme. In one example, a carbon–carbon double bond was left untouched whereas a saturated inactivated carbon was oxidized, as in natural enzymes (14).

Of course, enzymes use geometry to control not only positions of reaction in a substrate but also selectivity in the formation of chiral centers and selectivity among substrates. It is fair to say that chemists were inspired to develop ways to produce optically active compounds as single enantiomers by observing how nature produces single enantiomers of amino acids and sugars, and essentially every natural compound that can exist as two enantiomers. Again, the secret in enzymatic reactions is geometric control of the processes where the chiral centers are produced because of the geometry of the enzyme-substrate complex. Such enantiomeric selectivity is the target of much modern synthesis.

A particular example from our laboratory is observed in compound 5, in which a basic amino group held rigidly on a mimic of the coenzyme pyridoxamine phosphate is able to convert ketoacids to amino acids with high enantioselectivity (15). This reaction is modeled closely to the way in which transaminase enzymes achieve the same goal.

Enzymes can perform bifunctional acid–base catalysts

Any enzymes use coenzymes to achieve the detailed transformations they catalyze, but the enzyme proteins themselves also supply important elements of the catalysis. Enzyme proteins are the source of the entire catalytic effect when coenzymes are not involved. As a common process, acid and base groups in enzymes perform proton transfers that are critical to the catalytic mechanism. A particularly informative example is observed in the enzyme ribonuclease A, which catalyzes the cleavage of RNA (16). The catalytic process (Fig. 4) involves the imidazole ring of the amino acid histidine that removes the proton from the 2-hydroxyl of the ribose. A different protonated histidine transfers a proton to the RNA to promote the cleavage process. Studies with D2O–H2O mixtures established that the two proton transfers occur at the same time (17).

This reaction taught chemists the importance of simultaneous bifunctional catalysis, if it can be achieved. To avoid entropy problems, such bifunctional catalysts should have both catalytic groups in the same molecule, which are held in the correct position to participate in the simultaneous process. In ribonuclease A, the curling up of the peptide chain brings histidine-12 and histidine-119 close enough to let them participate in this way, even though they are separated by over 100 residues in the peptide chain. In a second step of the biologic process, the cyclic phosphate produced in the first step is hydrolyzed by the enzyme, again using the same two histidine side-chain imidazole groups.

We created a mimic of the cleavage of the cyclic phosphate by this enzyme by attaching two imidazole groups to a cyclodextrin molecule in well-defined positions (18). In water, the cyclodextrin bound a cyclic phosphate substrate—net, however, one derived from RNA—and performed the cleavage of the cyclic phosphate by using the two imidazoles as the enzyme does. That is, one functioned as a base, which delivered water...
to the phosphorus, and the other one, protonated, which functioned as a proton donor group in the hydrolysis reaction. The best catalyst in this process was one in which the proton was delivered to the phosphate anion group, which formed a five-coordinate phosphorane intermediate. In a subsequent fast step, this went to the final hydrolysis product, with the proton ending up on the OH group that is formed in the hydrolysis.

We also examined the reaction rate in different mixtures of D$_2$O and H$_2$O, and we saw that the proton transfers were simultaneous, not sequential (19). Remarkably, the detailed data were almost identical with those observed in the enzyme process, which raised the question of whether the enzyme also used a mechanism proceeding through a five-coordinate phosphorane intermediate, rather than going to the ring-opened product directly. We examined another biomimetic system to explore this question.

We studied the reaction of a tiny fragment of RNA, uridyldine 6, catalyzed by high concentrations of imidazole buffer (20) (Fig. 5). This buffer imitated the imidazole and imidazolium groups of the enzyme. In the enzyme, the effective concentration of these groups is high because of substrate binding right next to them. The products of the buffer-catalyzed reaction were uridine 7 and uridine cyclic phosphate 8, just as in an enzymatic cleavage process. However, we also saw that the buffer catalyzed an isomerization of the phosphate diester 6—originally attached to the 3-oxygen of uridine—to an isomer 9 in which it had migrated to the 2-oxygen.

The chemistry of phosphate reactions, and in particular the nonequivalence of the five groups in a phosphorane, requires that such a migration proceed through a phosphorane intermediate that undergoes pseudo-rotation to permit the migration (21). Our other kinetic studies made it clear that the cyclization process went through the same phosphorane intermediate.

We have proposed that the normal enzymatic hydrolysis process also proceeds through such a phosphorane intermediate, but isomerization does not occur because pseudo-rotation is not permitted in the enzyme (22). Complete agreement does not exist with our proposal (23). In this case, the biomimetic model system was not based only on nature’s enzyme; it pointed to possible aspects of the natural enzyme that had not been considered previously. Thus we had learned something from nature, and we returned the favor by learning something that could give insight into nature itself.

**Water exclusion by enzymes**

Natural enzymes are large molecules, which raised the question of what particular advantages such large catalytic systems could have. This question stimulated Klotz and Suh to study large synthetic polynucleotides as enzyme mimics (24). The compounds, derived by polymerizing aziridine, are available commercially in various sizes and with different degrees of polydispersity. Klotz and Suh used a large polyaziridine with extensive cross-links, and in some cases they added alkyl chains to form a hydrophobic region. In his independent research, Suh pursued this area, examining other polymers as well such as polystyrene (25). Also, Hollfelder et al. examined this system quantitatively and concluded that there was a medium effect in the polynucleotides with added hydrocarbon chains (26).
Generally, their studies involved hydrolytic and fragmentation reactions. We took up the study of such systems as mimics for transaminase enzymes and showed with the polyaziridines that we achieve very large accelerations of the conversion of ketocids to amino acids (27). This result reflected several ways in which these polyamines mimic enzymes.

First of all, transaminations involve many steps in which protons are added and subtracted from the reacting species and the reaction intermediates. In enzymes such as ribonuclease, normally imidazoles are used to perform such proton transfers. Their pKas are close to the operating pH for the enzyme, which makes them both the strongest base and the strongest acid that can exist in free form. A stronger base would be protonated by the medium and a stronger acid would be deprotonated. It can be shown that the catalytic result of increased acid or base strength does not make up for the loss of concentration of the free acid and base species.

The polyaziridines have their nitrogens so close to each other that they titrate over the range of pH 13 to pH 3, so they are half-protonated at pH 8 and have the strongest base and acid groups possible, just as in enzymes that use imidazole with a pK near neutrality. Of course the polyaziridines achieve this not with special basic groups, just with the effect of neighboring charges that make it harder to add more positive charge.

We also saw that the potency of the polyaziridines as enzyme mimics was increased strongly when hydrophobic chains were added to some of the nitrogens (described in Refs. 24–27 and 28–30), which was evident in two ways. First of all, with the hydrophobic core that these chains produced, we achieved strong binding in water solution of ketocids that carried hydrophobic components. The binding was much better, as reflected in a Michaelis K_m for indolepyruvic acid— that produced tryptophan on transamination— than for pyruvic acid that produced alanine, which is not surprising. Perhaps more interestingly, the hydrophobic chains increased the rate constants significantly for reaction of the complex (28), which reflects a medium effect that is invoked often for natural enzymes. Water is to some extent an enemy of rate because acids, bases, and substrates often need to lose their bound waters to react. In nature, this problem is solved by performing the reactions in the hydrophobic interior of the enzymes away from the external water. The hydrophobic chains in our enzyme mimic performed the same function, which excluded water from the reaction site.

The transformations themselves involved reactions of ketocids with a pyridoxamine unit, either covalently attached to the polymer or reversibly bound to the hydrophobic core (29), which converted the ketocids to amino acids, and the pyridoxamine was converted to a pyridoxal unit either covalently attached to the polymer or reversibly dissociated from the polymer. This reaction was modeled directly on the transamination process observed in natural enzymes. However, the second part of a full transamination in nature is the reaction of the pyridoxal with a different amino acid, which runs the transamination backward to form the pyridoxamine again while converting the new amino acid into its corresponding ketocid. We found that such a process was too slow in our biomimetic system and could not compete with the rapid aldol condensation of the ketocids with the pyridoxal.

To solve this problem, we used a mimic of a different enzyme, dialkylglycine decarboxylase (30). In this enzyme, pyridoxal phosphate reacts with an alpha-disubstituted glycine to perform an irreversible decarboxylation (Fig. 6) while converting the pyridoxal species to a pyridoxamine. We imitated this with our model transaminations using pyridoxal species that carried hydrophobic chains, and we were able to achieve as many as 100 catalytic turnovers. Thus, we could imitate one enzyme—the ordinary transaminases—by also imitating another enzyme that solved the turnover problem.

As we read more about disubstituted glycines, we learned that some such species are delivered by carbonaceous chondritic meteorites, and with partial enantioexcesses. This information offered a clue as to how biologic homochirality could have originated on earth (see below).

The mechanism used in the oxidative decarboxylation of alpha disubstituted glycines by an enzyme, which, in mimics, solved the problem of converting pyridoxal species to pyridoxamine species in biomimetic transaminase systems.
Cells as organized chemical systems

A living organism is more than just a collection of enzymes and other important species—it is an organized collection. The components of the cell are compartmentalized into organelles such as the nucleus, and the cell is structured additionally by internal and external membranes. This powerful type of system still waits to be fully mimicked.

Molecular machines

Living systems can convert chemical energy into mechanical motion. The most obvious examples are the motion in muscle contraction and the motion exhibited by flagella in bacteria and in sperm cells. These models from nature have stimulated a few chemists to try to mimic such properties with promising results (31, 32). We do not yet know what practical benefits such mimics will bring, but the creation of molecular machines by imitating those in nature is an exciting adventure. It is part of the fundamental change that chemistry itself is undergoing. Instead of being concerned only with the properties of pure chemical substances, the chemistry of the future will be concerned with the properties of organized multimolecular systems. It is an exciting prospect, which is stimulated by our observation of the wonderful properties of nature’s organized multimolecular system: the living cell.

Energy from sunlight

Nature performs a process, called photosynthesis, that is of great interest to chemists and that has stimulated several attempts to model it. Many ideas about the future sources of energy on earth involve the large amount of energy that sunlight delivers. Fossil fuels—petroleum and coal—are derived from plants that stored chemical energy derived from sunlight and the animals that ate those plants. Although producing artificial photosynthesis is a more practical way to obtain energy than “letting nature do it” with growing plants, however the details of photosynthesis still wait to be fully mimicked.

During photosynthesis, an electron of chlorophyll is pumped to a higher energy orbital by light absorption while a hole is created in the orbital that contained the electron previously. In photosynthesis itself, the excited electron produces the biological reducing agent required for the chemistry of photosynthesis, while also generating phosphate anhydrides such as ATP. Meanwhile, the hole from a second chlorophyll—whose photoreacted electron fills the hole in the electron-deficient first chlorophyll—picks up an electron from the oxidation of water, which generates molecular oxygen. However, biologic photosynthesis requires water and carbon dioxide; water in particular is in demand for many other uses. Thus, it is more appealing for human energy generation from sunlight to use the combination of an excited electron and a hole to produce an electrical current, to use the photovoltaic effect. This process does not consume water, so it can be carried out in deserts without competing with agriculture. Many scientists are now trying to produce practical photovoltaic devices. Perhaps not always consciously modeled in nature, the processes do involve the same step as in photosynthesis—the photoassication of an electron.

The chemical senses

Among biologic organisms, the ability to detect specific chemicals in the environment is universal—to find food, to avoid danger, and to find receptive mates. In higher animals, this ability is lodged to some extent in the sense of taste, but it chiefly occurs in the sense of smell. One of the most baffling aspects of such chemical senses is the ability of animals to detect, selectively, substances that were created newly by chemists. How could the genetic code produce a specific receptor for every possible compound, with compounds numbering in the millions and with more continually being created? The solution to this puzzle by nature, brilliantly discovered by Axel (33) and Buck (34) in their Nobel Prize-winning work, was multiplexing. In principle, a finite group of receptors could react to a compound, but not all with the same strength of reaction. Thus a fingerprint was created, which reflects the different strengths of reaction of each receptor to the chemical. Every chemical, old or new, could elicit a different pattern of such signals and thus be perceived differently.

Eric Arndt (35) has mimicked and adapted this principle for the detection and analysis of chemicals in a nonbiologic application. This principle promises the same ability to detect a great range of chemicals by using a multiplexed group of chemical detectors; but the different strengths of signals this group are transmitted to a computer, not to a brain.

The origin of life on earth

Nature has inspired chemists to mimic not just the details we understand, but also some important details about which we can only speculate. Chemists are making major efforts to understand how life could have begun on earth. Nature has presented us with the puzzle, in the form of biology as it now exists.

The L amino acids in proteins, the D carbohydrates, the nucleic acids, all are found to be homochiral, as observed as a single enantiomer. When the opposite enantiomers are found in biology, they are used for different functions, such as the D amino acids in some bacterial cell walls. Ordinary chemical reactions that create a new chiral center from optically inactive precursors normally produce a racemic mixture unless some chiral catalyst is present to direct the process. How did it get started before chiral catalysts such as enzymes were present?

Various theories exist, but one with considerable support takes note of the fact that some meteorites have brought to earth amino acids that are partially enantiomeric, with 3–15% enantiomeric excess of the L-amino acids (36). They are not normal amino acids, but they have a methyl group on the alpha carbon in place of the usual hydrogen so they cannot racemize. It is believed that they are formed as racemates in interstellar space, but then are deacetylated partially by circularly polarized light emitted by synchrotron processes at neutron stars.

Chemists have been inspired to study how these partially enantiomeric “seeds” could induce the formation of normal biologic molecules—amino acids and carbohydrates—under credible prebiotic conditions (36, 37), and unpublished work). Furthermore, chemists have been inspired to research how low levels of enantiopurity can be amplified into high levels—such
as a 95 to 5 ratio of L-phenylalanine starting with only a 50.5 to 49.5 ratio—under credible prebiotic conditions (38, 39).

Such studies—and others attempting to understand how spontaneous chemistry in the primitive earth could have formed living cells—are not directed indirectly by information about how life originated, but they are inspired by nature as we see it, which is populated as it is by life forms. The work does not aim to show how life originated, which is an historical question. However, it tries to develop the scientific evidence that can support a reasonable hypothesis of how it may have started.

**Artificial life**

A rather important goal of modern chemistry is directly inspired by nature: the hope to produce artificial life, or at least some systems that share many of the attributes of life. The ultimate goal of biomimetic chemistry is to mimic life itself, not in the form that it now has but in alternate forms. Chemists have generalized so many types of substances—novel polymers, novel carbohydrates and amino acids, novel hormones and enzyme inhibitors—that to generalize natural chemistry is one of their most characteristic activities. Can they generalize life itself? Time will tell.

**Other authors**

A good fraction of modern chemistry, which includes medicinal chemistry, derives some of its inspiration from a consideration of nature as a model. It is impossible, in this brief article, to describe more than a fraction of the efforts. For additional work in the spirit of biomimetic chemistry, the following is a list of some authors not yet mentioned whose independent work should be consulted. Even here, this list is certainly not complete: Jacqueline Barton, Steven Benkovic, Albrecht Bickle, Stephen Buchwald, Donald Cram, Peter Dervan, Francois Diederich, Berkessel, Thomas Bruice, Jik Chin, Jean Chmielewski, E. Matta, VC, Grubbs RH. Ruthenium(II) methathesis catalysts bearing an N-fluorophenyl-N-methyl-substituted unsymmetrical N-heterocyclic carbene. Organometallics 2007;26:2469-2472.


References


Further Reading


Physical Chemistry in Biology

Alan Cooper, WestChem Department of Chemistry, Glasgow University, Glasgow, Scotland, United Kingdom

doi: 10.1002/9780470048672.wecb451

Physical chemistry is that area of science that attempts to describe the properties and processes of matter in terms of underlying physical laws. Biology is no exception to these laws. This introductory article briefly sketches the role played by traditional physical chemistry in analyzing and understanding the behavior and control of biologic processes, focusing mostly on the molecular level. Thermodynamics and kinetics place restrictions on the kinds of reactions and processes that can take place as well as on how fast they occur. Electrochemistry, a branch of thermodynamics that deals with charge transfer processes, underpins much of redox biochemistry and membrane phenomena. Spectroscopy gives rise to a wealth of experimental techniques for structural and functional analysis, while quantum chemistry lurks in the background.

Thermochemistry and Thermodynamics

Thermodynamics is the area of science that relates to the interplay of heat and other forms of energy. At the molecular level, this science describes the balance between two generally opposing thermodynamic forces: the natural tendency for mechanical systems to move toward lower energies and the equally natural tendency for thermal Brownian motion to perturb this mechanical order. For open systems at constant pressure, this balance is expressed by the classic Gibbs free energy change (∆G):

\[ \Delta G = \Delta H - T \Delta S \]

Physical chemistry is the area of science that attempts to describe the chemical properties and processes of matter in terms of the underlying physical laws that govern the molecular world. It is a subject of enormous breadth that underpins most areas of science and technology, that lies at the very diffuse interface between physics and chemistry, and that consequently has important ramifications in biology. Throughout history, investigations of physical and biologic phenomena have been closely associated. Scientists, the "natural philosophers" of previous generations, were rightly curious about all aspects of the natural world, with none of the restrictive demarcation issues between "physical" and "biologic" sciences that came later.

Physical chemistry is that area of science that attempts to describe the chemical properties and processes of matter in terms of the underlying physical laws that govern the molecular world. It is a subject of enormous breadth that underpins most areas of science and technology, that lies at the very diffuse interface between physics and chemistry, and that consequently has important ramifications in biology. Throughout history, investigations of physical and biologic phenomena have been closely associated. Scientists, the "natural philosophers" of previous generations, were rightly curious about all aspects of the natural world, with none of the restrictive demarcation issues between "physical" and "biologic" sciences that came later.

Thermodynamics is the area of science that relates to the interplay of heat and other forms of energy. At the molecular level, this science describes the balance between two generally opposing thermodynamic forces: the natural tendency for mechanical systems to move toward lower energies and the equally natural tendency for thermal Brownian motion to perturb this mechanical order. For open systems at constant pressure, this balance is expressed by the classic Gibbs free energy change (∆G):

\[ \Delta G = \Delta H - T \Delta S \]

The objective here is to describe how the traditional field of physical chemistry applies to the study of biology, specifically to provide a broad and relatively nontechnical introduction for nonspecialists. However, a comprehensive description of the role of physical chemistry in biology is a mammoth task that cannot possibly be tackled adequately in a short article such as this. Consequently, I shall simply attempt to give a broad overview from the point of view of a former physicist, now physical chemist, working toward the physical understanding of biomolecular processes. I shall follow the traditional headings, such as might be found in conventional physical chemistry texts (some are listed below in Further Reading, but almost any standard text in any edition will do—like London buses, another one always comes along soon!) and shall note particular relevance to biologic systems, with a few specific examples and associated remarks as we proceed. This approach will be entirely unsatisfactory for most readers who require more detail, but it should act as an aide-memoir stimulating interest in additional details that may be found elsewhere in this volume.

Thermodynamics is the area of science that relates to the interplay of heat and other forms of energy. At the molecular level, this science describes the balance between two generally opposing thermodynamic forces: the natural tendency for mechanical systems to move toward lower energies and the equally natural tendency for thermal Brownian motion to perturb this mechanical order. For open systems at constant pressure, this balance is expressed by the classic Gibbs free energy change (∆G):

\[ \Delta G = \Delta H - T \Delta S \]
where $\Delta H$ is the change in “enthalpy” or heat content that describes changes in internal kinetic and potential energies of all molecules in the system, corrected for any pressure-volume work done on or by the surroundings during the process; $T$ is the absolute temperature (K or eV); and $\Delta S$ is the change in entropy of the system.

It is convenient (although sometimes confusing) to express the equilibrium constant ($K$) of a chemical process in terms of the “standard Gibbs free energy change”, $\Delta G^\circ = RT \ln K$, where $R$ is the universal gas constant. The standard Gibbs free energy change is the hypothetical free energy change for the process in which the reactants and products are imagined to be present in (arbitrarily) “standard states.” In solution, for example, the convention in (most of) chemistry is to adopt a standard-state concentration of 1 M (1 mole per liter) for all reacting species. This convention is clearly a long way from reality for biochemical systems, especially for reactions involving hydrogen ($\text{H}^+$) ions, so an alternative convention is frequently adopted here with $[\text{H}^+] = 10^{-7}$ M to reflect more closely the near-neutral pH of most biologic processes.

Under any conditions, the actual free energy change depends on concentrations of reactants and products:

$$\Delta G = \Delta G^\circ + RT \ln Q$$

where the reaction quotient, $Q$, is a quantity rather like the equilibrium constant, $K$, but involves the actual concentrations pertaining to the system rather than concentrations when it has reached equilibrium.

For example, for a hypothetical reaction $A + B \rightarrow C + D$ with $K = [C][D]/[A][B]$ compared with

$$Q = \frac{[C][D]}{[A][B]}$$

where $[A]$, $[B]$, $[C]$, and $[D]$ are the actual concentrations, compared with what they would be ($[A]^\circ$ $[B]^\circ$ $[C]^\circ$ and $[D]^\circ$) if the system was at thermodynamic equilibrium, in which case $\Delta G = 0$. (Note that in more rigorous treatments, one should replace concentrations with activities. The activity is related to concentration, but corrected for the nonideality that occurs from interactions not accounted for elsewhere in the analysis. This distinction can be significant for macromolecules in crowded environments inside living cells.)

Such basic thermodynamic considerations are essential to understanding biomolecular energetics. The $\Delta G$ places fundamental constraints on the amount of work (mechanical, chemical, electrical, etc.) that may be done by or on a system under ideal conditions. It follows from the above that any chemical reaction or process away from equilibrium can act as a source of free energy. The prime example here is the way that biology makes use of adenosine triphosphate (ATP) and its hydrolysis to ADP, as a ubiquitous free energy currency. ATP is produced during by a wide range of catabolic (energy-producing) reactions such as glycolysis, and its hydrolysis is coupled to free energy transcription in most molecular machines that drive living processes—muscle contraction, flagellated motion, cell division, membrane pumps, nerve signals, and so forth. Yet perhaps surprisingly, and in contrast to much ill-informed speculation about “high energy phosphate bonds,” the hydrolysis of ATP is not particularly energetic: The standard free energy of hydrolysis ($\Delta G^\circ$) of ATP to ADP and inorganic phosphate (about -35 kJ mol$^{-1}$) at physiologic pH is relatively small compared with many chemical reactions, and actual free energy values will, in any case, depend on the relative concentrations of all reactants and products and the extent to which they are away from equilibrium. Rather, it would seem that it is the relative ease with which ATP synthesis and hydrolysis can be coupled to the biomolecular energy transduction machinery that has led to its ubiquity. Despite considerable advances in determining the molecular structures of many enzymes and other proteins involved in this machinery, no consensus view yet exists as to how ATP hydrolysis and work are linked at the molecular level. Thermodynamics is a harsh taskmaster here, and proposed mechanisms need to satisfy these fundamental constraints. As the fictional Homer Simpson once said: “In this house we obey the laws of thermodynamics.”

In statistical thermodynamics, the Gibbs free energy ($G$) is related to the statistical probability (p) of finding the system in a particular state: $G = -k_B T \ln p$, where $k_B$ is the universal Boltzmann constant. Consequently, changes in Gibbs free energy ($\Delta G$) tell us about the relative probabilities that processes will occur in the absence of other interventions: Reactions with a positive $\Delta G$ involve transition to a less likely (smaller p) state and, therefore, are unlikely to occur of their own accord; however, reactions with a negative $\Delta G$ involve transition to a more probable state (larger p) and may proceed. Technically, the latter are said to be thermodynamically “spontaneous” processes—a term that can lead to confusion because many such processes, although thermodynamically feasible or spontaneous, will be limited by kinetic barriers and may not take place on a reasonable timescale.

The Boltzmann statistical definition of entropy, $S = k_B \ln W$, shows how this quantity, often expressed as a measure of “disorder” in a system, is more rigorously defined in terms of the number of ways ($W$) in which a system might adopt a particular energy. It follows that in a molecule system, the higher the energy, the greater the number of ways to partition that energy among the various forms of kinetic (translation, rotation, vibration, etc.) and potential energies. As a consequence, systems with high entropy (larger W) tend to be more likely, overall. For closed systems, such as the entire universe for which the total energy is constant, therefore, the total entropy must always be increasing ($\Delta S > 0$). However, for regions within such a system that can exchange energy with their surroundings, this mandatory increase is not necessarily the case, and it is the Gibbs free energy described above that determines the way things proceed. Obviously, this occurrence is relevant to living systems that, at first sight, seem to violate the rule of increasing W (or S). However, in the wider context, the flux of energy through such systems contributes to the whole, and thermodynamics remains ineluctable. Such concepts of nonequilibrium thermodynamics and dissipative structures have been
a major development in twentieth-century physical chemistry (1, 2).

One consequence of this statistical view of thermodynamics must play a significant role in our understanding of processes that involve biologic macromolecules (3–4–5–6–7). Although normally insignificant for macroscopic objects, thermally induced fluctuations in structural and thermodynamic properties of mesoscopic systems—systems between microscopic and macroscopic, typically relating to objects a few nanometers in size—such as individual proteins or other macromolecules, will be crucial to their function. (Mesoscopic: intermediate between micro- and macroscopic, typically relating to objects a few nanometers in size.) A dramatic illustration of this importance, first highlighted by Max Perutz (8), comes from the structures of the oxygen carrier proteins myoglobin and hemoglobin. Structures for these molecules were among the first to be determined by the emerging techniques of protein crystallography, and they were a major breakthrough in structural biology. Yet these high resolution structures immediately revealed a puzzle. In the (static) X-ray pictures, the oxygen-binding centers are buried deep inside the globular protein structure, with no apparent pathways within the protein structure through which individual oxygens may flow. Incidentally, this discovery also resolves another paradox, because burial of the oxygen-binding iron (heme group) within the fluctuating structure allows binding of O2 to take place with exclusion of water, which would otherwise facilitate irreversible oxidation (“rusting”) of the iron.

At a more basic level, measurements of the energetics of biochemical processes and their couplings to metabolic processes, following the rules of thermodynamics, underpin our understanding of how organisms can function. This understanding operates at all levels, from the “calorific value” of foodstuffs in nutritional biochemistry and food science to the coupling of ATP synthesis/hydrolysis to hydrogen ion gradients (the “proton-motive force” or “chemiosmotic” principle (9)) at one extreme, to entire ecosystems at the other (10). Experimental physical chemistry has played a role here from the very beginning by supplying the calorimetric tools to measure heat and related thermodynamic quantities in biologic systems and organisms. Indeed, one of the first quantitative applications of calorimetry was the use by Lavoisier of his ice calorimeter in the late eighteenth century to measure metabolic heat from a guinea pig, which established the crucial link between respiration and combustion. More recently, advances in instrumentation have led to the widespread use of microcalorimeters for the determination of biomolecular thermodynamics and, much more generally, as a generic analytic technique. It is now possible to measure directly the heat effects associated with noncovalent processes in quite dilute biomolecular systems (11).

Kinetics

Thermodynamics describes what could possibly happen at equilibrium. Kinetics tells us how fast we are getting there. The world around us, including the subset that we call biology, is never at thermodynamic equilibrium. Living organisms succeed because of the very careful way in which the rates of biochemical processes are controlled. Consequently, the study of kinetics has had a major impact on biologic science.

Catalysis and control is crucial here. For a chemical reaction to take place, it is generally necessary that the molecules collide in the correct orientation and with sufficient energy to overcome the activation energy barrier \(E_A\), which leads to reaction products. This reaction is encapsulated in the classic empirical Arrhenius rate law:

\[
R = A \cdot \exp(-E_A/(RT))
\]

where \(R\) is the gas constant and where the pre-exponential factor, \(A\), can be related to the apparent collision frequency and the exponential term reflects the Boltzmann probability that the colliding species will have sufficient energy.

A more rigorous description is given by transition state theory (12, 13):

\[
R = (k_B T/h) \cdot \exp(-\Delta G^*/RT)
\]

where \(h\) is the Planck constant and \(\Delta G^*\) is the (notional) Gibbs free energy of activation to the transition state. This equation strictly derives from a quantum statistical mechanical treatment of two-body collisions, so its application to more complex systems (especially those that involve biologic macromolecules in water) needs to be interpreted with due caution.

It follows from this equation that the rates of chemical reactions can, in general, be affected by manipulation of either collision factors or activation parameters or both. And, of course, reaction rates will be very sensitive to temperature change—a key consideration in the evolution of complex pathways of interdependent chemical processes that are necessary for maintenance of biologic organisms.

A biologic reaction rates can be affected by molecular diffusion processes that dictate the rates at which collisional encounter complexes occur before reaction. This affect usually shows up in the way reaction rates depend on the physical form of the reactants (gas, liquid, solid, solution, etc.), particularly on concentrations for reactants in gas or liquid phases. A disortion of reactants onto surfaces can enhance the effective concentra- tions of reactive species and/or reduce the dimensionality of the diffusion process. Classic work by Eigen and Richter (14) showed how restricting diffusion to one or two dimensions can dramatically increase potential reaction rates, and this principle has been applied to the kinetics of protein translocation along DNA chains, for example. See References 15 and 16 for more information.

Enzyme catalysis completely exploits all these aspects of the physical chemistry of reaction kinetics. Despite the great enhancement in reaction rates and specificities that can be achieved by biologic catalysts, it is now generally accepted that no new underlying physical principles are involved, just that enzymes are much better at using and optimizing the various factors required for any chemical reaction—particularly...
Spectroscopy and Photochemistry

The interaction of electromagnetic waves (photons) with matter is crucial to many biologic processes. This interaction is also the basis for a wide range of spectroscopic techniques in physical chemistry that probe the structure and function of molecules, many with applications and implications for biology. Absorption and emission of electromagnetic radiation invariably involves changes in electric or magnetic dipoles, such as relative displacements of electrical charges in atoms or molecules, or reorientation of nuclear magnetic moments, and the frequencies (energies) with which these transitions take place dictates the region of the electromagnetic spectrum involved and the sorts of processes that may occur.

Electromagnetic radiation in the ultraviolet (approximately 180–340 nm) or visible (approximately 340–800 nm) region generally interacts with matter through excitation of electronic transitions in atoms or molecules. This interaction gives rise to the characteristic colors of molecules and materials and is the basis for a multitude of analytical techniques in biochemistry and elsewhere. Life, itself, is dependent on the electronic transitions that exploit fundamental effects in electronic energy transfer and transduction (20). Changes in electronic absorbance properties give rise to color changes that underpin many analytical techniques in biochemistry and medicine, and differences in absorbance of left- or right-circular polarized light by chiral structures that exploit fundamental effects in electronic energy transfer and transduction (20). Changes in electronic absorbance properties give rise to color changes that underpin many analytical techniques in biochemistry and medicine, and differences in absorbance of left- or right-circular polarized light by chiral structures that exploit fundamental effects in electronic energy transfer and transduction (20).

The electronic excited state is inherently unstable and can decay back to the ground state in various ways; some of which involve (re)emission of a photon, which leads to luminescence phenomena (fluorescence, phosphorescence, and chemiluminescence) (22). Some biologic molecules are naturally fluorescent, and phosphorescence is a common property of many marine and other organisms. Fluorescence is photon emission caused by an electronic transition to ground state from an excited singlet state and is usually quite rapid. Phosphorescence is a much longer-lived process that involves formally forbidden transitions from electronic triplet states of a molecule. Fluorescence measurement techniques can be extremely sensitive, and the use of fluorescent probes or dyes is now widespread in biomolecular analysis. For example, the large increase in fluorescence of a dye molecule (e.g., ethidium bromide) when bound to double-helical DNA is widely used in molecular biology to detect and locate DNA fragments. Chemiluminescence occurs when the electronically excited state develops as a result of chemical reaction. Chemiluminescence is the basis for many biologic light shows, including the flashing lights of firefly tails (using enzyme luciferase-catalyzed hydrolysis of ATP) and many marine organisms.

Vibrational motions of chemical bonds that involve changes in electric dipole moment or polarizability can be detected by IR absorbance or Raman scattering techniques. The strong infrared (IR) absorbance by water makes conventional IR spectroscopy less generally useful in biology, although this disadvantage is overcome by Raman spectroscopy, which relies on the inelastic scattering of high intensity visible (laser) light. Resonance Raman spectroscopy has been particularly useful in picking out the vibrational spectra of biologic chromophores in otherwise complex mixtures, and the differential scattering of circularly polarized light (Raman optical activity) has been developed as a high resolution probe of chiral features, particularly in biologic macromolecules (23).

Phenomena associated with the reorientation of nuclear or molecular magnetic dipoles in applied magnetic fields have led to dramatic advances in molecular spectroscopy and imaging techniques. Nuclear magnetic resonance spectroscopy (NMR) is now routinely applicable to the study of the structure and dynamics of biomolecules, large and small (24), and NMR imaging techniques are used for whole-body diagnostics and related investigations. Electron paramagnetic resonance (EPR) techniques that rely on the reorientation of electronic spins are less generally applicable because of the relative scarcity of appropriate paramagnetic species in biology. However, specific “spin probes” are used in some instances, for example, in the investigation of biologically significant free radical chemistry, and new probes and improved technologies are being developed.

Most spectroscopic processes mentioned so far involve relatively low energy transitions that do not (usually) affect the covalent structure of the sample. However, absorption of higher energy photons (UV, and others) can lead to higher energy-excited electronic states in which chemical transformations can occur. Such photochemistry can be harmful to biologic organisms, particularly at the DNA level (melanomas, etc.), but can also form the basis for useful techniques and therapies. Photochemistry is also, of course, central to biologic processes such as photosynthesis (20) and visual transduction (25). And the speed with which photoprocesses can be initiated is exploited in many biophysical techniques for studying structure and dynamics, for example, flash photolysis and fluorescence recovery after photobleach (FRAP).

Mass spectrometry is one physical technique that does not (at least directly) involve electromagnetic radiation. However, some sample desorption and ionization processes do use high intensity pulses of laser light in techniques such as MALDI (Matrix-Assisted Laser Desorption Ionization) that have proved very useful in mass analysis of proteins and other biologic macromolecules. High resolution mass spectrometry derives from atomic/molecular beam studies in which the trajectories of ionized particles in a vacuum can be manipulated by static
Quantum Theory and Bonding

Quantum mechanics is probably the most successful theory of the twentieth century, if not ever. It explains the structure and spectroscopic properties of atoms and molecules with remarkable precision and is the foundation for our understanding of chemical bonding. It is therefore surprising, perhaps, that it has not yet had any great impact on our understanding of biologic processes. This occurrence is mainly a matter of scale. Quantum theory is remarkably good at describing the electronic structures of atoms and molecules. And, to the extent that biologic (macromolecules are held together by the same covalent bonds as any other kinds of molecules, the quantum mechanical nature of chemical bonding is inherent to their structure and properties. However, knowing the wave function for the human genome, even if we could calculate it, would be of little use to a molecular biologist. Solving the Schrödinger equation for a polypeptide would not likely tell us how it folds or what its function might be; these problems, for the time being at least, do not need the degree of precision at the atomic and electronic level afforded by quantum theory.

This principle can be illustrated by the wave–particle duality aspects of quantum theory. The de Broglie wavelength (\(\hbar/mv\)) for a typical macromolecule (10 kDa) traveling at thermal velocities (>50 m s\(^{-1}\)) is around 0.01 Å or less. Consequently, quantum effects deriving from the wave-like behavior of matter and the Heisenberg Uncertainty Principle are unlikely to be of significance at the level of resolution (typically 1 Å) currently available for biologic molecules, and classic mechanics is adequate in most cases, which is not to say that quantum effects are entirely absent, however. Quantum mechanical tunneling has been observed in some instances and may be of importance in some enzyme-catalyzed reactions. Tunneling or barrier penetration is a uniquely quantum phenomenon whereby the wave-like nature of particles at the atomic and subatomic level allows them to pass through, rather than over, energy barriers. In some enzyme reactions, this phenomenon is manifested by the deviations from classic Arrhenius kinetic behavior at very low temperatures. Classical, at absolute zero (0 K) all molecular motion stops and chemical reactions should cease, because the molecules no longer have sufficient thermal energy to overcome activation barriers. However, for some reactions, especially those that involve hydrogen transfer, for example, finite reaction rates have been observed at liquid helium temperatures (27, 28). Room temperature tunneling effects are also significant in electron transfer processes within or between proteins (29, 30). The possible importance of significant quantum effects in wave-like energy transfer processes in the early stages of photosynthesis has recently been reported (31).

More generally, electronic and vibrational spectroscopic properties of biologic molecules will, of course, be subject to the underlying rules of quantum mechanics. However, except in special circumstances, the spectra of biologic molecules in solution rarely show the discrete quantized energy-level structure seen in simpler atomic or small-molecule systems in vacuum. This rare display is because the significant conformational flexibility and dynamic solvent environment gives rise to heterogeneous and homogeneous spectral broadening effects that mask the underlying energy-level structures.

At the theoretical level, full quantum mechanical calculations on biologic macromolecules are not computationally feasible, nor would they be particularly helpful in understanding macromolecular properties without proper inclusion of the solvent water or other biologic matrix on which these properties so intimately depend. However, ab initio quantum mechanical calculations on smaller systems that represent crucial steps in an enzymic reaction, for example, can be helpful in understanding specific processes within macromolecules or in estimating intermolecular forces and stereochemical effects in molecular mechanics simulations that are not experimentally accessible.

Symmetry and Group Theory

Group theory is a very powerful mathematical technique for analyzing molecular structures in terms of their symmetry properties. These properties can be characterized in terms of the various symmetry operations, symmetry elements, and space groups or point groups into which they fall. A symmetry operation is any action (translation, rotation, reflection, etc.) that leaves the object looking the same. A symmetry element is the point, line, or plane about which the symmetry operation is performed. A point group is the list of all possible symmetry operations that leave at least one point in the object unchanged (i.e., without spatial translation). In contrast, a space group is the set of all operations that move a molecule (or object) to another position in space. Symmetry considerations can be very useful in determining, for example, spectral characteristics of small molecules or functional groups; for instance, the different selection rules regarding IR absorbance or Raman scattering bands in vibrational spectroscopy of small molecules are equally applicable to such group vibrations in larger systems. But most biologic macromolecules and larger complexes lack sufficient symmetry for the full rigor of group theory to be applicable in any generally useful way. Two significant exceptions exist. The classic work of Caspar and Klug on the structures of icosahedral viruses was based on symmetry considerations, many of which derived from the architectural symmetries in the building designs of Buckminster-Fuller (32). And space groups are, of course, central to the analysis of protein crystals, where determination of the space group is the first essential step in crystallographic resolution of protein (and other structures) by X-ray diffraction techniques.

Electrochemistry

Processes that involve the transfer of charge can be manipulated by electrical potential or, conversely, can generate voltages that...
can drive other reactions or can be used for analytical purposes. Electron transfer between molecules or chemical groups (i.e., oxidation/reduction or "redox" processes) can be configured as electrochemical cells for numerous purposes, and the thermodynamics of such systems is key to understanding biologic redox processes such as in various intermediate stages of photosynthesis and respiration. Ion transport across membranes or against chemical potential gradients is involved in all aspects of electrolyte balance in biologic systems, both at the microscopic level of individual cells or organelles and at the macroscopic level of whole organisms. It is the basis for the transmission of nerve signals and for the use of membrane potentials for ATP synthesis (9).

The key expression here is the Nernst equation that, under ideal conditions, relates the electrical potential (E) of a system participating species:

\[ E = E^\circ - \frac{RT}{nF} \ln \frac{Q}{Q^\circ} \]

where \( E^\circ \) is the standard electrochemical potential of the process, \( n \) is the number of moles of charge transferred per mole of reaction, \( F \) is the Faraday (96,500 coulomb/mole), and \( Q \) is the "reaction quotient" that involves the relative concentrations (activities) of reactants and products.

This process is the basis for several analytical devices, electrochemical cells, pH- and ion-specific electrodes, and other sensors in which concentrations and electrochemical properties can be determined in terms of the voltage (electrochemical potential) developed across membranes or other partitions separating the two halves of a redox reaction. Various microelectrode technologies employ this approach to study electrochemical potential gradients in single cells, in tissues, and across biologic membranes. Chemical potential gradients across membranes can also drive chemical reactions. This approach is the thermodynamic rationale for chemiosmotic and protonmotive force effects in which \( H^+ \) or other ion gradients across membranes can provide the driving free energy for synthesis of ATP for example (9).

The molecular machinery that involves membrane-associated proton pumps and ATP-synthesizing enzymes must still be worked out in detail, but the basic principles of electrochemistry must apply.

The mobility of ions in an electric field, especially in water, formed the basis of early work in the physical chemistry of solutions. It has now blossomed into the wide range of multi-purpose electrochemistry techniques that are the mainstay of most biochemistry and molecular biology laboratories today. Protein gel electrophoresis, in which protein mixtures may be separated on the basis of their size and charge, is a standard analytical and quality assurance methodology. The same is true for DNA and RNA biochemistry. Sequencing the human genome, for example, would not have been possible (or at least greatly hindered) without the development of the high-resolution electrophoresis methods capable of separating strands of DNA that differ in length by just one nucleotide.

### Experimental Techniques

Perhaps the greatest contribution of physical chemistry to biology lies not in the theoretical fundamentals, although important, but in the experimental techniques that have developed from physical chemistry. Textbooks are written with the benefit of hindsight, and can give the impression that theory comes first, with experiments playing a subsidiary supporting role. Reality is somewhat different. Most scientific development comes from curiosity-driven observation and experimentation, and experimental techniques developed to study the physics and chemistry of matter are applicable equally to biologic systems. Indeed, the need to study biomolecular processes has acted as a spur to the development and applications of physico-chemical methods, which benefits both sides. Here is a list of just some techniques derived from physical chemistry that have applications in biology:

1. In no particular order: UV/visible spectroscopy; fluorescence spectroscopy; circular dichroism; Raman spectroscopy; rapid reaction kinetics; stopped-flow and flash photolysis methods; isothermal and scanning (micro)calorimetry; specific electrode technologies; electrophoresis; chromatography; light scattering and diffusion methods; viscosity and rheology; analytical ultracentrifugation; NMR spectroscopy and imaging; X-ray and neutron diffraction; Langmuir–Blodgett films and other surface technologies; mass spectrometry; and EPR spectroscopy . . .

### And the Rest . . .

A short introductory article such as this cannot possibly cover completely such a wide-ranging subject, and much has been left out. Flick through any physical chemistry textbook and you will find many other topics that we have missed, all of which possess one or more important roles in biology. Here are just a few (with a hint of some biologic significance in parentheses):

- Properties of gases (physiology of oxygen uptake); acid-base equilibrium, buffering in aqueous solution (pH control in cells and organisms); colloid, detergents, and micelles (biologic membranes); surface chemistry (bioadhesion and biocatalysis); polymer structure and dynamics (polypeptides, polynucleotides, polysaccharides); crystallography and diffraction techniques (structural biology); noncovalent interactions (hydrogen bonding); and so on.

The list could continue, but it illustrates the key role that physical chemistry has played, and continues to play, in our understanding of biology.

### References

Normal regulation of the cell cycle ensures the passage of genetic material without mutations and aberrations. Proper completion of each phase is critical to the initiation of the following phase, and the pathways that cell division occur in an ordered, sequential, and irreversible procession. The two major cell-cycle events that are regulated tightly are DNA replication and cell division. Progression through each phase transition is regulated by extracellular signaling, transcription factors, cyclin-dependent kinases (CDKs), and checkpoints, which prevent uncontrolled cell division.

Cyclin/CDK complexes are the primary factors responsible for the timely order of cell-cycle progression, which include entry into S phase, initiation of DNA replication, and mitotic entry. Each phase of the cell cycle and the different cyclin/CDK complexes, as well as other important factors that regulate cell-cycle progression and checkpoints, will be discussed.

The cell cycle is the sequence of events by which growing cells duplicate and divide into two daughter cells. In mammalian cells and other eukaryotes, cell division represents a process of highly ordered and tightly regulated molecular events. The cell cycle is composed of five phases in mammals, including G0, G1, S, G2, and M phases. Replication of DNA occurs during S phase and division occurs during M phase. During the two gap phases, G1 and G2, cells produce RNA and proteins required for the subsequent S and M phases, respectively. Cells in a resting, quiescent state are in G0 phase. Stimulation by external growth factors or mitogens triggers quiescent cells to reenter the cell cycle in G1 by activating numerous signaling cascades, and it leads to the sequential activation of cyclin-dependent kinases (CDKs). Activation of CDKs requires interaction with a cyclin partner, T-loop phosphorylation at T160 (CDK2) or T161 (CDK1) catalyzed by CDK activating kinase (CAK), and dephosphorylation at T14 and Y15 by CDC25 dual phosphatases. The inhibitory phosphorylations at T14 and Y15 are catalyzed by the serine/threonine kinase Wee1 and threonine/tyrosine kinase Myt1, and these cause misalignment of the glycine-rich loop and the ATP phosphate moiety. CDKs phosphorylate multiple substrates. The proper regulation of CDKs is necessary for orderly cell-cycle phase transitions. A general representation of the key players and events during the cell cycle can be observed in Fig. 1.

Numerous checkpoints also exist to ensure normal cell-cycle progression and transmission of an unaltered genome. These checkpoints are conserved signaling pathways that monitor cell growth conditions, cell-cycle progression, and structural and functional DNA defects; they are critical for cell survival or death. Checkpoint responses induce and sustain a delay in cell-cycle progression, and activate machinery to respond to changes in cell growth conditions, repair DNA, and stall replication. When cellular damage cannot be repaired, these checkpoints can induce apoptosis. The mammalian checkpoints include the quiescent checkpoint, G1/S checkpoint, replicative checkpoint, G2 checkpoint, mitotic checkpoint, and the DNA damage checkpoints. Improper checkpoint control promotes tumorigenesis through increased mutation rates, aneuploidy, and chromosome instability. The following sections will give an overview of the regulation of the various phases of the mammalian cell cycle, activation of specific checkpoints, and the molecules involved in the mechanisms that regulate these processes.
Cell Cycle, Regulation of

Figure 1  Regulation of the mammalian cell cycle by cyclin/CDKs. Activation of growth factor receptors in G0 leads to activation of many signaling cascades that lead to the expression of cyclin D. Progression into S phase is mediated by Rb and E2Fs that lead to the initiation and progression of DNA synthesis through cyclin E/CDK2 activity. On completion of DNA replication, cyclin B/CDK1 activity promotes phosphorylation of substrates required for entry into mitosis and eventual cytokinesis, which produces two identical daughter cells.

From Quiescence to the Point of No Return

G0-G1 transition

After cell division, the daughter cell enters into G0 phase where it becomes ready to divide again before entering into G1. In most cases, the newly formed cell increases in size and mass for division to occur again, by enhancing ribosome biosynthesis (1). This task is accomplished by phosphorylation of the S6 ribosomal subunit by S6 kinase (2). This kinase is regulated by members of the PI3K family, including TOR, PDK1, and PI3K, which are activated by insulin receptor signaling (3, 4). These family members phosphorylate the translational inhibitor 4E-BP1, leading to dissociation of the initiation factor eIF4E, which promotes cyclin D and Myc translation (5). In the absence of growth factors, these kinases are inactive and cannot signal progression from quiescence to G1. Acetylation and phosphorylation of the tumor suppressor p53 also seem to be involved in maintaining cellular quiescence (6, 7).

G1 phase

In the presence of growth factors during the G0 and G1 phases, ras and mitogen-activated protein kinase (MAPK) cascades are activated and subsequently regulate cell cycle progression (8). MAPK regulates cyclin D expression directly by controlling the activation protein-1 and ETS transcription factors, which transactivate the cyclin D promoter (9, 10). Consequently, the MAPK cascade activates cyclin D-dependent kinases (CDK4 and CDK6) and regulates cell proliferation. Additionally, the MAPK cascade regulates directly the synthesis of the CIP/KIP family of CDK inhibitors (CKIs), specifically p21(CIP) and p27(KIP), which regulate CDK activity negatively and influence cyclin D/CDK4/6 complex formation in G1 (11, 12).
The growth factor-dependent synthesis of D-type cyclins occurs during the G1/S transition and peak in concentration in late G1 phase (13). These proteins have a very short half-life and are degraded rapidly after removal of mitogenic stimulation. The INK family of CKIs primarily inhibits cyclin D/CDK4/6 complexes. Only when the concentration of cyclin D exceeds that of the INK proteins can these cyclin D/CDK4/6 complexes overcome their inhibition (14, 15).

In early to mid G1 phase, active cyclin D/CDK4/6 complexes phosphorylate the three Rb pocket proteins (pRb, p130, and p107), which results in their partial repression (16–18). E2F proteins (E2F1-6) and cyclin E/CDK2 form heterodimers with a related family of DP proteins (DP1-3), and can act as both activators and repressors of transcriptional activity. In G0, and early G1, Rb is in an active, hypophosphorylated form. Active Rb represses the activity of the E2F transcription factor family by binding directly to the transactivation domain of E2F proteins and recruiting histone deacetylases, methyltransferases, and chromatin remodeling complexes to E2F-regulated promoters (19, 20). This activity results in the modification of histones, compaction of chromatin structure, and prevention of promoter access by transcription machinery (20). Phosphorylation of Rb by cyclin D/CDK4/6 complexes during G1 releases histone deacetylase, which alleviates transcriptional repression partially (13, 19, 20). As a result, the E2F/DP transcription factors activate the transcription of cyclin E and many genes responsible for the G1/S transition and DNA synthesis including CDK2, cyclin A, cyclin E, RPA1, MAT1, PCNA, DHFR, c-Myc, DNA polymerase-ε, p20(MMT), and CDC25A (21).

G1/S transition

Cyclin E expression in mid to late G1 results in the formation of cyclin E/CDK2 complexes, which are required for S-phase entry and the initiation of DNA replication. Cyclin E/CDK2 also phosphorylates Rb, except on different residues than those catalyzed by cyclin D/CDK4/6 complexes (22). Cyclin E/CDK2 phosphorylation of Rb promotes the dissociation of E2F transcription factors from Rb, which results in complete relief of transcriptional repression (23). Thus, Rb inactivation occurs through the sequential phosphorylation by CDK4/6 and CDK2. Additional E2F and cyclin E/CDK2 activity increases through a positive feedback mechanism because cyclin E is one of the many genes activated by E2F (24). Cyclin E/CDK2 activity enhances this positive feedback even more by promoting the degradation of its own inhibitor, p27kip1. These complexes have been shown to phosphorylate p27kip1 at Thr187, which promotes its association with the Skp-Cullin-F-boxp1 (SCFp1) complex to target p27kip1 for ubiquitination and proteasomal degradation (25). Cyclin D/CDK4/6 complexes have been hypothesized to sequester the bound CDK inhibitor p27kip1 away from cyclin E/CDK2 complexes to facilitate their activation (26). However, recently p27kip1 was shown to be phosphorylated by Src-family tyrosine kinases at Y88, which reduces its steady-state binding to cyclin E/CDK2. This action facilitates p27kip1 phosphorylation at Thr187 by cyclin E/CDK2 to promote its degradation (27, 28). Thus, rather than cyclin D/CDK4/6 sequestering of p27kip1, these tyrosine kinases may be responsible for activation of p27kip1-bound cyclin E/CDK2 complexes at the G1/S transition.

The c-myc proto-oncogene encodes another transcription factor involved in many processes, which include E2F regulation (29). Its expression is induced by mitogenic stimulation, promotes S-phase entry in quiescent cells, and increases total cell mass. Myc activates the transcription of cyclin E, CDC25A, and several other genes (30). The Myc-induced proliferation mechanism activates cyclin E/CDK2 activity directly through increased cyclin E levels and CDC25A activity, which removes Thr14 and Tyr15 inhibitory CDK2 phosphorylation catalyzed by Wee1/Myt1 (31). Additionally, this activity is enhanced indirectly through Myc by mediating the sequestration of p27kip1 from cyclin E/CDK2 into cyclin D/CDK4/6 complexes, which in turn promotes the cyclin E/CDK2 catalyzed phosphorylation and degradation of p27kip1 (32). Cul-1, a component of the SCFp1 complex, was shown to be a transcriptional target of Myc, which may explain the link between p27kip1 degradation and Myc activation (33).

Cyclin E/CDK2 also phosphorylates p20(MMT), which is a protein involved in the regulation of histone gene expression. This phosphorylation is a major event that occurs as cells begin to enter S phase (34). The phosphorylation of p20(MMT) by cyclin E/CDK2 is required for histone gene expression activation at the onset of S phase (35). Once cells have passed through the restriction point, they are committed to initiate DNA synthesis and complete mitosis. Cell-cycle progression continues independently of the presence of growth factor stimulation after passage through the restriction point.

Regulation of DNA Synthesis and Mitotic Entry

S phase

At the G1/S transition, the cell enters S phase in which DNA synthesis occurs and each chromosome duplicates into two sister chromatids. During S-phase entry, the initiation of replication occurs at sites on chromosomes termed origins of replication. Replication origins are found in two states within cells: a pre-replicative complex (pre-RC) that is present in G1 before DNA replication initiation, and the post-replicative complex (post-RC) that exists from the onset of S phase until the end of M phase (36). At the onset of S phase, an increase in cyclin A expression and cyclin A/CDK2 activity occurs (37, 38), and the protein kinase GSK-3β phosphorylates cyclin D and signals its relocalization to the cytoplasm where it is degraded by the proteasome (39, 40). Cyclin A/CDK2 activity controls each round of DNA replication, and this dictates the state of the replicative complexes. Low CDK activity permits the assembly of the pre-RC to form a licensed origin at the end of M phase, whereas the increase in CDK activity during the G1/S transition triggers initiation of DNA replication and converts origins to the post-RC form (41). Replication of the pre-RC is prevented...
Cell Cycle, Regulation of

by high CDK activity, which acts to inhibit re-replication events that would result in numerous copies of chromosomes.

The initiation of DNA replication requires both the assembly of the pre-RC complex at origins of replication and the activation of these complexes by CDKs and other kinases to initiate DNA synthesis (42–44). Numerous proteins are required for pre-RC formation of the DNA replication initiation complex, which include the Origin Recognition Complex (ORC), cdc6/18, cdt1, the GINS complex, and mini chromosome maintenance (MCm) proteins (43). ORC proteins (ORC-1-6) bind directly to replication origins as a hexamer and facilitate the loading of other components of the pre-RC (45, 46). The cdc6/18 and cdt1 proteins play a central role in coordinating chromatin licensing. They bind directly to the ORC complex independently of each other (47). Here, they facilitate cooperatively the loading of the Mcm proteins (MCm2-7), which form a hexameric ring-complex that possesses ATP-dependent helicase activity (48, 49). Cyclin E/CDK2 is recruited to replication origins through its interaction with cdc6, and this event regulates cdt1, cdc6/18, and MCm loading, which makes chromatin replication competent. A fiber binding of the Mcm proteins, the affinity of both cdc6/18 and cdt1 for the ORC is reduced, and they dissociate (48, 49). Then, cyclin A/CDK2 phosphorylates cdc6 to promote its export from the nucleus and cdt1 to target its ubiquitination (48, 49). Then, cyclin A/CDK2 phosphorylates cdc6 to promote its export from the nucleus and cdt1 to target its ubiquitination by the SCF^CMCc complex (50, 51). In this way, after initiation and release of these factors from the ORC, cyclin A/CDK2 activity acts to prevent re-replication by inhibiting reformation of the pre-RC. However, cyclin E/CDK2 activity acts primarily to promote the initiation of DNA synthesis (52).

Dbf4-dependent kinase (DDK) contains the kinase subunit cdc7, and it is required for DNA replication initiation (53). DDK targets Mcms for phosphorylation, thereby increasing the affinity of these proteins for cdc6, which is a factor required for the initiation and completion of DNA replication (54). The GINS complex, which consists of the four subunits Sld5, Psf1, Psf2, and Psf3, is required for the initiation and progression of eukaryotic DNA replication (55). The GINS complex associates with Cdc45 and the Mcm proteins to activate their helicase activity. As a result of GINS and cdc45 binding to the Mcm complex, the DNA is unwound, which results in single stranded DNA (ssDNA) (49, 57). Replication protein A (RPA) is recruited to single stranded DNA, and it is required for the subsequent binding and activation of DNA polymerase-α (58–60). The GINS complex also interacts with, and stimulates the polymerase activity of the DNA polymerase-α-prime complex (61).

G2 phase

After completion of DNA duplication, the cell enters the second restriction point of the cell cycle, referred to as the G2 phase. Similar to what happens during G1, in this second gap phase the cell halts to synthesize factors required for initiation and completion of mitosis and to check for any aberrations that result from DNA synthesis (62, 63).

Cyclin B/CDK1 is the primary regulator of the G2/M transition, and its activity is required for entry into mitosis. It was termed the mitosis-promoting factor (MPF) because it was originally shown to be essential for Xenopus oocytes maturation after hormonal stimulation, and it was found subsequently to be equivalent to a mitosis-promoting activity (64). CDK1 activity is regulated primarily by localization of cyclin B, CDC25C activity, and p21^CIP levels, which are controlled by checkpoint machinery (65). Cyclin B/CDK1 complexes remain inactive until their activity is required for mitosis entry in late G2. Toward the end of S phase, cyclin B expression is increased. However, during the onset of G2, cyclin B is retained in the cytoplasm by its cytoplasmic retention signal (CRS), and the CK1 p2A^T inhibits CAP-mediated activation of cyclin/CDKs (66). Additionally, Wee1 and Myt1 phosphorylate T14 and Y15 on cyclin B/CDK1 in the cytoplasm to keep these complexes inactive even when CDK1 is phosphorylated by CAK (67). The transcription factor F53 also mediates the inhibition of cyclin B/CDK1 activity by promoting p21 expression, and it down-regulates expression of CDK1 (63, 68). Furthermore, cyclin A/CDK2 phosphorylates and inactivates members of the E2F transcription family in G2 to suppress cell growth during this gap (69–71).

G2/M transition

During the G2/M transition, the localization of cyclin B changes dramatically and regulates CDK1 activity (72). The CRS is phosphorylated by MAPK and polo-like kinase 1 (Plk1), which promotes its nuclear translocation (73, 74). Concomitant with nuclear import, cyclin B is phosphorylated to a greater extent to prevent association with CEN1, which promotes its nuclear retention (75, 76). This relocalization occurs at the onset of mitosis toward the end of the G2/M transition when the cell is ready to begin the mitotic process (77). Activation of cyclin B/CDK1 in late G2 is achieved by preventing the access of cytoplasmic Wee1/M1 and to the complex and by promoting shuttling of the CDC25 phosphatases to the nucleus, where they dephosphorylate and activate CDK1 (78–80). Cyclin B/CDK1 complexes phosphorylate CDC25A to promote its stability and CDC25C to promote its activity (81, 82). Both CDC25A and CDC25C activate CDK1 even more, which results in a positive feedback loop that sustains cyclin B/CDK1 activity in the nucleus to signal mitotic entry (82, 83). ERK-MAP kinases also regulate cyclin B/CDK1 activity by phosphorylating CDC25C at T48 (84). ERK2/2 activation of CDC25C leads to removal of inhibitory phosphorylations of cyclin B/CDK1 complexes and is required for efficient mitotic induction. Thus, MAPKs are involved in the positive feedback loop that leads to cyclin B/CDK1 activation.

The increase in nuclear cyclin B/CDK1 activity promotes phosphorylation of nuclear substrates that are necessary for mitosis, such as nuclear envelope breakdown, spindle formation, chromatin condensation, and restructuring of the Golgi and endoplasmic reticulum (55, 85, 86). Numerous cyclin B/CDK1 substrates have been defined, which include nuclear lamins, nucleolar proteins, centrosomal proteins, components of the nuclear pore complex, and microtubule-associated proteins (87–89). Cyclin B/CDK1 complexes also phosphorylate Mcm4 to block replication of DNA, the TFIIH subunit of RNA polymerase II to inhibit transcription, and the ribosomal 56 protein kinase to prevent translation during mitosis (90–92).
Regulation of Cell Division

The centrosome

Normally, the centrosome is composed of two centrioles and the pericentriolar material. It functions not only as a microtubule nucleation center, but also as an integrated regulator of cell-cycle checkpoints. Recent data indicates it is required for cell-cycle progression (93). The centrosome duplication process begins in late G2 and is regulated primarily by CDK2 activity (94). Cyclin A/E/CDK2 phosphorylates the Mps1p kinase and nucleoplasmin, which are two centrosome associated proteins. CDK2 activity is required for Mps1p stability and Mps1p-dependent centrosome duplication (95). Cyclin E/CDK2 phosphorylates nucleoplasmin at T199, releasing it from unduplicated centrosomes, which are a requirement for centrosome duplication (96). Completion of centrosome duplication and initiation of their separation occur in G2 and are dependent on cyclinA/E/CDK2 activity. These processes are necessary for proper spindle formation and for balanced chromosome separation during mitosis.

The Aurora kinase family members play a role in centrosome function, spindle assembly, and chromosome alignment, and they are essential for mitosis. Specifically, Aurora-A activity is maximal during G2/M; it regulates mitotic spindle assembly, centrosome separation, and it facilitates the G2/M transition by phosphorylating CDC25B at the centrosome, which is an important event for cyclin B localization to the nucleus (97). Aurora-B activity is maximal from metaphase to the end of mitosis and regulates chromatin protein modification, chromatid separation, and cytokinesis (98). During mitosis, a complex process of degradation and phosphorylation regulates Aurora kinase activity to ensure proper mitotic advancement. Aurora-A is activated mainly by autophosphorylation (99), Ajuba (100), TPX2 (101), and HEF1 (102), whereas INCENP is thought to inhibit Aurora-A/B (103). Both Aurora-A and B are degraded rapidly at the end of mitosis.

M phase

The mitotic phase is divided into five phases, which include prophase, prometaphase, metaphase, anaphase, and telophase. During prophase, nucleoli disappear, chromatin condensation takes place, and the mitotic spindle is formed at centrosomes that contain centrioles. In prometaphase, fragmentation of the nuclear envelope occurs and mitotic spindles extend from the poles toward the center of the cell. At metaphase, centrioles pair at opposite poles, and the chromosomes align in the cell center along the metaphase plate. Then, microtubules bind to the kinetochores located at the centromeres of each chromatid of the chromosomes. The transition from metaphase to anaphase is triggered by MPF inactivation through the degradation of cyclin B by the E3 ubiquitin ligase anaphase-promoting complex (APC/C) (104). APC/C is required for activation of the ubiquitin ligase activity of APC/C, which promotes degradation of securin. Subsequently, a release mechanism activates the protease known as separase, which cleaves cohesion and promotes sister chromatid separation and anaphase entry (105, 106). This mechanism induces the separation of chromatids in anaphase as microtubules from each pole pull them apart through their kinetochore. Because of cyclin-B/CDK1 inactivation in late anaphase, the major ubiquitin ligase activity is switched from APC/C-Cdc20 to APC/C-Cdh1. The latter continues to regulate many proteins whose degradation is required for cell-cycle progression, including Cdc20, which becomes one of its targets and a substrate of the Aurora kinase. (107–111).

In telophase, nuclei for each daughter cell form at the two poles, and the mitotic spindle apparatus disappears. Furthermore, nuclear membranes, nuclear lamina, nuclear pores, and nucleoli are reformed. The cell is now ready for cytokinesis, which is physical division of the cytoplasm. The cytokyme divides as actin/myosin filaments contract and pinch off the plasma membrane, which results in two daughter cells that enter into G0 or G1 for another round of division. The main checkpoint that exists during M phase in mammalian cells is the spindle checkpoint; it is in place to ensure proper mitotubule assembly, proper cell division, and that each daughter cell receives one copy of DNA.

Spindle checkpoint

The spindle checkpoint is activated when microtubules fail to attach to the kinetochores of each sister chromatid and/or when misalignment of chromosomes occurs along the metaphase plate (112–114). This mechanism blocks entry into anaphase and ensures proper segregation of the chromatids to opposite spindle poles. Misregulation of this checkpoint results in aneuploidy daughter cells after division (115, 116). Checkpoint proteins associated with kinetochores monitor microtubule-kinetochore attachment and tension; these proteins regulate this checkpoint by preventing cdc20 binding to the APC/C (117–119).

The main spindle checkpoint proteins include Mad1, Mad2, BubR1, Bub3, Mps1p, and CENP-E. These proteins act both independently and dependently of their interaction with kinetochores. Association of Mad2 with kinetochores and cdc20 requires the presence of Mad1 (120). At the kinetochores, Mdm2 is converted to a form that can bind and sequester cdc20 away from the APC/C, which results in its inhibition (121). Additionally, formation of the mitotic checkpoint complex BubR1/Bub3/Mad2/Cdc20 occurs independently of interaction with unattached kinetochores, and signals anaphase to wait by binding and inhibiting the APC/C (122, 123). An unattached kinetochore activates a kinase cascade that involves the dual-specificity kinase Mps1p and the serine/threonine kinase BubR1 and Bub3 that amplifies this wait signal (124, 125). Furthermore, Bub1 interacts directly with the kinesin-like protein CENP-E to regulate microtubule tension at kinetochores, which is also involved in regulation of the spindle checkpoint (126, 127). Thus, this checkpoint serves to inhibit the APC/C indirectly through cdc20 sequestration and directly through association with the mitotic checkpoint complex, and to regulate the tension at kinetochores required for anaphase entry (128).

DNA Damage Checkpoints

In addition to checkpoints that ensure normal cell-cycle progression, numerous DNA damage checkpoints exist in mammalian
cells. These checkpoints exist to regulate the highly conserved mechanisms that control DNA replication and mitosis to ensure mutations within the genome are not passed on to the daughter cells. Misregulation of these pathways is associated with genomic instability and cancer development. The key players involved in the DNA damage checkpoint cascade (Fig 2) include the DNA damage sensors ATM (Ataxia Telangiectasia M mutant), ATR (ATM and Rad3 Related), Rad1, Rad9, Hus1, and ATRIP, and the effectors Chk1/2 (Checkpoint Kinase 1/2), and CDC25.

**G1/S-phase checkpoint**
The primary DNA damage checkpoint is the G1/S checkpoint, which acts to prevent the replication initiation of damaged DNA. During G1 and even after passage through the restriction point (but prior to initiation of DNA synthesis), DNA damage activates two checkpoint-signaling pathways sequentially, and both pathways function to inhibit CDK2 activity. The first pathway initiated is p53-independent and is rapid and short-lived (129). This pathway results in phosphorylation and degradation of CDC25A (130, 131). DNA damage leads to the activation of ATM and ATR, which phosphorylate and activate Chk1 and Chk2 (132, 133). CDC25A is phosphorylated by these kinases, which target its ubiquitination and proteasomal degradation (134). As a result, the inhibitory phosphorylations of CDK2 are increased, which diminishes CDK2 activity. Ultimately, this lack of CDK2 activity inhibits the cdc45 leading to pre-RCs and the subsequent initiation of DNA replication to halt the cell cycle and to allow time to repair damaged DNA (133).

The second pathway activated in the presence of DNA damage prior to initiation of DNA synthesis acts in a p53-dependent manner. As stated above, the tumor suppressor p53 is a transcription factor, which acts primarily to increase expression of the CK1/p21^{*WAF1} during DNA damage. Like CDC25A, the activation of ATM/ATR promotes the phosphorylation of p53, which enhances the stability of p53 by preventing efficient interaction with the E3 ubiquitin ligase MDM2, which is a protein responsible for targeting p53 degradation (135). This mechanism leads to the transcription and accumulation of p21, which silences CDK2 activity to prevent cell-cycle progression and to allow for DNA repair (136). MDM2 is also a target of p53 transcription, which creates a negative feedback loop with p53 (137). After repair of damaged DNA has been completed, the checkpoint is turned off and progression into S phase resumes.

**S-phase checkpoints**
Cells that have passed the G1/S checkpoint are ready to begin S phase and DNA replication. The S-phase checkpoints are a group of three mechanistically distinguishable checkpoints (138) of which two respond directly to DNA damage. One is independent of ongoing replication and is activated in response to DNA double-stranded breaks (DSBs), which is known as the intra-S-phase checkpoint. The second checkpoint, the replication checkpoint, responds to replication fork stalling caused by the collision of replication machinery with DNA damage, direct inhibition of polymerases, or depletion of dNTPs. Although these two checkpoints respond to different forms of stress, both checkpoints prevent cell-cycle advance, inhibit ongoing replication, prevent origin firing, and stabilize the replication fork so that repair and replication resumption can occur. The third type of S-phase checkpoint is the SIM checkpoint. Currently, this checkpoint is not understood as well as the previous two, but it is known to prevent entry into mitosis when replication is stalled or incomplete. It acts to preserve genomic stability by preventing premature chromatin condensation and breaks at common fragile sites.

**The replication checkpoint**
The replication checkpoint is activated when the replication machinery encounters DNA damage or when the replicative polymerase is inhibited and stalled (139, 140). This checkpoint stabilizes stalled replication forks and signals for DNA damage repair while preventing exit from S phase. Stalling causes uncoupling of the helicase from the polymerase, which leads to DNA unwinding without subsequent new strand polymerization. This action leads to accumulation of ssDNA, which is a trigger for checkpoint activation (141–143). ssDNA is also believed to activate other checkpoints which include those initiated by DNA repair mechanisms such as nucleotide excision repair (144, 145) or resection of DSBs generated during homologous recombination (146, 147). The ssDNA is coated by RPA proteins (148, 149), which set up a scaffold for the recruitment and localization of DNA damage sensors in S phase. ATR is central to the replication checkpoint and is recruited to RPA-coated Figure 2 Brief model of DNA damage checkpoint signaling. DNA damage elicits a conserved response headed by the ATM and ATR kinases. Phosphorylation cascades and localization of mediators to sites of damage allows for signaling to the effector kinases Chk1 and Chk2. Chk1/2 elicit cell-cycle arrest through phosphorylation-dependent degradation of the Cdk2 family of phosphatases. Parallel activation of p53 by both ATM/ATR and Chk1/Chk2 leads to upregulation of the CDK inhibitor p21, which enforces cell-cycle arrest to a greater extent. See text for in-depth discussion of the checkpoint pathways.
ssDNA through its interaction with its binding partner, ATM-P (150–152). In addition, other sensors of DNA damage, including Rad17, which is an RIF-like clamp loader, and the 9-1-1 complex, which is a heterotrimeric clamp composed of Rad9, Rad1, and Hus1, are recruited to RPA coated ssDNA and serve to activate ATR and to help recruit and activate downstream mediators of the checkpoint (153–155). After ATR activation and recruitment/activation of other sensors, numerous proteins are recruited to the site of damage and act as mediators of the DNA damage-signaling cascade. Most mediators are involved in the activation of the effector kinase Chk1 (156). One mediator, Claspin, is recruited to sites of damage, is phosphorylated by ATR, and recruits Chk1 subsequently. Direct interaction between Claspin and Chk1 is required for phosphorylation and activation by ATR (157–160). Other mediators include BRCA1 and BRCA1C-terminal motif (BRCT)-containing proteins. These mediators form large multimeric complexes and are often visualized as nuclear foci by immunofluorescence microscopy (156, 161). MDC1 (a mediator of DNA damage-checkpoint protein 1) recruits mediators of the checkpoint, such as S3BP1 and NBS1 (162-164). These proteins function to maintain foci oligomerization and to promote ATR mediated phosphorylation of its substrates, which include all of these mediators and SMC1 (Structural maintenance of chromosomes 1). SMC1 is part of the cohesin complex and is required for sister chromatin cohesion in S phase (165, 166). Finally, Chk1 is recruited to the nuclear foci that contain the large scaffold of BRCT-containing proteins and is activated in an ATM,Rad179-1BRCA1C/Claspin dependent fashion (157–159). Then, Chk1 facilitates the checkpoint by phosphorylating family members (167) and p53 (see above for more detail on these events); this results in cell-cycle arrest, DNA repair, and survival choices.

The intra-S-phase checkpoint

Unlike the replication checkpoint, the intra-S-phase checkpoint does not require replication to be activated (138, 168). At the head of this checkpoint is the ATM protein kinase, which is a member of the PI3K family of protein kinases (including ATR and DNA-PK), ATM and the intra-S-phase checkpoint are activated by the detection of DSBs, which can be achieved without direct interaction of the replication machinery with sites of damage. Another interesting difference between the replication checkpoint and the intra-S-phase checkpoint is that activation of the latter does not alter the progression of active replication units, only the inhibition of late origin firing (169). Thus, the intra-S-phase checkpoint causes delays in, but not complete arrest of, S-phase progression (138).

Although the sensors of DSBs are not definitively known, two protein complexes serve as excellent candidates because of their ability to enhance ATM activity. These complexes are the MRN (Mre11-Nbs1-Rad50) complex and the Rad179-1-1 complex (discussed above). The MRN complex has nuclelease activity that localizes to DSBs independently of ATM. At sites of damage, it plays a role in activation of ATM, efficient phosphorylation of ATM substrates, and recession of DSBs (170-172). Although much of the checkpoint from here out involves the same mediators including S3BP1, BRCA1, MDC1, and SMC1, it has two more distinct features compared with the replication checkpoint.

The first feature involves the recession of DSBs, which activates a parallel ATR/ssDNA signaling cascade similar to that discussed above (146, 173, 174). The second feature involves the activation of Chk2. Unlike Chk1, which is only present in S and G2 phases, Chk2 is present throughout the entire cell cycle (175, 176). Chk2 also differs from Chk1 in that it must dimerize to be fully active (177–179), and in response to DNA damage, it becomes soluble in the nucleus and dissipates from damage sites as a mechanism to enhance signaling (180, 181). When phosphorylated by ATM, Chk2 plays similar roles as Chk1 in the degradation of CDC25 family members and phosphorylation of p53.

Although the replication and intra-S-phase checkpoints have distinct mechanisms of activation and signaling, the final goal is the same: to delay or to inhibit S-phase progression providing time and signaling events that lead to DNA repair, so that mutations are not transmitted to daughter cells in the ensuing mitotic division.

S/M checkpoint

The S/M checkpoint can be activated by replication inhibition or when DNA replication is not completed (182-186). This checkpoint signals through the ATR/Chk1 pathways and prevents premature chromatin condensation (PCC) and entry into mitosis (183, 185, 187). Depletion of ATR in Xenopus egg extracts or Chk1 in embryonic stem cells results in premature entry into mitosis prior to completion of replication (183, 185).

In addition, different regions of the genome replicate at different rates, and common fragile sites are known to be late replicating regions. These common fragile sites are often left unreplicated during mitotic entry (188–191). PCC causes breaks when fragile sites are not fully replicated (189). Therefore, mitotic delay is required to ensure the proper replication of the entire genome to prevent breaks that might occur because of PCC. Both ATR (187) and Chk1 (188) are involved in the stability of common fragile sites, which indicates that the S/M checkpoint is required to maintain genomic stability by ensuring proper replication prior to mitotic entry.

G2/M-phase checkpoint

The G2/M checkpoint acts to ensure that cells that experience DNA damage in G2 or that contain unresolved damage from the previous G1, or S phase do not initiate mitosis. Much like the G2 checkpoint and in contrast to the S checkpoints, cell-cycle arrest or delay that results from the G2 checkpoint involves a combination of acute/transient and delayed/sustained mechanisms. The acute/transient mechanisms involve the rapid posttranslational modification of effector proteins, whereas the delayed/sustained mechanism involves the alteration of transcriptional programs (192).

Of all molecules targeted in the G2/M checkpoint, cyclin B/CDK1 seems to be the most important as its activity stimulates mitotic entry directly. DNA damage in the G2 phase activates ATM/ATR pathways (as described above), which results in
mitotic entry to allow for repair of DNA lesions (68). The more delayed and prolonged mechanisms by which the checkpoint silences Cdk1 activity is through the activation of the p53 pathway. Activation of p53 is achieved by phosphorylation by ATM/ATR or Chk1/Chk2 and results in nuclear localization, tetramerization, and stimulation of p53 transcriptional activity toward p21CIP1, p27KIP1, and p16INK4a. BRCA1 can cooperate to achieve maximal inhibition of Cdk1 and to prevent mitotic entry to allow for repair of DNA lesions (68).

The centrosome also regulates the G2/M DNA damage response, and numerous checkpoint proteins are associated with the centrosome (194). Centrosome separation is regulated by the kinases Nek2 and Plk1, and this process is inhibited by DNA damage in an ATP-dependent manner. ATM activation leads to Plk1 and Nek2 inhibition, which results in deregulation of the centrosome (195). By this mechanism, centrosome separation is inhibited, and it contributes to maintaining the G2/M checkpoint (196). Plk1 is also known to phosphorylate and activate CDC25C (197). Thus, Plk1 inhibition also results in CDC25C inhibition, inactivation of cyclin B/CDK1, and a halt in cell-cycle progression. Normally, cell-cycle progression resumes when DNA damage repair is completed; otherwise, apoptosis prevents genomic instability and allows cell-cycle arrest and apoptosis. However, data from Saccharomyces cerevisiae, Xenopus, and human cells, suggests that pathways to re-enter cell-cycle progression exist even when unrepaired DNA damage is present. This process of “checkpoint adaptation” has been shown to allow mitotic entry in response to ionizing radiation in human cells in a Plk1-dependent manner, and it may promote carcinogenesis and genomic instability (198, 199). It has been speculated that activation of centrosomal cyclin B/CDK1 plays a central role in this process, and it may occur through Plk1-mediated degradation of Wee1 and/or inhibition of Chk1 activity that leads to stabilization of CDC25 (200). Although its function is not well understood, checkpoint adaptation has been proposed to move cells into a phase where they can die, allow progression into other phases where difficult DNA damage can be repaired, and even exist to allow natural evolution (201).

**Conclusions**

The mammalian cell cycle is controlled by numerous factors involved in regulation of CDKs and checkpoint responses. Although many proteins involved in the pathways that lead to activation or inactivation of these have been elucidated over the years, much remains to be explored. Although most CDKs control the cell division cycle, regulation of the cell cycle is clearly more than progression from growth to DNA synthesis to division and transmission of genetic material. Growing evidence exists for the role of CDKs in controlling the balance between senescence, cell growth, checkpoint activation, and apoptotic signaling. Clearly, the inability to respond properly to DNA damage and cellular stress through checkpoint activation and apoptosis has a role in oncogenic potential and therapeutic considerations. The identification of novel factors and signal cascades that mediate the regulation of the cell cycle will lead to new drug targets in the fight against cancer and numerous other diseases.

**References**

Cell Cycle: Regulation of

62. Zou L, Stillman B. Assembly of a complex containing Cdc6-Rpa, replication protein A, and Mcm2-7p at replication origins controlled.


118. Pinsky BA, Biggins S. The spindle checkpoint: tension versus
115. Bharadwaj R, Yu H. The spindle checkpoint, aneuploidy, and
112. Malmanche N, Maia A, Sunkel CE. The spindle assembly check-
111. Reis A, Levasseur M, Chang HY, Elliott DJ, Jones KT. The CRY
110. Prinz S, Hwang ES, Visintin R, Amon A. The regulation of Cdc20
109. Bembenek J, Yu H. Regulation of the anaphase-promoting com-
106. Waizenegger I, Gimenez-Abian JF, Wernic D, Peters JM. Regula-
105. Yu H. Regulation of APC-Cdc20 by the spindle checkpoint. Curr.
103. Palacios V, Solari R, Okumura Y, Boston JA, Earnshaw WC. The
101. Tsai MY, Wiese C, Cao K, Martin O, Donovan P, Ruderman
87. Kops GJ, Weaver BA, Cleveland DW. On the road to can-
69. Busico L, Chessa M, Draxta G, Donzell M. Cdc20 phospho-
66. Wu X, Bye A, Olen J, Levine A. The p53-mdm-2 autoreg-


188. Durkin SG, Arlt MF, Howlett NG, Glover TW. Depletion of CHK1, but not CHK2, induces chromosomal instability and breaks at common fragile sites. Oncogene 2006;25:4381-4388.

Further Reading

See Also
Cell Cycle
Cell Division, Small Molecules to Study
DNA Damage: An Overview
DNA Damage, Sensing of
DNA Replication, an Overview
The process of regulated or programmed cell death (PCD) executed through genetically encoded intrinsic cellular machinery is widely accepted to represent one of the key cellular responses to extrinsic and intrinsic stimulation. Extensive analysis of PCD carried out during the last decade clearly established that proper execution of PCD is important for normal mammalian development and also for homeostasis of the adult organism. Deregulation of PCD has been linked to development of many severe human diseases like cancer, autoimmunity, stroke, and some neurodegenerative diseases. Although apoptosis, which is the first discovered form of PCD, has been and remains the mainstay of PCD research, better understanding of the process of apoptosis led to the surprising discovery that other forms of PCD also exist and play multiple important roles in health and disease. Small-molecule inhibitors of apoptotic and nonapoptotic cell death have been successfully developed and proved very useful in defining the mechanisms and functional role of various PCD processes. Furthermore, some molecules have been developed extensively and represent emerging new therapies for human pathologies. In this article, we will discuss the major PCD-related protein targets of chemical inhibitors and describe the major classes of small molecule inhibitors developed to this point.

Introduction

The paradigm of programmed cell death (PCD) has emerged in the last decade as the critical mechanism of normal development and homeostasis of multicellular organisms. Apoptosis, which was the first discovered form of PCD (1), remains the central topic in the field of PCD research. Alterations to apoptotic signaling either through genetic or small-molecule means can cause significant developmental abnormalities, including mortality, and they can also lead to the development of serious pathologies in adult animals, such as cancer and autoimmunity.

Apoptosis is associated with several highly uniform and characteristic morphological changes, such as cell shrinkage, cell membrane blebbing, condensation of nuclear chromatin, micronuclei formation, extensive vacuolization of cytoplasm, and disintegration of the cell into small fragments (apoptotic bodies) (2), which are reflective of the highly conserved nature of the apoptotic execution machinery. In general, two different general pathways of apoptotic cell death initiation exist: extrinsic and intrinsic (Fig. 1). The extrinsic pathway is activated by external stimuli, such as engagement of death domain receptors (DRs) with their cognate ligands. The intrinsic pathway is stimulated by intracellular stress, such as DNA damage. The pathway of DR-induced apoptosis has been studied and characterized extensively. It involves recruitment of the proforms of cysteine proteases, caspase-8 and caspase-10, into the receptor-induced death inducing signaling complex (DISC), which results in their autocatalytic cleavage and activation. Apical caspases can trigger execution of apoptosis directly by processing and activation of the “effector” caspases (caspase-3, caspase-6, and caspase-7) (3). Effector caspases execute apoptosis by cleaving various cellular substrates, which leads to the orderly cell demise. In some types of cells, DR signaling relies on a mitochondrial amplification step, which is carried out through caspase-8 and caspase-10 mediated cleavage of the BH3 domain-only Bcl-2 family member Bid (4–6). Processed Bid translocates to the mitochondria, where it causes release of cytochrome c from the intermembrane space through induction.
Cell Death, Biological Mechanisms and Small Molecule Inhibitors of Caspases

Death-receptor signaling may involve direct caspase-8-mediated caspase-3 activation (type 1 cells) or a Bid-cleavage-dependent mitochondrial amplification step (type 2 cells).

Death-receptor signaling may involve direct caspase-8-mediated caspase-3 activation (type 1 cells) or a Bid-cleavage-dependent mitochondrial amplification step (type 2 cells). The intrinsic pathway of apoptotic death also proceeds through release of cytochrome c from mitochondria, and different forms of stress use different BH3-only factors for signaling.

Although apoptosis remains the mainstay of PCD research, rapidly accumulating evidence indicates that programmed or intrinsically regulated cell death can occur through additional pathways independent of caspase activation and other proapoptotic factors. Multiple alternative processes of regulated cell death have been described, including autophagic cell death, mitotic catastrophe, necroptosis, oncosis, and so on. In most cases, mechanisms of these processes are much less understood than that of apoptosis. Here, we will discuss some emerging small-molecule regulators of these pathways, which have proven helpful in understanding the processes of nonapoptotic cell death. We have omitted several important protein factors from our discussion, such as PI3-kinase, Akt, and p53, and the process of autophagy. Although corresponding cellular pathways make important contributions to the regulation of cell death, they play much broader role in cellular regulation and deserve separate discussion.

Caspases

The family of cysteine proteases, which are called caspases, plays a major role in the execution of apoptotic cell death. Many studies suggest that increased apoptosis and caspase activity contribute to tissue damage in both acute (e.g., myocardial infarction, stroke, sepsis, spinal cord injury) and chronic (e.g., Alzheimer’s, Parkinson’s, Huntington’s disease) human diseases. Caspase family members are also prominently

Figure 1 Schematic representation of the receptor-mediated (extrinsic) and the intracellular stress-mediated (intrinsic) pathways of caspase activation.
involved in inflammatory responses and are required for processing and secretion of proinflammatory cytokines (15). For example, deficiency in caspase-1 or caspase-11 leads to significant protection from septic shock (16, 17). Thus, inhibition of caspase activity has emerged as a promising direction for cytoprotective therapies. All caspases, with the exception of caspase-9, are expressed in the form of catalytically inactive single-chain zymogens that contain a large and a small subunit (18). Activation of the proform occurs by proteolytic cleavage that releases N-terminal pro-domain and separates large and small subunits. An active enzyme is a heterotetramer composed of two large and two small subunits with two identical active sites, which are formed with the contribution from both large and small subunits (19). A small molecule activator of pro-caspase-3, PAC-1 (Fig. 2c), has been identified recently in an in vitro screen of 20,500 compounds (20). This molecule displayed an EC\textsubscript{50} = 220 nM in an in vitro assay and EC\textsubscript{50} < 1 \mu M in cancer cell lines. Furthermore, this molecule was significantly more toxic to primary colon cancer cells compared with matched normal cells (at least 10-fold difference), which suggests that activation of procaspase-3 may represent a new approach for selective killing of cancer cells. Indeed, administration of PAC-1 was found to attenuate growth of three different types of cancer in vivo. However, the precise mechanism of caspase activation by PAC-1 has not yet been described. Unlike many other proteases, caspases possess a high degree of substrate specificity, that is, they display an almost absolute requirement for an aspartic acid residue in the P1 position of a substrate. This requirement stems from several hydrogen bonds that Asp forms within the caspase substrate-binding pocket (21). Furthermore, three preceding amino acid residues (P2–P4) of the substrate contribute to substrate recognition by specific caspase family members, although substrate selectivity of caspases is typically not absolute (22). Based on their substrate preferences, caspases can be divided into three groups (23). Group 1, including caspase-1, -4, and -5, prefers a hydrophobic amino acid in the P2 position.
acid in the P4 position. Group 2 displays strong preferences for Asp in the P4 position and includes caspase-2, -3, and -7. Activity of group 3 caspases, which consists of caspase-6, -8, -9, and -10, is less dependent on the identity of a P4 residue.

Most caspase-inhibitor design strategies target their active sites and are based on caspase substrate preferences. Tetrapeptide inhibitors, which are based on identified sequences of the four amino acid recognition motifs, were shown to inhibit caspase family members selectively (24) (Fig. 2a). Overall, a typical peptide-based caspase inhibitor consists of three major structural components:

1. the “warhead” moiety that interacts with an active site Cys of the caspase
2. Aasp in P1 position (invariable for all peptide-based inhibitors); and
3. P2-P4 sequence, which provides some selectivity toward individual caspase sub-classes (21) (Fig. 2b).

Selectivity of peptide substrates is mainly defined by P4 and to a lesser extent by P2 and P3 residues (Fig. 2b) (12, 22).

The warhead is an electrophilic group that reacts with the nucleophilic Cys of the active site to form reversible or irreversible adducts. Use of aldehyde, semicarbazone, or thiomethylketone groups leads to reversible caspase inhibitors, whereas fluoroacetamidinoketones, chloromethylketones and acyloxymethylketones are used to generate irreversible inhibitors (21). Although peptide inhibitors can be very useful in defining the functional role of apoptosis and caspases in particular cell death paradigm in vitro and even, in some cases, in vivo (25), peptide inhibitors possess several general disadvantages, including toxicity of the leaving group, limited half-life in cells and in vivo, modest selectivity toward individual caspases, and lack of oral bioavailability. These disadvantages limit their use mostly to cell-based studies. Specificity of these molecules toward caspases is somewhat limited, as for example, an irreversible pan-caspase inhibitor DAV-Dmk found to inhibit unrelated cysteine proteinases, such as cathespin B and H (26).

Several promising nonpeptide inhibitors of caspases have also been developed, which may overcome many limitations of tetrapeptide molecules. Lee et al. (27, 28) used high-throughput screening to identify 5-nitrosoin as a caspase inhibitor, which was optimized significantly to obtain selective, low-nanomolar inhibitors of caspase-3 and -7 (Compound 1, Fig. 2d) and caspases -2, -7 and -9 (Compound 2, Fig. 2d). In contrast to peptide inhibitors, these compounds displayed high activity despite the lack of interaction with 51 residue of caspase substrate pocket, and X-ray crystallography suggested that they primarily interacted with the S2 subsite. These compounds inhibited apoptosis in several different cell types, which include camptothecin-treated Jurkat T cells and chondrocytes and cyclophosphamide-treated neutrophils (27, 28).

Okamoto et al. (29) pursued development of peptidomimetic caspase-1 inhibitors based on the crystal structure of caspase-1/A-C-Tyr-Val-Ala-Asp-H complex, which resulted in an inhibitor displaying potent activity in an in vitro caspase assay (EC\textsubscript{50} = 38 nM) and blocking IL-1\textbeta processing in the cells with EC\textsubscript{50} = 230 nM. Furthermore, this molecule blocked IL-1\textbeta release in mice in a dose-dependent fashion. X-ray analysis revealed that naphthoyl- and methyl groups of the methanesulfonamidecarboxyl group, limited half-life in cells and in vivo, even, in some cases, with IL-1\textbeta. VX-740 was found to reduce inflammation and disease symptoms significantly in patients during Phase 2 clinical trials (33, 34). However, clinical trials also demonstrated significant liver toxicity associated with long-term dosing of this molecule, which is likely associated with activity of the warhead group. This finding led to premature termination of clinical trials. Synthesis of improved and highly selective caspase-1 inhibitor, VX-765, which is a prodruk resulting in 4-hydroxybutyrolactone “warhead” moiety, has also been reported recently (35, 36). This molecule efficiently blocked release of inflammatory cytokines in the cells and in vivo; however, no data from clinical trials has been reported thus far.

IDUN Pharmaceuticals (San Diego, CA) also reported development of a potent peptidomimetic irreversible oxamylpeptide inhibitor IDN-6556 that uses 2,3,5,6-tetrafluoroxyphenoxyacetylketone warhead (Fig. 2b) (37, 38). This molecule preferentially accumulates in the liver, which results in pronounced liver protection in animal models (for example, after Fas-induced liver injury (39)), and showed significant promise in clinical trials of acute alcoholic hepatitis, human liver preservation injury, and chronic hepatitis C (38, 40, 41).

To overcome limitations of peptidomimetic inhibitors, such as limited central bioavailability caused by accumulation in the liver, a novel strategy for rapid identification of nonpeptidomimetic caspase inhibitors, tethering has been recently proposed (42, 43). Tethering is based on covalent capture of sulfhydryl-containing small molecules that interact within the active site of caspase-3. For this assay, caspase is modified to contain a free thiol-bearing “extender” attached covalently to the active site cysteine, which is used for small molecule capture. Using this approach, a caspase-3 extender complex was screened against a library of fragments modified to contain a free sulfhydryl group. Selected fragments were subsequently combined with a reversibly binding form of extender, which resulted in rapid selection of potent and reversible caspase-3-specific inhibitor with K\textsubscript{i} of 2.8 nM. Additional chemical modifications enhanced its potency into low nM range (Fig. 2g) (42).

**BCL-2 Family**

Members of the Bcl-2 protein family are key regulators of the mitochondrial step in apoptotic pathway (44). Upregulation of antiapoptotic Bcl-2 family members is commonly observed in many types of cancers and is well established to play a major role in substrate recognition by caspases.

Vertex Pharmaceuticals (Cambridge, MA) reported development of several additional caspase-1 peptidomimetic inhibitors, such as VX-740 (Pranalacasan, Fig. 2e), for the treatment of rheumatoid arthritis (30–32). VX-740 is a pro-drug, which is converted in vivo into an aldehyde active form with K\textsubscript{i} of 1 nM against caspase-1. VX-740 was found to reduce inflammation and disease symptoms significantly in patients during Phase 2 clinical trials (33, 34). However, clinical trials also demonstrated significant liver toxicity associated with long-term dosing of this molecule, which is likely associated with activity of the warhead group. This finding led to premature termination of clinical trials. Synthesis of improved and highly selective caspase-1 inhibitor, VX-765, which is a prodruk resulting in 4-hydroxybutyrolactone “warhead” moiety, has also been reported recently (35, 36). This molecule efficiently blocked release of inflammatory cytokines in the cells and in vivo; however, no data from clinical trials has been reported thus far.

IDUN Pharmaceuticals (San Diego, CA) also reported development of a potent peptidomimetic irreversible oxamylpeptide inhibitor IDN-6556 that uses 2,3,5,6-tetrafluoroxyphenoxyacetylketone warhead (Fig. 2b) (37, 38). This molecule preferentially accumulates in the liver, which results in pronounced liver protection in animal models (for example, after Fas-induced liver injury (39)), and showed significant promise in clinical trials of acute alcoholic hepatitis, human liver preservation injury, and chronic hepatitis C (38, 40, 41).
role in apoptosis evasion of cancer cells under chemotherapeutic treatment conditions (45). The Bcl-2 family can be further subdivided into three classes of proteins. The proapoptotic members Bax and Bak activate apoptosis through formation of a pore in the outer mitochondrial membrane, which results in cytochrome c release and activation of the apoptosis (46). The primary function of antiapoptotic proteins Bcl-2, Bcl-xL, Bcl-w, Mcl-1, and Bfl-1 is to inhibit the functions of Bak and Bax by preventing their oligomerization (47). The members of the third “BH3-only” group have homology with the other family members only in the BH3 domain and serve as upstream sensors of apoptotic signaling. Once activated by an apoptotic signal, BH3-only proteins are proposed to act through two different mechanisms. Some BH3-only factors, which are termed “sensitizers” (Bad, Bik), may act primarily by inhibiting antiapoptotic Bcl-2 family members through BH3-mediated binding to the hydrophobic cleft formed by BH1, BH2, and BH3 domains of antiapoptotic factors. A other subgroup of BH3-only factors, termed “activators,” including BID and Bim, were proposed to activate Bax and Bak directly, which induces their oligomerization (48).

Initial studies of BH3-dependent heterodimerization found that an isolated 16 a.a. BH3 peptides derived from several Bcl-2 family proteins can bind antiapoptotic Bcl-2 family members with submicromolar affinity (49) and agonize Bcl-xL heterodimerization with proapoptotic proteins Bax and Bad (50, 51). The ability of synthetic BH3 peptides to trigger apoptosis was first demonstrated in a cell-free system based on extracts of Xenopus eggs (52). In these studies, BH3 peptides derived from Bak, Bax, and Bad were all found to induce apoptosis through rapid activation of caspases. Because BH3 peptides cannot permeate the cell readily, several strategies were used to generate cell-permeable BH3 peptide-based proapoptotic agents. Wang et al. (53) demonstrated that attachment of decanoic acid allows generation of cell-permeable BH3 peptides. One such peptide, termed CPM-1285 and containing the BH3 domain of mouse Bad, was shown to compete with a fluorescein-labeled Bax BH3 peptide for binding to Bcl-2 (IC50 of 130 nM) in vitro and to trigger apoptosis in human myeloid leukemia HL-60 cells. In another approach, Bcl-2 BH3 peptide was fused to an internal “crosslinks” into Bcl-xL peptides, which led to stabilization of a helical structure, increased cell permeability, and bioavailability. Subsequently, a “stapled” Bcl-xL BH3 peptide efficiently induced apoptosis in leukemia cells in vitro and in vivo, and this activity was increased even more by membrane targeting of the peptide (57). The first proapoptotic inhibitor of Bcl-2 family proteins was identified by Wang et al. (58) in 2000 using a virtual-screening strategy. This method relies on the high-resolution three-dimensional structure of a targeted receptor protein and computer-aided techniques to screen a large number of organic compounds for a potential ligand. Virtual screening of more than 190,000 organic molecules resulted in identification HA14-1 (Fig. 3b) and subsequent in vitro binding assay demonstrated the interaction of HA14-1 with the surface pocket of Bcl-2 with an IC50 value of 9 µM. Subsequently, multiple research groups reported activation of apoptosis by HA14-1 in a variety of cell types through mechanisms related to regulation of the Bcl-2 family as well as retardation of glioblastoma tumor growth in vivo when this molecule was combined with etoposide (59). The group of Dr. Shaomeng Wang (University of Michigan) also successfully used computational strategies to identify a number of different submicromolar small molecule antagonists of antiapoptotic Bcl-2 family members (TW-37, Fig. 3c, IC50 = 290 nM (60)); pyrogallol-based inhibitors, Fig. 3d, IC50 = 110 nM (Bcl-2) (61); flavanoid compound B1-33, Fig. 3e, IC50 = 17 nM (Bcl-2) (62)). These molecules were all found to trigger apoptosis in tissue culture, with some molecules displaying very potent effect (IC50 for Bcl-2 in MDA-MB-231 breast cancer cells = 110 nM). Furthermore, one of these molecules, TW-37, was found to enhance the antitumor effect of standard chemotherapy (cyclophosphamide-doxorubicin-vincristine-prednisone, CHOP) in mouse lymphoma model (63). Overall, these data suggest that computational approaches can be very powerful in designing proapoptotic inhibitors of the Bcl-2 family.

Using a competitive binding assay based on fluorescence polarization (FP), Depterev et al. (64) screened a chemical library of 16,320 compounds to identify two classes of small molecule inhibitors of Bcl-xL, which are termed BH3I-1 and BH3I-2 (Fig. 3a). These compounds were shown to inhibit BH3 peptide binding to Bcl-xL and Bcl-2 with Ki values in the low micromolar range (Ki of 2.4-15.6 µM) as determined by NMR titration assays (64, 65). The NMR titration experiments suggested that BH3I molecules directly interact with the BH3 binding pocket in disrupting Bcl-xL heterodimerization. These compounds were found to induce apoptosis in Jurkat cells through disruption of Bcl-2/Bcl-xL heterodimerization measured in intact cells (64). In addition to affecting Bcl-2-dependent regulation of outer mitochondrial membrane permeability, BH3I-2 was also found to induce damage to inner mitochondrial membrane, likely also through interaction with Bcl-2 (63). In another report, small scale FP-based screening of polyphenols identified gossypol (Fig. 3f) as a novel Bcl-2 and Bcl-xL inhibitor (66). This molecule was found to possess significant activity as a sensitizer when combined with CHOP in a mouse lymphoma model (67) and ionizing radiation in prostate cancer xenograft (68). Using a similar FP screen, PKC inhibitor cheterythrine (Fig. 3g) was identified as a low micromolar Bcl-xL/BH3 inhibitor (69). Curiously, NMR analysis and molecular docking suggested that unlike BH3I-1, this and related sauraphagine molecules may bind to the hydrophobic cleft of Bcl-xL, but rather to the BH groove and BH1 domain, respectively. This analysis suggests a distinct mechanism of Bcl-xL domain displacement, which was proposed to explain increased cytotoxicity of these molecules compared with BH3I-1 (70). Researchers at Abbott Laboratories used a different approach for discovering high-affinity protein ligands, the “structure-activity relationships by nuclear magnetic resonance” (“SAR by NMR”) (71). In this method, the relatively large site is divided into two smaller half-sites that are targeted individually by small molecules. The BH3 binding region of Bcl-xL was first divided into two smaller half-sites that are targeted individually by small molecules.
molecules. The two lead molecules are then chemically linked to improve affinity. In this approach, although the two molecules that target each half displayed $K_i$ values of only 0.3 and 4.3 mM, the combined molecule displayed a $K_i$ of 36 nM against Bcl-2. Subsequent chemical modifications to improve affinity and decrease nonspecific binding to human serum albumin yielded ABT-737 molecule with a $K_i < 1$ nM for Bcl-xL, Bcl-2, and Bcl-w (Fig. 3i) (72). ABT-737 was found to kill the cells efficiently through Bcl-2 or Bcl-xL-dependent mechanism. Furthermore, ABT-737 induced cytochrome c release from isolated mitochondria, which was dependent on inhibition of Bax and Bak by Bcl-2 (72). In other words, ABT-737 was found to antagonize prosurvival activity of Bcl-2 aimed at inhibition of Bax and Bak. Furthermore, Oltersdorf et al. (72) demonstrated that ABT-737 can act as a selective cancer therapeutic drug that displays potent single-agent efficacy against small cell lung cancer (SCLC) cells and cells from lymphoid malignancies, which are known to express high levels of Bcl-2, with EC50 as low as 10 nM. In mouse xenograft models, ABT-737 treatment provided significantly improved survival in mice injected with either lymphoma or SCLC cell lines (73). Curiously, very high specificity of ABT-737 binding to Bcl-2 may actually limit efficacy of this molecule as it has been found not to inhibit activity of the antiapoptotic Bcl-2 family member, Mcl-1, which results in reduced activity in multiple cancer cell lines (74, 75).

Another small-molecule pan-Bcl-2 inhibitor that mimicks BH3-only proteins, GX15-070 (developed by GeminX, Montreal, Canada) (Fig. 3k), was shown to be a potent apoptosis inducer in breast cancer, chronic lymphocytic leukemia, multiple myeloma, and mantle cell lymphoma cell lines (76-79).
GeminiX is currently conducting several clinical trials of GX-15-070 in multiple cancer types as a single agent and in combination with other agents.

The laboratory of Dr. David Hackenberry discovered that increased sensitivity of Bcl-2 and Bcl-xL-expressing cells to mitochondrial respiratory chain inhibitor, antimycin A, is caused by its direct interaction with Bcl-2 family members (80). Furthermore, this effect is retained by 2-methoxy antimycin A (2MAA) (Fig. 3j), an analog lacking ability to inhibit complex 3 of the respiratory chain. Curiously, activity of 2MAA seems very different from that of all of the abovementioned Bcl-2/Bcl-xL inhibitors, as 2MAA displays preferential toxicity toward Bcl-2/Bcl-xL-overexpressing cells. This activity may be caused by the unique ability of 2MAA to antagonize Bcl-xL-dependent changes in cell metabolism, namely reduction in oxidative phosphorylation and activation of glycolysis (88). Such "gain-of-function" Bcl-xL antagonists may be very beneficial for preferentially inducing cell death in Bcl-2-overexpressing cancers.

The Iap, Xiap, and Smac Mimetic Peptides

The inhibitor of apoptosis proteins (IAPs), which are characterized by the presence of one or more baculovirus IAP repeat (BIR) domains, are a family of endogenous apoptosis inhibitors that possess multiple antiparotic activities, including binding and inhibition of active caspases 3, 7, and 9. By inhibiting the downstream caspases 3 and 7, IAPs block the convergence point of multiple caspase activation pathways and thus inhibit apoptosis induced by various stimuli (81). At least eight human IAP members have been identified, of which XIAP (X-linked IAP) and survivin have received the most attention as therapeutic targets (82).

Survivin is a bifunctional protein that acts as a suppressor of apoptosis and plays a central role in the regulation of cell division. Survivin is preferentially expressed in malignant cells, and its expression is frequently responsible for radioresistance of malignancies (83, 84). However, this effect may not be linked to the direct regulation of apoptosis, but rather to the regulation of cell division. Cell-cycle-dependent transcriptional regulation of the survivin gene (85) as well as posttranslational modifications, including phosphorylation by the p38MAPK (86), were found to be essential for the cell-cycle control. Based on the finding that pharmacologic inhibition of mitotic phosphorylation of survivin accelerated the protein destruction and counteracted its function (87), CDK inhibitors such as flavopiridol (Fig. 4a) and purvalanol A (Fig. 4b), which is a more specific p34cdc2 inhibitor, were tested in tumors induced by arsanilic acid with taxol, which induces hyperphosphorylation of survivin (88). Administration of CDK inhibitors resulted in escape from the mitotic block imposed by taxol, activation of mitochondria-dependent apoptosis, and anticancer activity in vivo (87). The stability and function of survivin depends on physical interaction between its BIR domain and ATPase domain of the molecular chaperone heat shock protein 90 (HSP90). Targeted antibody-mediated disruption of the survivin-Hsp90 complex in cancer cells resulted in professional degradation of survivin, mitotic arrest, and mitochondria-dependent apoptosis (89). A structure-based rational screening for antagonists of the survivin-HSP90 complex identified a cell-permeable peptidomimetic derived from the Lys98-Leu99 sequence of the survivin called sheepdin (90). Sheepdin inhibited HSP90 chaperone function by competing with ATP binding and destabilized several HSP90 client proteins, including A1I, CDK6, and telomerase, to induce cell death via apoptotic and nonapoptotic mechanisms in various tumor cell lines. Sheepdin (79–83), which is a cell-permeable five-residue peptide that contains the Lys98-Gly99 sequence of sheepdin essential for HSP90 binding (91), induced rapid killing of different types of human acute myeloid leukemia (AML) cell lines, but not of normal mononuclear cells. Moreover, sheepdin (79–83) efficiently inhibited the growth of AML xenograft tumors without systemic or organ toxicity (91). More recently, a combined structure- and dynamics-based computational design strategy using sheepdin as a scaffold identified the nonpeptidic small molecule that targeted the HSP90 function, 5-aminomimidazole-4-carboxamide-1-β-D-riboside monophosphoramide (AICAR, also a known activator of AMP kinase (AMPK)) (Fig. 4c) (92). AICAR was shown to destabilize several HSP90 client proteins in vivo, including survivin, and to exhibit antiproliferative and proapoptotic activity in multiple tumor cell lines, but not in normal human fibroblasts. Finally, a small molecule that selectively inhibits survivin gene transcription and protein expression has been identified, YM155 (Fig. 4d), and is currently being evaluated in a Phase 2 study for patients with stage 3 and stage 4 melanoma. It showed marked antiproliferative activity in the nanomolar range in a broad spectrum of human tumor cell lines and induced tumor regression in lymphoma, prostate cancer, and non-small cell lung cancer xenografts (93).

XIAP is the best characterized human IAP and is the only member of this family shown to inhibit both the initiator (caspase-9) and executor (caspases-3 and -7) caspases directly. Structural and functional studies of XIAP have demonstrated that a groove in its BIR3 domain is required for binding and inhibition of caspase-9, whereas two surfaces of the BIR2 domain and the juxtaposed linker region bind and inhibit active caspases-3 and -7 (94, 95). The natural inhibitor of XIAP, cIAP1, and cIAP2, the proapoptotic protein SMAC/DIABLO, is released into the cytosol from the mitochondrial intermembrane space during apoptosis activation (96, 97). XIAP dimers cooperatively bind and inhibit both BIR3 and BIR2 domains of IAPs and thus relieve their caspase inhibitory function (98). Peptides that correspond to the four N-terminal amino acids of SMAC (AVP1) were shown to be sufficient for binding to XIAP and preventing XIAP-mediated inhibition of caspase-9 (99, 100). On the other hand, the SMAC peptides cannot relieve the inhibition of caspase-3 by XIAP, because they do not change the XIAP conformation around the linker region (101). When delivered into the cells either by conjugation with the TAT protein-transduction domain (102) or to polyarginine tail (103), SMAC peptides sensitized the SHEP neuroblastoma
Cell Death: Biological Mechanisms and Small Molecule Inhibitors of XIAP and Omi/HtrA2

Figure 4

Structures of small molecule inhibitors of (a, b) CDK, (c) HSP90, (d) survivin gene expression, (e–l) XIAP, and (m) Omi/HtrA2. (a) flavopiridol, (b) purvalanol, (c) 5-aminoimidazole-4-carboxamide-1-beta-D-ribofuranoside (AICAR), (d) YM155, (e–h) SMAC-mimetics, (i) Smac mimetic tetraoxyl thioether, (j) benzoquinone embelin, (k) aryl sulfonamide, (l) XIAP inhibitor of Polyphenylurea series, (m) Omi/HtrA2 inhibitor Ucf-101.

(102) and non-small cell lung carcinoma H460 (103) cell lines to apoptotic cell death induced by chemotherapeutic drug treatment. In addition, cell-permeable SMAC peptide delayed tumor growth of lung cancer and glioma xenografts (102, 103). These studies provided the proof-of-concept that small-molecule SMAC-mimics can be effective as the anticancer agents (82).

Nonpeptidic small molecule inhibitors of BIR3 domain of XIAP with micromolar binding affinities were synthesized at Abbott Laboratories (Abbott Park, IL) using structure-based design (104). Having conserved the first amino acid residue of SMAC peptide alanine, the substituted five-membered heterocycles such as thiazoles and imidazoles were identified to serve as a replacement for peptide fragments of the lead (Fig. 4e). Several research groups have also reported the discovery of nonpeptidic XIAP inhibitors active in the cell. Using a high-throughput fluorescent polarization assay, pentapeptides competing with the binding of SMAC-like protein H10 to BIR3 domain of XIAP with affinities in 40–60-nM range were identified (105). Tripeptide peptidomimetics (Fig. 4f) based on these leads were shown to inhibit the interactions between the BIR3 domain of XIAP and SMAC, caspase-9, and SMAC derived peptide (105). When the cytotoxicity of selected peptidomimetics was assessed in various cancer cell lines (105), compounds revealed a wide range of potencies from low nanomolar activity in some cell lines to no activity at 50 µM in most others. The toxicity exhibited by peptidomimetic BIR3 ligands in the xenograft cell lines (breast cancer cell lines BT-549 and MDA-MB-231, melanoma cell line SK-MEL-5, and human myeloid leukemia HL-60 cell line) was observed in the absence of additional apoptotic stimulation (105). Furthermore, the selected peptidomimetics were found to slow the growth of tumors in a MDA-MB-231 breast cancer xenograft model (105). Another research group reported synthesis of a SMAC-mimetic, which was approximately 23 times more potent than SMAC peptide in binding the BIR3 domain of XIAP (Fig. 4g) (106). This compound efficiently inhibited the growth of etoposide-treated Jurkat leukemia T cells, which protects the cells from etoposide-induced apoptosis.

The broad-spectrum peptidomimetic IAP family inhibitors were also recently developed (107). Designed based on (7, 5)-bicyclic scaffold, these SMAC-mimetics were found to antagonize the protein interactions that involve XIAP, melanoma...
IAP (ML-IAP), cIAP1, and cIAP2. The most potent SMAC mimetic was more specific for cIAP1-BIR3 and ML-IAP-BIR with \( K_i \approx 50 \text{ nM} \) (Fig. 4h). The compounds were demonstrated to activate caspase-3 and -7, to reduce cell viability in assays using MDA-MB-231 breast cancer cells and A2058 melanoma cells and to enhance docosahexaenoic acid-induced apoptosis in MDA-MB-231 cells.

Computer-based rational drug design was successfully used to synthesize a tetrazoyl thioether (Fig. 4j), a dimeric SMAC mimetic that binds both the BIR3 and BIR2 domains of XIAP with nanomolar affinity (108). Furthermore, this molecule also cross-reacted with cIAP1 and cIAP2. The potency of this dimeric inhibitor is consistent with the recently reported synergistic BIR2/BIR3 inhibition by dimeric SMAC peptide, which is explained by close proximity (4-5 angstrom) between BIR2 and BIR3 binding sites (109). The tetrazoyl thioether sensitized cells to the death induced by death receptor ligands (tumor necrosis factor alpha (TNFα) and TNF-related apoptosis-inducing ligand (TRAIL)) and promoted the activation of caspase-8. This result was unexpected because XIAP is not known to inhibit caspase-8 directly. Rather, recent analyses suggested that killing by this molecule is primarily mediated by targeting cIAP1 and cIAP2, which promotes their autoubiquitination and degradation (110). This, in turn, leads to stabilization of NIK kinase, NF-kB activation, and TNFα production, triggering apoptosis in an autocrine mode. Furthermore, cIAP degradation promotes TRAF2-dependent RIP1 recruitment to TNFR1, which promotes formation of RIP1-dependent caspase-8-activating complex. These unexpected findings suggest that cIAPs may be more important targets for SMAC mimetic compounds than XIAP, for which this class of molecules was originally developed.

Virtual screening was also used to identify the BIR3 domain inhibitors (111). In this case, a library of Chinese herbal remedies was docked into the BIR3 domain main inhibitors (111). In this case, a library of Chinese herbal remedies was demonstrated to reduce cell viability in assays using MDA-MB-231 breast cancer cells and A2058 melanoma cells and to enhance docosahexaenoic acid-induced apoptosis in MDA-MB-231 cells.

Computer-based rational drug design was successfully used to synthesize a tetrazoyl thioether (Fig. 4j), a dimeric SMAC mimetic that binds both the BIR3 and BIR2 domains of XIAP with nanomolar affinity (108). Furthermore, this molecule also cross-reacted with cIAP1 and cIAP2. The potency of this dimeric inhibitor is consistent with the recently reported synergistic BIR2/BIR3 inhibition by dimeric SMAC peptide, which is explained by close proximity (4-5 angstrom) between BIR2 and BIR3 binding sites (109). The tetrazoyl thioether sensitized cells to the death induced by death receptor ligands (tumor necrosis factor alpha (TNFα) and TNF-related apoptosis-inducing ligand (TRAIL)) and promoted the activation of caspase-8. This result was unexpected because XIAP is not known to inhibit caspase-8 directly. Rather, recent analyses suggested that killing by this molecule is primarily mediated by targeting cIAP1 and cIAP2, which promotes their autoubiquitination and degradation (110). This, in turn, leads to stabilization of NIK kinase, NF-kB activation, and TNFα production, triggering apoptosis in an autocrine mode. Furthermore, cIAP degradation promotes TRAF2-dependent RIP1 recruitment to TNFR1, which promotes formation of RIP1-dependent caspase-8-activating complex. These unexpected findings suggest that cIAPs may be more important targets for SMAC mimetic compounds than XIAP, for which this class of molecules was originally developed.

Virtual screening was also used to identify the BIR3 domain inhibitors (111). In this case, a library of Chinese herbal remedies was docked into the BIR3 domain in silico, which resulted in selection of natural compound benzoxazinone emebilin (Fig. 4i). Subsequent studies confirmed that emebilin binds to the XIAP BIR3 domain, which resulted in the inhibition of its interaction with caspase-9, and induced apoptosis in the prostate cancer cells expressing high levels of XIAP. In stably XIAP-transfected Jurkat cells, emebilin was shown to overcome the protective effect of XIAP effectively, which enhanced etoposide-induced apoptosis. At the same time, this molecule had a minimal effect in Jurkat cells transfected with vector control.

The BIR2 domain of XIAP has also been specifically targeted for inhibition. The linker region immediately to the N-terminus of the BIR 2 domain binds the catalytic domain of caspase-3 and blocks the active site of the enzyme through steric hindrance (94, 112, 113). This interaction is relatively weak and is stabilized by a stronger interaction between the binding groove of BIR 2 and a site on a small subunit of caspase 3 (94, 112). Using a high-throughput enzymatic de-repression assay based on caspase-3 proteolytic activity, Wu et al. (114) identified a series of aryl sulphonamide inhibitors of XIAP (Fig. 4k).

These molecules bind the BIR2-linker region and were found to sensitize resistant cell lines to death triggered by the acti-
Poly(Adp-Ribose) Polymerase (Parp)

The polyADP-ribosilation reactions play important roles in many cellular processes, which include regulation of DNA repair, transcriptional control, cellular transformation, and cell death (123). PARP is an enzyme-sensing single and double strand DNA nicks, which catalyzes addition of ADP-ribose units to DNA, histones, and various DNA repair enzymes (using NAD$^+$ as a substrate) to promote DNA repair. Mouse knockout studies showed that combined deletion of just 2 of the 17 PARP family members (PARP-1 and PARP-2) is sufficient to block DNA repair (124, 125). Curiously, multiple studies suggested that combining inhibition of DNA repair with the use of DNA-damaging agents hypersensitizes cancer cells to cell death, which prompted development of PARP inhibitors as the general sensitizing anticancer agents (126). In addition, tumors deficient in DNA repair-associated factors BRCA1, BRCA2, and ATM, were all shown to be hypersensitive to PARP inhibitors, which suggests that PARP inhibitors can be useful in killing these types of cancer cells (127, 128).

At the same time, overactivation of PARP, which is frequently observed in various pathologies, including cardiovascular, neurological, and inflammatory diseases, was shown to result in the depletion of NAD$^+$, leading to the loss of ATP and necrotic cell death. Cell-based studies showed that overactivated PARP-1 mediates both mitochondrial-dependent apoptosis and necrosis (129, 130). Consistent with this notion, genetic deletion of PARP rendered mice resistant to experimental stroke (131), providing rationale for developing PARP inhibitors as the cytoprotective agents.

Several PARP-1 inhibitors were designed based on 3-amino-benzamide (132) (Fig. 5a), but these molecules lacked specificity and potency (133). A newer early inhibitor, benadrostin, showed high activity in vitro for PARP-1, but still required high concentrations (10-100 uM) for chemopotentiation (137, 138). Subsequently, significant effort has been spent on structure-based drug design utilizing information generated using known inhibitors, which ultimately led to the identification of a number of potent PARP inhibitors. Five of these molecules are currently in clinical trials for oncologic indications, that is, AG014699 (139) (Fig. 5d), KU59436 (140–142), and BSI-201 (BiPar, Brisbane, CA), INO-1001 (140–142), and GPI 21016 (MGI Pharma, Bloomington, MN). ABT-888 (143, 144) (Fig. 5e) is expected to enter clinical trials shortly (145). All of these molecules are low-nanomolar PARP inhibitors, which sensitize cancer cells effectively to chemotherapy or radiotherapy at nanomolar concentrations. In animal studies, all of these molecules are well tolerated and effectively synergize with multiple DNA-damaging anticancer agents.

Conversely, PARP inhibitors showed significant promise as cytoprotective agents in animal models of inflammation, stroke, Parkinson’s disease, spinal cord injury, and myocardial infarction (146–148). In addition, PARP inhibitors showed activity in inhibiting various types of injury associated with type 1 and 2 diabetes, including neuropathy, retinopathy, and nephropathy, as well as beta cell death in a streptozotocin-injection model of type 1 diabetes (149). Overall, PARP has emerged as a very promising therapeutic target, especially in treating cancer, whereas a general lack of success in developing cytoprotective treatments has made this direction of PARP inhibitor development more challenging.

Necroptosis

Necroptosis or programmed necrosis is a novel type of regulated nonapoptotic cell death that was recently described by several research groups (150–153). It was found that in some cell types, stimulation of DR with their cognate ligands (TNFα, FasL, or TRAIL) under specific conditions where apoptosis is inhibited leads to the cellular demise with necrotic morphological features. Similar observations were also reported for cell death induced by some oncogenes (Ras, cMyc) (154, 155), chemoattractive agents (etoposide, camptothecin, and staurosporine) (156–158), and viral and bacterial agents (159, 160).

This unique type of cell demise shares the characteristics of both apoptosis (as a regulated form of cell death) and unregulated pathologic necrosis (by the cellular morphology) (161). Thus, discovery of necroptosis may offer an opportunity for therapeutic targeting the pathologic necrosis, because, in contrary to previously accepted views, it may represent a regulated and, therefore, specifically, inhibitable form of cell death. A clivation of necrosis-like death under apoptosis-suppressive conditions has been observed in various mouse models of acute pathologic death, including experimental pancreatitis (162) and multiple organ failure (163).

The signaling pathway of necroptosis is just beginning to emerge with Holier et al. (150) establishing that Ser/Thr kinase activity of DR-associated adaptor molecule, RIP1, is a key specific upstream activator of necroptosis. RIP1 kinase is a client protein of the molecular chaperone heat shock protein 90 (Hsp90), and inhibition of Hsp90 in the cells by the small molecule geldanamycin was found to result in efficient proteasome-mediated degradation of the RIP1 (164). As a result, geldanamycin (Fig. 5f) has been found to inhibit the activation of necroptosis in human Jurkat T cells (150). However, this effect is not specific to necroptosis as geldanamycin also was shown to block the RIP1-dependent NF-kB activation efficiency (164), which is independent of the RIP1 kinase activity, unlike necroptosis (165). In another study, Temkin et al. (166) have suggested that RIP1 may translocate to the mitochondria and lead to disruption of the VDAC/AntiCyclophilin D complex through an unidentified intermediate mechanism.

Although the mechanisms of necrotic cell death downstream of RIP1 are mostly unclear, inhibition of certain cell-signaling pathways has been found to attenuate necrotic cell death. Overproduction of reactive oxygen species (ROS) is a hallmark of necrotic cell death and is a prominent part of necroptosis in some systems (161). ROS production in conditions of

© 2008, John Wiley & Sons, Inc.
necrotic cell death is mediated by the mitochondrial respiratory chain complexes 1 and 2 (167) and/or through formation of RIP1/Rac1/nicotinamide adenine dinucleotide phosphate (NADPH) oxidase complex resulting in an oxidative burst (168).

The antioxidant 3-tert-butyl-4-hydroxyanisol (BHA) (Fig. 5g) has been found to be effective in blocking necroptosis in mouse fibrosarcoma L929 cells and in mouse embryonic fibroblasts (169), but not in human Jurkat T cells (161). Based on the data that closely related antioxidant butylhydroxytoluene (BHT) (Fig. 5h) was not as effective in inhibiting necrotic cell death as BHA, it was proposed that additional inhibitory activities of BHA, like inhibition of mitochondrial Complex 1 and/or of lipid peroxidation, may be critical for its inhibition of necroptosis (169). Consistent with the role of the respiratory chain in ROS generation during necroptosis, inhibitors of complexes 1 (rotenone, amyntal, Fig. 5j, 5i) and 2 (thenoyltrifluoroacetone, Fig. 5k), but not of complex 4 (cytochrome c oxidase) of the mitochondrial respiratory chain, have provided marked attenuation of cell death (167).

Autophagy is an important large-scale cellular catabolic process (see Reference 170 for review), and it is prominently activated as a part of necroptosis in many systems (161, 171-173). However, inhibition of autophagy with 3-methyladenine has been found to inhibit necroptosis in some cell lines, such as mouse fibrosarcoma L929 cells, but not in the other cell lines (human Jurkat T cells or mouse embryonic fibroblasts), which suggests cell type-specific contribution of autophagy to necrotic cell demise (161).

Activation of acid sphingomyelinase (A-SMase) and ceramide production were also shown to contribute to DR-induced necrosis (174). Inhibition of A-SMase activation by small-molecule inhibitors D609 (Fig. 5l) and desipramine (Fig. 5m) has been reported to attenuate necroptosis (174). D609 inhibits A-SMase activation by small-molecule inhibitors D609 (Fig. 5l) and desipramine (Fig. 5m) has been reported to attenuate necroptosis (174). D609 inhibits A-SMase induction indirectly by inhibiting the upstream-acting phosphatidylcholine-specific phospholipase C, whereas desipramine causes rapid and irreversible degradation of A-SMase. Finally, inhibition of NADPH oxidase C-Jun N-terminal kinase (JNK) axis of the pathway using siRNA tools has been shown to
attenuate necroptosis (168). Although it has not yet been tested directly, it is likely that small-molecule inhibitors of NADPH oxidase, such as diphenyleneiodonium chloride (DPI) (Fig. 5n), and of H/K ATPases, such as SK&F60025 (Fig. 5o), may interfere with necrosis activation. It is important to note that none of the above-mentioned inhibitors is specific for necroptosis, because their target proteins are involved in a wide range of cellular regulatory networks. An additional analysis of the mechanisms of specific activation of the downstream pathways of necroptosis by RIP1 kinase will be important for developing more specific strategies for necrosis inhibition.

The first specific inhibitor of necroptosis, necrostatin-1 (Nec-1), was identified by Degterev et al. (161) in a cell-based screen of ~15,000 compounds. Nec-1 (Fig. 5p) efficiently blocked necrotic death of human monocyctic U937 cells stimulated with TNFα in the presence of broad-spectrum caspase inhibitors zVAD-fmk and other instances of necrotic death (161). Although Nec-1 did not inhibit either activation of apoptosis or NF-κB by TNFα, it completely eliminated all the manifestations of cellular necrosis. Furthermore, optimized derivatives of this molecule were reported with EC50 in the cells of 50 nM (Fig. 5q) (175). A different screening resulted in identification of the other potent inhibitors of necroptosis, termed Nec-3 and Nec-5 (Fig. 5r, s) (175, 176), which created a unique panel of nanomolar inhibitors of this process. Based on implication that necroptosis is responsible, at least partially, for the pathologic necrosis, the necrosatins were tested for the cytoprotective effects in vivo rodent models of acute organ injury, which included cerebral (161) and cardiac (177) ischemia and brain trauma (178). Necrostatins were found to provide significant cytoprotective effect and functional improvement in multiple paradigms of acute injury. In the case of brain ischemia, Nec-1 displayed protection when administered up to 6 hours after 2-hour middle cerebral artery occlusion, which suggested that necroptosis may represent a delayed and, hence, therapeutically targetable injury component. Furthermore, recent data showed that Nec-1 also inhibits necrotic cell death provoked in response to other pathologic stimuli in cellular assays, which includes high doses of glutamate (179), plant sterols (180), and the chemotherapeutic agent shikonin (180). These results suggest that importance of necroptosis likely extends beyond DR signaling, and it may represent a major novel component of acute pathologic injuries.

Mitochondrial PTP

The mitochondrial permeability transition (MPT) is the loss of the inner mitochondrial membrane impermeability to solutes caused by opening of the MPT pore (MPTP). In turn, this action results in a loss of mitochondrial function and provides a common mechanism implicated in activation of mitophagy/autophagy, apoptosis, and necrosis in different cell systems. Although the composition of MPTP is not fully settled, multiple studies suggest involvement of adenine nucleotide translocase (ANT) in the inner mitochondrial membrane, voltage-dependent anion channel (VDAC or porin) in the outer membrane, and cyclophilin D (CypD) in the matrix.

Involvement of other proteins such as benzodiazepine receptor, hexokinase, creatine kinase, and Bax has also been proposed (181–183). The first identified potent inhibitor of MPT is the cyclosporine A (CsA) (184) (Fig. 5t), which inhibits the interaction between CypD and ANT (180). Using isolated mitochondria, CsA was shown to inhibit MPT at submicromolar concentrations (185, 186). CsA is also a potent inhibitor of necrotic death in the cells, for example, induced by oxidative stress and in vivo, notably in models of ischemia/reperfusion injury of liver (187), brain and central nervous system (188, 189), and myocardium (190). This finding suggests that inhibition of MPT may be a promising general direction for treating ischemia/reperfusion injury. This notion is supported by the resistance of observed in CypD-deficient mice to this form of injury (191, 192). The major drawbacks of therapeutic use of CsA are its transient and incomplete PTP inhibition as well as immunosuppressive side effects (193). In addition to CsA, other cytoprotective agents were also found to act as PTP inhibitors. These agents include cyclosporin A, rapamycin, and promethazine (Fig. 5u, v), with both of the latter exerting neuroprotective effects in vivo (194, 195). However, in the case of these agents, the exact mechanism of the cytoprotective effects in vivo is yet to be established.

Conclusion

After the groundbreaking discoveries establishing that the cell death can result from intrinsic cellular regulation, rather than excessive external stress, significant focus has been placed on characterization of the molecular mechanism of this process. Several interesting targets, some of which are described in our review, have been identified, and multiple classes of small molecules have been developed and optimized successfully for use as the research tools and, in particular cases, even as the drug development candidates for various disorders, for which efficient treatments are currently unavailable. Compounds that target the enzymatic activities of cell death regulators, such as caspases, PARP, and HtrA2 inhibitors, as well as compounds that modulate the protein–protein interactions and protein conformational changes were developed successfully. Inhibitors have been identified using a variety of different screening approaches from tethering and SAR-by-NMR to rational drug design and high-throughput screening using enzymatic assays or whole-cell phenotypic readouts. Therefore, chemical inhibitors of cell death represent a very extensive and diverse small-molecule development effort, which attests to the importance of cell death regulatory pathways. In some instances, for example, of necroptosis inhibitors, small molecules provided very useful means for uncovering functional significance of the regulation of novel modes of cell death, even in the absence of the specific protein target information. In the other cases, for example, of the PARP inhibitors, small molecules proved very useful in showcasing the importance of the activity of their targets in the cells and in vivo, prompting additional analysis of the biological regulation associated with these protein factors. Finally, in the case of well-validated targets, such as caspases or Bcl-2 family, specific
small-molecule inhibitors confirmed validity of targeting these regulators and the apoptotic pathway in general for the treatment of human disease. They provide exciting new lead compounds for therapeutic development. Overall, these findings emphasize that the availability of tools and methods for fast and efficient identification of small molecule modulators of important biological processes, stemming from rapid development of synthetic chemistry and small molecule screening technologies, provide a very valuable compliment to biologic research.

References


Cell Death: Biological Mechanisms and Small Molecule Inhibitors of PARP-1


Cell Death: Biological Mechanisms and Small Molecule Inhibitors of

See Also

Kinases, selective and Nonselective Inhibitors of Mitochondria: Topics in Chemical Biology
Cysteine Proteases and Cysteine Protease Inhibitors
Protein-Protein Interactions, Tools to Study
DNA Damage, Sensing of
Cell Division, Small Molecules to Study

Michael A. Lampson and Tarun M. Kapoor,
Rockefeller University, New York, New York
doi: 10.1002/9780470048672.wecb060

Cell division is the process by which a cell creates two genetically identical daughter cells. To maintain genomic integrity, a complex and highly regulated sequence of events ensures that the replicated chromosomes are partitioned equally between the daughter cells. For more than 50 years, strategies designed around small-molecule inhibitors have played a critical role in advancing our understanding of this essential process. Here we introduce a series of questions in the biology of cell division and illustrate how small molecules have been used to design experiments to address these questions. Because of the highly dynamic nature of cell division, the temporal control over protein function that is possible with small molecules has been valuable particularly in dissecting biologic mechanisms.

Biology of Cell Division

Cell division is the process by which a cell creates two genetically identical daughter cells. Each chromosome is replicated before mitosis begins, and after the division, the daughter cells inherit exactly one copy of each chromosome. A complex series of events have been divided broadly into two processes: mitosis, in which the identical sister chromosomes are separated and transported to opposite ends of the cell, and cytokinesis, in which the cell physically divides to create two daughter cells (Fig. 1). Preservation of genome integrity requires that the cell divide only after accurate segregation of sister chromosomes in mitosis. Failure in either chromosome segregation or in the timing of critical events in cell division leads to the loss or gain of whole chromosomes in the daughter cells, which is a condition known as aneuploidy that is strongly associated with developmental defects and human diseases such as cancer (reviewed in Reference 1). For over a century, cell division research has focused on mechanisms that physically segregate the chromosomes in mitosis and that divide the cell in cytokinesis, as well as those that control both processes in space and time. Some fundamental questions are how forces are generated to move chromosomes and to cleave the cell, how chromosome movements and cell cleavage are coordinated in space to ensure that each daughter cell inherits exactly one copy of each chromosome, and how events are coordinated in time so that chromosome segregation is complete before cell cleavage. Progress toward addressing these questions has integrated a structural picture of the machinery of cell division with an understanding of the molecular mechanisms that regulate that machinery.

Examining Cell Division with Small-Molecule Inhibitors

Our understanding of biologic processes such as cell division often develops from discovering or designing ways to perturb the process and observe the effects of the perturbation. Although genetic approaches have been used widely for this purpose, small-molecule inhibitors offer some distinct advantages. Small molecules provide a high degree of temporal control over protein function, generally acting within minutes or even seconds, and often are reversible, allowing both rapid inhibition and activation. The ability to design perturbations on short time scales has proven valuable particularly in examining dynamic biologic processes. All cell division takes place in approximately 1 hour, with many events on second-to-minute time scales, and several small-molecule inhibitors (Table 1) have played an integral role in dissecting the mechanisms of mitosis and cytokinesis.

The nature of spindle fibers

Progress through mitosis is linked closely to chromosome movements (Fig. 3a). Replicated chromosome pairs first move to the center of the cell. After all chromosomes are positioned correctly at metaphase (Fig. 3a III), the sister chromosomes split...
Figure 1 Overview of mitosis. (a) (i) Chromosomes are replicated before mitosis, and sister chromosomes are held together. (ii) The spindle forms and chromosomes attach to spindle fibers. (iii) Chromosomes move to the center of the spindle at metaphase. (iv) Sister chromosomes separate at anaphase and move in opposite directions. (v) The cell divides as the cleavage furrow forms between the separated chromosomes. (vi) Two daughter cells form, each with exactly one copy of each chromosome. (b) Structure of colchicine, a small molecule that targets microtubules.

Table 1 Small-molecule inhibitors used to study cell division

<table>
<thead>
<tr>
<th>Small molecule</th>
<th>Source</th>
<th>Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colchicine</td>
<td>Natural product from meadow saffron (genus Colchicum).</td>
<td>Depolymerizes microtubules.</td>
</tr>
<tr>
<td>Benomyl</td>
<td>Synthetic, used as agricultural fungicide.</td>
<td>Depolymerizes microtubules, effective in yeast.</td>
</tr>
<tr>
<td>Nocodazole</td>
<td>Synthetic, identified as antihelminthic compound (47).</td>
<td>Depolymerizes microtubules.</td>
</tr>
<tr>
<td>Taxol</td>
<td>Natural product from bark of the yew tree (genus Taxus), identified for antifungal and antileukemic activity (56).</td>
<td>Inhibits microtubule dynamics.</td>
</tr>
<tr>
<td>Monastrol</td>
<td>Synthetic, identified as cell-division inhibitor (24).</td>
<td>Inhibits kinesin-5 motor protein.</td>
</tr>
<tr>
<td>Hesperadin</td>
<td>Synthetic, identified as cell-division inhibitor (36).</td>
<td>Inhibits Aurora kinases.</td>
</tr>
<tr>
<td>AKI-1</td>
<td>Synthetic, identified as kinase inhibitor (57).</td>
<td>Inhibits Aurora kinases.</td>
</tr>
<tr>
<td>Cytochalasin B</td>
<td>Natural product from the fungus Helminthosporium dematoidium (58).</td>
<td>Depolymerizes actin filaments.</td>
</tr>
</tbody>
</table>

apart at anaphase (Fig. 1a iv) and move to opposite sides of the cell before the cell divides into two daughter cells (Fig. 1a v, vi). Because of these coordinated movements, each daughter cell receives exactly one copy of each replicated chromosome. Correlated with chromosome movements is the appearance of a fibrous structure known as the mitotic spindle, initially observed in fixed samples, which forms at mitosis and disappears after the chromosomes have separated. One great challenge in the study of cell division has been to understand the organization and function of the mitotic spindle, particularly in relation to chromosomes movements. The physical properties of the spindle fibers and how they might drive chromosome movements, as well as their molecular components, have been understood in part through use of the small molecule colchicine (Fig. 1b). Exploiting the fact that the spindle fibers are optically anisotropic, or birefringent, with different indices of refraction in different directions (i.e., parallel or perpendicular to the fiber axis), Inoue (2) developed a sensitive polarized light microscope
to visualize directly the fibers in a living cell (2). He used this method to examine the behavior of the fibers after exposure to colchicines, a small molecule that was known to disrupt spindle function. Colchicine eliminated the birefringence over a time course that ranged from a few minutes to an hour depending on the concentration, which indicates a loss of the fibers (3). After the removal of colchicine, the fibers recovered to their original state. The inhibition of protein synthesis during the recovery demonstrated that the fibers were assembled from an available pool of material rather than by synthesis of new proteins (4). These findings, together with similar results obtained by chilling cells to eliminate the fibers (5), suggested that the spindle fibers consist of oriented polymers (hence the birefringence) in equilibrium with free molecules in solution. Treatment with colchicine or chilling pushes the equilibrium toward the depolymerized state, whereas the removal of colchicine or rewarming allows the system to return to its original state.

The same experimental paradigm, which combines the observation of spindle fibers with the perturbation of their function in living cells, was used to demonstrate the potential functional significance of the spindle fiber dynamics. When spindles were treated with low concentrations of colchicine, which did not eliminate the birefringence immediately, the fibers slowly contracted and pulled the chromosomes toward one side of the spindle, which was anchored at the cell surface. After the removal of colchicine, the fibers elongated as chromosomes moved away from the pole (3, 6). This finding demonstrated that polymerization and depolymerization of the fibers could generate force by coupling to chromosome movement. These experiments exploited both the reversibility of colchicine and the ability to manipulate the behavior of the fibers by controlling the compound concentration, both properties of many small molecules.

In the Inoue studies, colchicine was used as an experimental tool to probe spindle function, but its target and mechanism of action were unknown. As 100 nM colchicine was sufficient to arrest cultured cells in mitosis, the implied tight binding suggested that the inhibitor might be used to isolate a colchicine–protein complex. A breakthrough came when colchicine was labeled with ^3H with high specific activity to demonstrate reversible binding to cellular sites (7). When the labeled colchicine was tested with a variety of cells, tissues, and organelles, high-binding activity was observed with the mitotic spindle, cilia, sperm tails, and brain tissue (8, 9). A common feature of these preparations is that all are enriched in intracellular fibers called microtubules, the same fibers observed by Inoue, which suggested that the target of colchicine was a subunit of microtubules. The colchicine-binding activity was extracted by dissolving isolated sea urchin sperm tails, purified by gel filtration and sedimented over a sucrose gradient to identify a single component 10. The same component was isolated from mammalian brain, shown to bind GTP, and biochemically characterized as a GTP-binding protein (11). The protein was named tubulin because it was believed to be the primary constituent of microtubules (12).

It has been known for over 100 years that treatment with colchicine arrests cells in mitosis (reviewed in Reference 13). Other small molecules since then have been identified that also block mitosis by targeting microtubules. The potential of these compounds as cancer therapeutics was demonstrated by the vinca alkaloids, such as vincristine and vinblastine, which have been used in the clinic for 40 years. At high concentrations (10–100 nM), these compounds depolymerize microtubules, which eliminates the mitotic spindle. At lower concentrations that are used clinically, microtubules remain stable but microtubule dynamics are suppressed. A number of small molecules that inhibit microtubule dynamics, taxol, also arrests cells in mitosis and is used widely to treat a variety of cancers (reviewed in Reference 14). The mitotic arrest induced by these drugs eventually leads to cell death (15) through mechanisms that are only beginning to be understood (16–18).

**Progression Through Mitosis**

It is clear from observing chromosome movements that cell division occurs in an ordered sequence of events (Fig. 2a). First chromosomes attach to spindle microtubule fibers and move to the spindle equator. Only after completion of this step do sister chromosomes separate at anaphase and move to opposite sides of the cell, followed by division into two daughter cells. Events must occur in this order for successful chromosome segregation. If the cell enters anaphase prematurely, before chromosomes have attached properly to the spindle, the sister chromosomes will not segregate equally, which leads to aneuploid daughter cells. Therefore, mechanisms that determine the timing of anaphase onset are critical for the success of mitosis.

One hypothesis for how anaphase onset might be regulated is that it could either be washed out, as the effect is reversible, or used at a low dose, so that cells would survive the treatment. With high benomyl (70 μg/ml), which prevents spindle formation, cells arrested in mitosis (Fig. 2a). After the removal of benomyl, cells recovered, proceeded normally through mitosis, and continued to grow (Fig. 2b) (21). Alternatively, with low benomyl (15 μg/ml), spindle assembly was slowed and mitotic progression through the cell cycle was blocked as cells arrested in mitosis (Fig. 2c). However, mitotic arrest was only transient at these drug concentrations, and cells eventually continued through mitosis. This finding demonstrated that feedback control mechanisms make essential contributions to the regulation of mitosis.
Benomyl removed

(a) Wild-type cell arrests in mitosis.

(b) Mutant cell defective in feedback control fails to arrest.

(c) (d) Cells dead due to catastrophic chromosome mis segregation.

Wild type cell arrests in mitosis.

Mutant cell defective in feedback control fails to arrest.

Benomyl

Benomyl removed

Benomyl

Genes identified in the benomyl screens, which generally are well conserved from yeast to mammals. The importance of several of these genes for faithful chromosome segregation has been confirmed in transgenic mice, in which reduced expression increases both aneuploidy and cancer susceptibility, and human tumor cells have been reported to carry mutations in Mad1, Mad2, Bub1, and BubR1, a related vertebrate protein (reviewed in Reference 1). Additionally, human germ-line mutations in BubR1 have been linked to mosaic variegated aneuploidy, a condition associated with a high risk of cancer (23).

Benomyl was used in the mad and bub screens not because of its specific protein target but because of the perturbation of spindle assembly. In principle, the same experiments could have been done without knowing the protein target or by targeting a different component of the spindle. The generality of the spindle checkpoint has been demonstrated through the use of monastrol, a small-molecule inhibitor of the mitotic kinesin Eg5, which was identified in a screen for small molecules that arrest cells in mitosis without targeting tubulin (24). Monastrol treatment arrests cells in mitosis with monopolar spindles because Eg5 is required to separate the spindle poles. The inhibition of
Mad2 by the microinjection of inhibitory antibodies overrides the checkpoint so that cells enter anaphase in the presence of monastrol with monopolar spindles (25). This finding indicates that the principle of feedback control generally applies to spindle perturbations through highly conserved mechanisms. Inhibitors of Eg5 currently are in development as anti-cancer drugs because, like taxol and the vinca alkaloids, they arrest cells in mitosis by activating the spindle checkpoint. Recent studies have demonstrated that the efficacy of drugs targeting either Eg5 or microtubules requires a prolonged, checkpoint-dependent mitotic arrest (17, 26). A compromised spindle checkpoint, for example through reduced expression of Mad2, confers resistance to these drugs.

Primary signals for checkpoint activation: attachment or tension?

Whereas the benomyl screens established the existence of the spindle checkpoint and identified some of the key components in checkpoint signaling, a fundamental question that remained unanswered was what exactly is monitored by the checkpoint. Two general models have been proposed. One is that the checkpoint monitors the attachment of spindle microtubules at the kinetochore, a structure that forms on each chromosome to mediate microtubule binding. Unattached kinetochores keep the checkpoint active and delay anaphase (27). A second model is that the checkpoint monitors force across the centromere, the region of the chromosome where kinetochores assemble (28). When both sister kinetochores are attached correctly, they are pulled in opposite directions by the microtubule fibers and the centromere is under tension (Fig. 3a). In this model, the absence of centromere tension would keep the checkpoint active. Small molecules that target tubulin have provided a way to test these models experimentally. Nocodazole depolymerizes microtubules, which creates unattached kinetochores (Fig. 3c), whereas taxol stabilizes microtubules but inhibits their dynamics, which decreases the centromere tension (Fig. 3b) (29).

To determine the effects of these microtubule perturbations on spindle checkpoint signaling, intracellular localization of Mad2 was examined. At early stages of mitosis (Fig. 3a i, ii), Mad2 localizes to kinetochores. As cells progress through mitosis, however, Mad2 disappears from kinetochores, and at anaphase onset (Fig. 3a iv) none of the kinetochores have detectable Mad2. These findings suggest that the presence of Mad2 on kinetochores serves as a signal to delay anaphase (29-31).

When microtubules are depolymerized with nocodazole (Fig. 3c), Mad2 localizes to all kinetochores, which indicates the activation of the checkpoint. If microtubule dynamics are suppressed with taxol while maintaining kinetochore attachments (Fig. 3b), Mad2 localizes to only a few kinetochores (29). This finding suggests that checkpoint signaling, as determined by Mad2 localization, is sensitive to attachment but does not respond directly to centromere tension. The interpretation of these experiments is complicated, however, because tension required for kinetochores to bind the full complement of microtubules; loss of tension may activate the checkpoint indirectly (32).

Experiments in yeast suggested that a member the Auroa kinase family, p110, is required to activate the spindle checkpoint in response to loss of tension but not loss of microtubule attachments (33). Understanding the function of Auroa kinases is particularly important because they have been linked to oncogenesis, and Auroa kinase inhibitors currently are in development as cancer therapeutics (34, 35). In mammalian cells, the inhibition of Auroa kinase activity with small-molecule inhibitors has been shown to bypass the mitotic arrest induced by taxol but not that induced by nocodazole, which is consistent with the idea that the kinase is required specifically in a tension-sensitive mechanism of checkpoint activation (36, 37). The interpretation of these results is complicated, however, because Auroa kinases are implicated also in regulating kinetochore-microtubule binding (38, 39). An alternative interpretation is that Auroa kinase inhibition overrides the taxol-induced arrest through effects on kinetochore-microtubule attachments.

Correcting errors in chromosome–spindle attachments

Feedback control of anaphase onset, or mitotic checkpoint signaling, is one mechanism that contributes to ensuring accurate chromosome segregation. Delaying anaphase in response to unattached kinetochores, however, is not sufficient. Chromosomes must attach to spindle microtubules in a particular...
orientation. For each replicated chromosome pair, the sister kinetochores attach to opposite poles of the spindle so that when sister chromosomes separate at anaphase, they are pulled to opposite sides of the cell. Other attachment states can occur, for example if both sister kinetochores are attached to the same spindle pole or a single kinetochore is attached to both poles. If these errors are not corrected, sister chromosomes will not segregate properly at anaphase (40).

Error correction is thought to occur by stabilizing correct attachments while destabilizing incorrect attachments (41). Experiments in yeast showed that the inhibition of the Ipl1/Aurora family of kinases prevents error correction by stabilizing incorrect attachments while destabilizing incorrect attachments (38, 42), but how the active kinase corrected attachment errors was not known. This problem was particularly difficult to address because attachment errors are observed infrequently in the presence of active Aurora kinase (43). Experimental approaches that accumulated attachment errors through inhibition of Aurora kinase, for example by genetic mutation (42), did not permit subsequent kinase activation to examine error correction. Reversible small-molecule Aurora kinase inhibitors present a solution to this problem because they can be used to inhibit kinase function and subsequently removed to activate the kinase.

To devise a strategy to address the question of how attachment errors were corrected, several issues needed to be addressed. First, Aurora kinases have been implicated in multiple processes in mitosis (44). Ideally, kinase inhibition temporarily would be controlled to isolate experimentally the error correction process. Second, the microtubules fibers that attach chromosomes to the spindle are highly dynamic, and the error correction likely involves some regulation of these dynamics. Live imaging would permit the analysis of microtubule dynamics with high temporal and spatial resolution. Finally, analysis of microtubule dynamics is difficult if individual fibers are obscured by other microtubules in the spindle. By creating conditions in which the improperly attached chromosomes are positioned away from the spindle body, individual fibers could be observed clearly.

An assay was developed, using several small-molecule inhibitors, in which all of these issues were addressed (Fig 4a). First cells were arrested in mitosis with monopolar spindles using the Eg5 inhibitor monastrol (Fig 4a). In the monopolar spindles many chromosomes have a particular attachment error in which both sisters are attached to the single spindle pole, referred to as synthetic attachment (46). After the removal of monastrol, the spindle becomes bipolar, all attachment errors are corrected, and the chromosomes segregate normally at anaphase. To test if Aurora kinase activity is required for correction of the attachment errors, an Aurora kinase inhibitor was added immediately after the removal of monastrol (Fig 4a ii). The advantage of adding the Aurora kinase inhibitor at this point is that Aurora kinase activity is unperturbed for all the preceding stages of mitosis. This assay was performed with two structurally unrelated Aurora kinase inhibitors, A-K1-3 and hesperadin, to control for possible off-target activities of the small molecules (Fig 4b).

Using cells that express GFP-labeled tubulin, both chromosome and microtubule dynamics were examined at high resolution by multimode fluorescence and transmitted light microscopy during spindle bipolarization in the presence of the Aurora kinase inhibitor. The syntelic attachment errors persisted as the spindle bipolarized, which directly demonstrates that Aurora kinase activity is required for the correction of these errors. Notably, some of the improperly attached chromosomes were positioned away from the spindle body, which allowed clear observation of the attached microtubule fiber, unobstructed by other spindle microtubules (Fig 4c). At this point, the Aurora kinase inhibitor was removed to examine how the active kinase might correct the syntelic attachment errors (Fig 4a iii, iv). Aurora kinase was shown to be fully active in vivo 30–60 minutes after removing the inhibitor, as determined by measuring the phosphorylation of histone H3, a known Aurora kinase substrate. On the time scale of kinase activation, improperly attached chromosomes remained attached to the microtubule fibers and were pulled to the spindle pole as the fibers shortened (Fig 4b). After disassembly of the microtubule fibers, the chromosomes moved to their usual position at the center of the spindle as correct attachments formed. Properly attached chromosomes were not affected, which suggests the local regulation of microtubule dynamics by Aurora kinase activity.

This assay demonstrates some advantages of small-molecule inhibitors, particularly in combination with high-resolution live-cell microscopy. Mitosis is a highly dynamic process with many events occurring on time scales of minutes or seconds. Ideally, an experiment would allow both perturbation of protein function and observation of the effects of the perturbation on similar time scales. The use of reversible small-molecule inhibitors to manipulate protein function, together with live cell imaging, makes this possible. In the assay described here, the inhibitors were used effectively as switches to turn enzymes on and off for both the kinases Eg5 and Aurora kinases. With this high degree of temporal control, a mechanism for correcting chromosome attachment errors could be dissected without perturbing the preceding processes, such as those involved in spindle assembly.

Force generation in cytokinesis

Following inactivation of the checkpoint, the chromosomes separate at anaphase and are pulled in opposite directions by the attached microtubule fibers. At anaphase, the cell physically divides into two daughter cells in cytokinesis. Several key questions in cytokinesis have been addressed by using small-molecule inhibitors. First, how is the force generated to cleave the cell into two parts? Second, what mechanisms determine the position of the cleavage plane? Third, how is the timing of progression through cytokinesis controlled?

Cleavage of a cell into two equal parts is a dramatic event that requires coordinated generation of force around the entire perimeter of the cell (Fig 5a). As a clue to how this might be achieved, a filamentous structure was shown by electron microscopy to lie at the cleavage furrow, just below the plasma membrane (reviewed in Reference 47). These filaments, distinct from microtubules, were called microfilaments. A key
Figure 4  Correction of improper chromosome attachments by activation of Aurora kinase (45). (a) Assay schematic. (i) Treatment with the Eg5 inhibitor monastrol arrests cells in mitosis with monopolar spindles, in which sister chromosomes often are both attached to the single spindle pole. (ii) Hesperadin, an Aurora kinase inhibitor, is added as monastrol is removed. As the spindle bipolarizes with Aurora kinase inhibited, attachment errors fail to correct so that some sister chromosomes are still attached to the same pole of the bipolar spindle. (iii) Removal of hesperadin activates Aurora kinase. Incorrect attachments are destabilized by disassembling the microtubule fibers, which pulls the chromosomes to the pole, whereas correct attachments are stable. (iv) Chromosomes move from the pole to the center of the spindle as correct attachments form. (b) Structures of the Eg5 inhibitor monastrol and two Aurora kinase inhibitors, hesperadin and AKI-1. (c) Spindles were fixed after bipolarization either in the absence (i) or presence (ii) of an Aurora kinase inhibitor. Arrows indicate sister chromosomes that are both attached to the same spindle pole. Projections of multiple image planes are shown, with optical sections of boxed regions (1 and 2) to highlight attachment errors. Scale bars 5 µm. (d) After the removal of hesperadin, GFP-tubulin (top) and chromosomes (bottom) were imaged live by three-dimensional confocal fluorescence microscopy and DIC, respectively. Arrows and arrowhead show two chromosomes that move to the pole (marked by circle in DIC images) as the associated kinetochore-microtubule fibers shorten and that then move to the center of the spindle. Time (minutes:seconds) after the removal of hesperadin. Scale bar 5 µm.
step in understanding the function of microfilaments in cytokinesis and other processes was to observe a correlation between the presence of the filaments, their disruption by the small molecule cytochalasin (Fig. 5a), and the phenotype of cytochalasin treatment. Cytochalasin eliminated the microfilaments at the cleavage furrow and prevented contraction of the furrow at cytokinesis. Cytochalasin also inhibited several other forms of cellular or intracellular force generation, including cell motility, membrane ruffling, and nerve outgrowth (48, 49). Microfilaments were observed in all of these systems, and in every case the microfilaments were disrupted by cytochalasin and returned to their normal state as cells recovered after the removal of cytochalasin. Furthermore, the actions of cytochalasin and colchicine generally were mutually exclusive: Processes dependent on microtubules and therefore inhibited by colchicine were often insensitive to cytochalasin, whereas those inhibited by cytochalasin were generally insensitive to colchicine (49). These observations suggested that the two types of filamentous structures could function independently in the cell. Although the molecular target of cytochalasin was still unknown, the correlative evidence indicated that microfilaments played a fundamental role in the generation of forces at the cellular level.

The evidence seems overwhelming that microfilaments are the contractile machinery of nonmuscle cells" (49). Contractility in muscle was achieved through the action of the myosin motor, which uses energy from ATP hydrolysis to slide filaments made up of polymers of the protein actin. The relevance of this process to other cell types had not been demonstrated. A direct link between cytochalasin and actin was provided by the demonstration that cytochalasin decreases the viscosity of actin filaments purified from muscle (50). This experiment led to two important conclusions. First, cytochalasin interacts directly with actin. Second, "an interaction of cytochalasin with actin or actin-like proteins in vivo could account for the ability of cytochalasin to inhibit various forms of cell motility and contraction" (50). Thus, actin was shown to be the molecular target of cytochalasin and implicated as a critical component of the microfilaments involved in cytochalasin-sensitive processes, including contraction of the cleavage furrow at cytokinesis.

Spatial and temporal control of cytokinesis

The cleavage plane typically is positioned in the center of the cell so that cellular components are divided equally between the two daughter cells. Asymmetric divisions do occur, however, and are particularly important during development, when the location of the cleavage plane can determine the fate of the daughter cells. Models to explain the position of the cleavage plane relied on the presence of the bipolar microtubule array of the mitotic spindle, which would place the division plane in between the spindle poles. To test this idea directly, an experiment was designed using monastrol to determine if cytokinesis could occur in cells with monopolar spindles (25). Because the spindle checkpoint prevents anaphase onset in monastrol-arrested cells, inhibitory antibodies against Mad2 were microinjected to override the checkpoint. The microinjected cells entered anaphase and successfully completed cytokinesis (Fig. 6). This experiment demonstrated that a bipolar microtubule array is not required for cytokinesis. Furthermore, careful analysis of microtubule dynamics showed that a population of microtubules near the chromosomes was stabilized during anaphase in the monopolar spindles at the location where the cleavage plane formed. These findings suggested a model in which microtubule dynamics are regulated through association with chromosomes to determine the position of the division plane.

The monastrol experiment showed that the positioning of the cleavage plane is correlated with the position of a particular population of microtubules. How microtubules generate a signal to recruit components of the contractile machinery remained an outstanding question. It has been difficult to isolate experimentally the molecular events that occur in the short time between anaphase onset and the beginning of cytokinesis. To address this problem, a strategy was devised to arrest cytokinesis at a defined point, before contraction of the cleavage furrow (51). A small-molecule inhibitor of the ATPase activity of nonmuscle myosin II, the actin-based motor that generates the force to contract the cleavage furrow, was identified in a high-throughput screen. This inhibitor, called blebbistatin because it prevents myosin II-dependent membrane blebbing, blocks cytokinesis
This experiment demonstrated that the duration of cytokinesis myosin II at the cleavage furrow, increased substantially (51) the blebbistatin arrest. In the presence of MG132, the time that small-molecule proteasome inhibitor MG132 was added during the blebbistatin arrest is that their effects on a single process could be isolated without affecting the preceding processes. These experiments showed that signals from both Aurora and Rho kinases are required to localize myosin II to the cleavage furrow.

To investigate the timing of cytokinesis, small molecules have been used in several strategies to perturb the cleavage process. If cytokinesis is prevented by perturbing either actin or microtubules with cytochalasin or nocodazole, or by inhibiting myosin II with blebbistatin, a window of approximately 1 hour exists during which cytokinesis can occur if the inhibitor is removed (51–53). The existence of this window suggests that an irreversible step exists that prevents cells from reversing processes. Such irreversible steps can be mediated by the degradation of key proteins through proteolysis. To test whether proteolysis is required for the irreversible exit from cytokinesis, the small-molecule proteasome inhibitor MG132 was added during the blebbistatin arrest. In the presence of MG132, the time that cells remained in cytokinesis, as determined by the presence of myosin II at the cleavage furrow, increased substantially (51). This experiment demonstrated that the duration of cytokinesis is determined by ubiquitin-mediated proteolysis.

**Conclusion**

The experiments described in this review illustrate how small-molecule inhibitors have been used to design strategies to address fundamental questions in cell division. As our understanding of cell division advances, the use of small molecules should complement genetic and RNAi-based approaches. In particular, the temporal control over protein function that is possible with small molecules makes it possible to dissect the functions of proteins that are involved in multiple processes. Part of the complexity of cell division is that many proteins, for example the Aurora and Polo family kinases (44, 54), are implicated at multiple stages in both mitosis and cytokinesis. The perturbation of one stage often affects subsequent events, such as by activation of the spindle checkpoint, which limits analyses by preventing subsequent steps. Small-molecule inhibitors can be used to temporally isolate a specific process without perturbing the preceding events. These strategies will likely make important contributions to future investigations of cell division mechanisms.

It also is important to consider some of the limitations of small-molecule inhibitors, particularly in comparison to genetic approaches. With genetics, any gene can be targeted for mutation or deletion without directly affecting any other gene. The discovery of useful small-molecule inhibitors, however, is challenging. The specificity of small-molecule inhibitors also is difficult to demonstrate convincingly. Testing a kinase inhibitor against over 500 kinases in the human genome, for example, is a substantial undertaking. One way to address specificity is to use small molecules in focused assays, in which a narrowly defined biologic process is examined, so that off-target effects are unlikely to be relevant. In combination with this approach, the effects of different, chemically unrelated inhibitors that target the same protein could be compared, as the off-target activities are unlikely to be similar.

**Future directions**

The use of small-molecule inhibitors is limited by the availability of inhibitors and the assays that can be designed around them. The proteins that currently are known to be targeted by small molecules make up a small fraction of the proteome. The identification of new inhibitors will promote the application of small-molecule-based strategies to an increasing range of biologic problems. As methods are developed to monitor protein function with high temporal and spatial resolution, particularly in living cells, the scope for using small molecules also will increase. With recent advances in fluorescence-based probes, it has become possible to monitor numerous properties of living cells, including protease activity, posttranslational modifications, membrane potential, and pH, as well as mediators of intracellular signaling such as Ca2+ and cyclic AMP (55). The temporal control available with small-molecule inhibitors, combined with these high-resolution readouts, should be a powerful combination for examining dynamic biologic processes in living cells. In vitro methods also have been developed for measuring the enzymatic activities of single protein molecules. Observing the effects of a small-molecule inhibitor both at this level and in a more complex cellular context should provide new insights into protein function.

Both the identification of new inhibitors and the design of increasingly sophisticated assays to examine their effects also will contribute to the drug discovery process in several ways. First, studies with small-molecule inhibitors will advance our understanding of the effects of chemical inhibition, the mode of action of most drugs, which typically does not affect levels of the target protein. Approaches such as RNAi that act by

---

**Figure 6**

Assay to examine cytokinesis in the presence of a monopolar spindle (25). Treatment with monastrol, a small-molecule inhibitor of the kinesin Eg5, causes cells to arrest in mitosis with monopolar spindles because of activation of the spindle checkpoint. Microinjection of an antibody against the protein Mad2 inactivates the checkpoint so that cells divide with monopolar spindles.

---
reducing protein levels may have other effects, for example, preventing the formation of a multiprotein complex for which the depleted protein is required. Second, the discovery of new inhibitors may provide leads for drug development. For example, inhibitors that induce a mitotic arrest by targeting a protein that is specific to mitosis are potential leads for anticancer drugs that will have fewer side effects than tubulin poisons such as taxol and the vinca alkaloids. Third, assays that are designed to examine the effects of small molecules on specific cellular processes can be used to screen potential drugs for both on- and off-target activities. Finally, a deeper understanding of the basic molecular mechanisms of key processes should lead to improved therapeutic strategies.

References


Chromosomes are large molecules of DNA that organize genetic materials of an organism. Chromosomes need to be properly organized, packaged, and compacted to ensure their integrity and segregation to daughter cells during mitosis. In eukaryotic cells, chromosomes associate with histones and other proteins to form chromatin within the nucleus. The basic packaging unit of chromatin is the nucleosome. The nucleosome core particle consists of 147 base pairs of DNA that wrap around a core histone octamer. A linear array of nucleosomes is further folded into high-order chromatin structures. In addition, chromosomes are regulated during mitosis by the structural maintenance of chromosomes (SMC) family of proteins. Sister chromatids are physically tethered during DNA replication by the SMC1/3 cohesin complex to ensure equal segregation of chromosomes to daughter cells. These sister chromatids are compacted by the SMC2/4 condensin complex to form metaphase chromosomes.

DNA is the genetic material that encodes all information necessary for life. The DNA molecules that carry this information must be highly regulated to control and protect precisely the information they encode. In addition, cellular DNA needs to be replicated and divided equally to daughter cells during mitosis. For these reasons, DNA is elegantly compacted more than 100,000 times into a nucleoprotein structure known as chromatin. This compaction allows for the protection, regulation, and segregation of the genetic information encoded in DNA.

**Chromatin Structure**

Genomic DNA is organized into subunits called chromosomes (1). The human genome is split into 46 chromosomes. In eukaryotes, DNA is packaged and compacted with histones and other proteins to form chromatin. DNA wraps around core histones to form nucleosomes, the most fundamental level of DNA compaction (2). In addition to compacting DNA, histones also regulate the accessibility of DNA to various biological processes that use DNA as the template, including transcription, DNA replication, and DNA repair.

Nucleosome and chromatin fiber

The nucleosome core particle is the basic building block of chromatin (Fig. 1). It consists of 147 base pairs (bp) of DNA wrapped around the histone core in 1.7 left-handed, nonuniform superhelical turns (3). X-ray crystallography studies have revealed the structure of the nucleosome core particle to a resolution of 1.9 Å (4, 5). Each nucleosome core particle is made up of eight core histone molecules, one histone H3-H4 heterotetramer and two histone H2A-H2B heterodimers (4). Each core histone contains an amino-terminal unstructured tail and a three-helix histone fold (5). The histone folds in each dimer contain three distinct DNA-binding motifs that make extensive contacts with the minor groove of the DNA (4, 5). The helical periodicity or average number of base pairs per helical turn of DNA around the nucleosome core is 10.2bp (5, 6). The minor grooves of DNA wrapped around neighboring nucleosomes line up to form channels through which the flexible histone tails emerge from the nucleosome core particle (5). The major groove of nucleosomal DNA is exposed to allow recognition by nuclear proteins, such as transcription factors (5).

Nucleosomes are connected to one another by linker DNA of variable length and the linker-binding histone H1 protein (Fig. 1) (7). These long arrays of nucleosomes spontaneously condense to form helical arrays of nucleosomes, termed the 30-nm fiber after its apparent diameter (Fig. 1) (8). Additional condensation and compaction of chromatin occur through internucleosomal interactions. One important internucleosomal interaction required for chromatin fiber formation is the interaction of a highly acidic patch of histone H2A with the histone H4 tail (8). Ultimately, these internucleosomal interactions form interphase chromatin with an unknown architecture (Fig. 1) (9).
Histone tails

One striking characteristic of the nucleosome core particle is the protrusion of flexible amino-terminal tails from the core histone proteins (5). These histone tails make ideal targets for the regulation of the nucleosome core particle. During events such as transcription, the chromatin must be opened up to allow the extensive protein machinery to access DNA. A cetylation of lysine residues on the protruding core histone tails is one major posttranslational modification that promotes chromatin opening (10). Lysine acetylation of histone tails is associated with active transcriptional regions of the genome, whereas hypoacetylation correlates with transcriptional repression (11). Histone acetyl transferases (HATs) promote transcription by catalyzing acetylation of histone tails (11). Histone deacetylases (HDACs) mediate histone tail deacetylation and, therefore, repress transcription. In addition to acetylation, core histone tails are posttranslationally modified by a number of other moieties, such as serine/threonine phosphorylation, arginine/lysine methylation, lysine ubiquitination, lysine sumoylation, glutamic acid A DP ribosylation, arginine deimination, and proline isomerization (10). These modifications either promote or suppress access to chromatin during transcription, DNA repair, or chromatin condensation in unique ways (12). Recent work has identified several proteins that catalyze these histone modifications, such as the methyl transferases that methylate lysines and arginines, as well as the demethylases such as Jumonji C (jmc) domain-containing proteins and lysine specific demethylase 1 (LSD1) (13).

Chromatin assembly

Chromatin assembly is coupled with DNA replication. Immediately after DNA replication, the parental histones from the original chromosome are randomly divided between the two daughter strands of DNA (14). The remaining complements of histones are assembled from newly synthesized histones. Newly synthesized histones H3 and H4 are acetylated, and these acety groups are removed after their incorporation into DNA (15). A assembly of chromatin requires the functions of histone chaperones and ATP-dependent chromatin remodeling factors (16). Histone chaperones bind histones and facilitate their proper deposition onto DNA by preventing nonspecific histone-DNA interactions (17). Two major histone chaperones are CAF-1 and NAP-1. CAF-1 localizes to the replication fork by binding PCNA and facilitates the deposition of histones H3 and H4 onto the newly synthesized DNA strands (18, 19). Subsequently, NAP-1 facilitates the deposition of histones H2A and H2B to complete the nucleosome (20). Using in vitro nucleosome assembly and nuclease digestion mapping assays, it was shown that the periodic spacing of nucleosomes requires the function of ATP-dependent chromatin remodeling factors, such as the ACF/ISWI complex (16, 21).

Chromosome Cohesion

The propagation of genetic information during cell division is a complex and highly regulated process that involves extensive changes to the interphase chromatin. The somatic cell cycle is divided into four phases: G1, S, G2, and M. During S phase, DNA is replicated to form sister chromatids (22). Each sister chromatid contains the exact same genetic information. Therefore, it is imperative to ensure equal segregation of sister chromatids to daughter cells during mitosis (M phase). Unequal segregation of chromosomes results in daughter cells with wrong chromosome numbers, a condition known as aneuploidy. One important mechanism that ensures proper segregation of chromosomes to daughter cells is sister chromatid cohesion. The cohesin complex

Sister chromatid cohesion is primarily established by the cohesin complex (23). The cohesin complex is a multiprotein complex that contains the structural maintenance of chromosomes (SMC) family of proteins (24). SMC proteins contain the Walker A motif at their N-terminus and the Walker B motif at their C-terminus (Fig. 2a). These motifs are brought together by an intramolecular coiled-coil domain to form a functional ATPase domain, which is similar to other ATP-binding cas tases (ABC) ATPases, such as RAD50 (25). SMC proteins

Figure 1: Schematic drawing of how chromosomes are formed from nucleosomes to the mitotic chromosome. Double-stranded DNA (black lines) wraps around core histones (gray cylinders) to form the nucleosome. Histone tails (curved lines) protrude from the histone core and can undergo various posttranslational modifications that regulate chromatin state. Helical array of nucleosomes forms the 30-nm fiber. Scaffolding proteins (gray bars), such as condensin, further compact chromatin during mitosis to promote proper segregation to daughter cells.
Chromosome Formation

(a) Architecture of the SMC protein complexes. A functional ATPase head domain is formed by N-terminal Walker A and C-terminal Walker B motifs that are brought together by an intramolecular coiled coil. Two SMC proteins heterodimerize through their hinge domains. Kleisin proteins link the two head domains of the SMC heterodimer. (b) Architecture of the cohesin complex. The SCC1 kleisin protein links together the head domains of the SMC1–SMC3 heterodimer to form a ring. (c) Two models for how cohesin establishes functional sister-chromatid cohesion. Sister chromatids are shown as cylinders.

heterodimerize with one another through their hinge domains to form three unique complexes: cohesin (SMC1–SMC3), condensin (SMC2–SMC4), and SMC5–SMC6 (16, 19, 20). The head domains of the heterodimeric pairs of SMC proteins are linked by proteins known as kleisins (23). In addition to kleisins, SMC complexes contain additional non-SMC proteins. The cohesin complex is a multi-subunit complex that consists of the heterodimeric SMC1–SMC3 core, SCC1/MCD1/RAD21, and SA2 (SCC3 in yeast) (Fig. 2b) (23, 24). The function of SA2 is unknown. SCC1 is the kleisin subunit of the cohesin complex. SCC1 contains a winged helix domain at its C-terminus that mediates binding to the head domain of SMC1 (21). The binding of SCC1 to SMC1 facilitates the binding of SCC1’s N-terminal domain to the SMC3 head domain (26). In so doing, SCC1 links the two head domains of the SMC1–SMC3 heterodimer to form a ring-shaped structure. Cryoelectron microscopy has confirmed the ability of the cohesin complex to form a ring-shaped structure with an internal diameter of 30–40 nm (27).

Sister chromatid cohesion

The cohesin complex is loaded onto chromatin during late G1 phase of the cell cycle (28). The ATPase activity of SMC1–SMC3 is required for cohesin loading onto chromatin (29). The mechanism by which the ATPase activity promotes cohesin loading is unclear. In addition to the ATPase activity of SMC1–SMC3, cohesin loading is dependent on the SCC2–SCC4 complex (30). Both SCC2 and SCC4 contain Huntingtin, elongation A subunit, TOR (HEAT) repeats. The exact mechanism by which SCC2–SCC4 promotes cohesin loading is unclear. It has been proposed that SCC2–SCC4 stimulates the ATPase activity of SMC1–SMC3 (23, 29). The cohesin complex is concentrated around centromeres and at intergenic regions (31–33). Interestingly, in budding yeast, the chromosomal localization of SCC2–SCC4 does not overlap with that of cohesin, suggesting that cohesin redistributes to intergenic regions after SCC2–SCC4 facilitates its loading onto chromatin (27). The enrichment of cohesin at intergenic regions is thought to function in regulation of gene expression by insulating promoters from distant enhancers.

Loading of cohesin onto chromatin is itself insufficient to establish functional cohesion between sister chromatids. Establishment of sister chromatid cohesion occurs subsequent to cohesin loading and is tightly coupled to DNA replication in S phase. Several factors are required for the establishment of functional sister chromatid cohesion after the loading of cohesin onto chromatin. These factors include the acetyl transferase ECO1/CTF7, the replication factor C (RFC) complex (CTF18/CTF8), and the chromatin structure remodeling (RSC) complex (34–36). The precise mechanism by which chromatin-bound cohesin establishes sister chromatid cohesion is unresolved. Elegant studies have shown that cohesin binds chromatin by entrapping DNA within its ring-shaped structure. These studies inserted an artificial tobacco etch virus (TEV) protease cleavage site into
one cohesin subunit to allow inducible cleavage. Upon expression of TEV, the cohesin complex was rapidly removed from chromatin and sister chromatid cohesion was destroyed (27). In addition, using circular minichromosomes in which cohesin is loaded, cleavage of the circular DNA results in the removal of cohesin from these chromosomes (37). These studies show that DNA is entrapped within the cohesin ring. How DNA is encircled by the cohesin ring is unclear. Several possible mechanisms have been proposed, most of which involve the transient dissociation of proteins within the cohesin complex to open the ring structure. These include dissociation of the SMC1 and SMC3 hinge domains, SMC1 head domain and SCC1 C-terminus, or SMC3 head domain and SCC1 N-terminus. A further outstanding question is how cohesin actually mediates sister chromatid cohesion. Two equally viable models have been proposed (Fig. 2c). The embrace model proposes that a single cohesin complex encircles both sister chromatids within its ring-shaped structure as the DNA replication fork progresses (Fig. 2c, left). With a diameter of 30–40 nm, the cohesin ring is large enough to accommodate both sister chromatids (10 nm each) (32). However, it is unclear how the cohesin ring copes with the massive DNA replication machinery that would presumably need to move through the cohesin ring. Alternatively, a cohesin dimerization model proposes that each sister chromatid is trapped by separate cohesin complexes. In this model, the individually trapped sister chromatids are held together by the dimerization of the two cohesin complexes (Fig. 2c, right). However, intermolecular interaction between cohesin complexes has not been detected.

Mitotic Chromosome Condensation

In addition to cohesion between sister chromatids established during the S phase of the cell cycle, chromosomes are further condensed to allow for the movement and segregation of sister chromosomes to daughter cells during mitosis. The formation of stable, rod-like chromosomes during prometaphase of mitosis is mediated by several scaffolding proteins, such as topoisomerase II and the cohesin complex (38, 39).

The condensin complex

Chromosome condensation during mitosis is mediated by the condensin complex. Condensin has two subtypes: condensin I and condensin II (Fig. 3a) (23, 24). Both condensin I and condensin II contain the core SMC2–SMC4 heterodimer (34). Similar to SMC1 and SMC3, SMC2 and SMC4 dimerize through their hinge domains. An intramolecular coiled-coil domain brings together the Walker A and Walker B motifs at the N- and C-termini of SMC2 and SMC4. These functional AT-Pase heads are linked by kleisin proteins, CAP-H for condensin I and CAP-H2 for condensin II (24). The condensin complexes also contain additional non-SMC proteins. Condensin I contains CAP-D2 and CAP-G, whereas condensin II contains CAP-D3 and CAP-G2 (34). The functions of these non-SMC proteins are unclear. Electron microscopy studies showed that the condensin complex can adopt a ring-shaped structure similar to cohesin. Condensin can also adopt a lollipop-like structure, with the coiled-coil and hinge domains of SM C2–SM C4 projecting away from a protein mass that includes the head domains of SM C2–SM C4 and the non-SMC proteins (35). How or whether condensin I and condensin II vary in function is unknown. Both condensin I and condensin II are excluded from the nucleus during interphase. Condensin I associates on chromatin after nuclear envelope breakdown during mitosis, whereas condensin II associates with chromatin before nuclear envelope breakdown (40). By cytological examination, chromosome-bound condensin I and condensin II do not overlap (36). Depletion of condensin I or condensin II causes distinct abnormalities in chromosome morphology (41), suggesting that the two complexes may not be functionally redundant.

Chromosome condensation

The mechanism by which the condensin complex organizes chromosomes during mitosis into their fully compacted and stable structure is currently unresolved. Several studies in multiple organisms have established that inactivation of condensin does not result in completely decondensed chromosomes, but in compacted chromosomes that are not properly structured into rigid rods (42). The identity of those factors that mediate chromosome compaction in the absence of condensin is unknown. These
studies suggest that condensin mediates the organization of compacted chromosomes into higher order structures that give rise to familiar metaphase chromosomes (Fig. 1). Clues into how condensin organizes chromosomes come from its activities in vitro. The addition of DNA to purified Xenopus condensin stimulates its ATPase activity (43). Condensin promotes the positive supercoiling of closed circular DNA when in the presence of topoisomerase I and ATP, but not the nonhydrolyzable ATP analog, AMP-PNP (43). The supercoiling of DNA by condensin could promote mitotic chromosome formation, but whether this activity of condensin occurs in vivo is unclear.

The activity of condensin is cell cycle regulated. Condensin purified from mitotic extracts is much more active in promoting DNA supercoiling than condensin from interphase extracts (43). In yeast, cdc2 stimulates condensin activity by phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Furthermore, cdc2 is necessary for chromosome condensation (45). Likewise, inhibition of regulatory phosphorylation of condensin (44). Further...
Chromosome Formation


See Also
Nucleosome: Topics in Chemical Biology
Cell Cycle
Cell Cycle: Regulation of Topoisomerases, Chemistry of
DNA repair is the elimination of covalent DNA modifications and the correction of base mismatches. There are six basic repair categories: direct repair, base excision repair, nucleotide excision repair, recombination, cross-link repair, and mismatch repair.

**Direct Repair**

In direct repair, the chemical modification that constitutes the lesion is reversed without removing and replacing nucleotides. There are four known direct repair enzymes: photolyase, spore photoproduct lyase, methylguanine DNA methyltransferase, and AlkB family oxidative demethylases (1).

**Photolyase.** Photolyase harnesses blue light (400–500 nm) energy to repair ultraviolet (UV) (200–300 nm)-induced DNA damage. Photolyase is a 55–65-kDa flavoprotein that contains two noncovalently bound chromophores (2, 3). These chromophores are a flavin in the form of FADH− and a pterin in the form of methenyltetrahydrofolate (5,10-MTHF) in most photolyases and in rare cases 8-hydroxydeazariboflavin. There are two types of photolyases: cyclobutane photolyase, which repairs cyclobutane pyrimidine dimers (Pyr−→Pyr−), and 6–4 photolyase, which repairs pyrimidine 6-4 pyrimidone photoproducts. The structures and mechanisms of action of both types of enzymes are similar (2). They bind damaged DNA in a light-independent reaction and carry out catalysis by light-initiated cyclic electron transfer. In cyclobutane photolyase, FADH− transfers an electron to Pyr− to generate a radical pair, FADH− and Pyr−−Pyr−−. The cyclobutane radical undergoes bond rearrangement to yield two canonical pyrimidines concomitant with back electron transfer to regenerate the catalytically competent FADH−. Then the enzyme dissociates from the repaired DNA (Fig. 1). The (6–4) photolyase acts by a similar mechanism except that the (6–4) photoproduct is first converted to a four-membered oxetene ring thermally (1, 3) (Fig. 1). Although humans and other placental mammals do not possess photolyase, they do possess two proteins with a high degree of structural similarity to photolyase but no repair activity (1). These proteins, named cryptochrome 1 and 2, are essential for the regulation of the circadian clock (4, 5).

**Methylguanine DNA Methyltransferase.** Methylguanine DNA methyltransferase (MGMT), which repairs the O6-methylguanine and O6-methylthymine, is a 20–30-kDa protein that has no cofactors and is ubiquitous in nature (1). It recognizes DNA damage either by three-dimensional diffusion or, as recently claimed, by two-dimensional diffusion along the duplex with 3′ scanning bias, and it forms a low stability complex with the backbone of the DNA at the active site (7). It then flips out the O6-methylguanine base into its active site (7). The methyl group is then transferred to a nucleophilic cysteine residue in the active site in a SN2 manner (8). The repair reaction generates methyl-cysteine and inactivates the enzyme (Fig. 3) (9–11). The protein cannot be regenerated and is degraded after acting in the reaction. In the reaction the enzyme acts as an alkyl transfer reagent and not as a true enzyme (12). The MGMT encoded by the ada gene is largely responsible for the so-called “Adaptive Response” in Escherichia coli (13).

**AlkB Family Oxidative Demethylases.** These enzymes repair 3-methylcytosine and 1-methyladenine that are introduced by S9-type methylating agents. They use molecular oxygen, 2-oxoglutarate, and Fe(II) to oxidize the methyl group and release it in the form of formaldehyde. The enzyme is also capable of repairing RNA. It has been shown that the human homologues of AlkB, ABH2 and ABH3, prefer dsDNA and ssDNA, respectively, and that in the case of RNA, ABH2 prefers...
Figure 1: Reaction mechanism of DNA photolyases: (A) mechanism of cyclobutane photolyase and (B) mechanism of (6-4) photolyase. Both photolyases harness blue light energy to remove UV-induced damage and contain two noncovalently bound chromophores. They bind UV-damaged DNA in a reaction that is light independent and carry out catalysis in a light-initiated cyclic electron transfer. In (6–4) photolyase, the (6–4) photoproduct is converted to a four-membered oxetane ring thermally ($kT$) before the photochemical reaction.
double-stranded substrates, whereas ABH3 prefers single-stranded substrates. AikB is widespread in nature, and at least eight genes encoding the enzyme have been identified in humans (Fig. 4) (14).

Base Excision Repair

In base excision repair, the damaged nucleotide is removed in two steps. First, a glycosylase cleaves the glycosidic bond of the damaged base and releases it; then a combination of AP lyase/P endonucleases cleaves the phosphodiester bonds on either side of the resulting abasic site (AP = apurinic/apyrimidinic site) to release the deoxyribose. The resulting gap is filled in by a polymerase and sealed by DNA ligase. Base excision repair as a general rule works on non-bulky DNA lesions that do not cause gross structural alterations in the duplex (15).

Some glycosylases are simple glycosylases, which only catalyze the hydrolytic removal of a base to form an abasic site, whereas other glycosylases cleave off the base using a lyase mechanism and catalyze an AP lyase reaction. In this AP lyase reaction, a Schiff's base is formed from the 1′ aldehyde of the AP deoxyribose. This labilizes the 2′ hydrogen, resulting in a β-elimination reaction, and leaving a 5′-phosphomonoester and a 3′ unsaturated sugar phosphate residue. Lyase reactions are usually associated with glycosylases that remove oxidized bases, but not with glycosylases that remove alkylated or normal bases (15).

Depending on the initial events in the removal of the damage, in humans repair may proceed by two possible mechanisms: short patch repair or long patch repair. In short patch repair, the repair patch is a single base; in long patch repair, it is 2–10 nucleotides long. In the short patch monofunctional glycosylase pathway, APE1 makes a 5′ incision at an AP site and the 5′-baseless sugar phosphate is trimmed by the dRP lyase activity of Polβ to generate a ligatable 5′ end, which is ligated by the Ligase III/XRCC1 complex. Long patch repair occurs when hydrolytic glycosylases or spontaneous base loss generates the AP site. In long patch repair, APE1 cleaves the 5′ phosphodiester bond, and then repair synthesis and nick translation are carried out by DNA Polymerase δ or ε aided by RFC/PCNA clamp loader/polymerase clamp. DNA synthesis displaces several nucleotides. The flap that is generated is

Figure 2. Reaction mechanism of spore photoproduct lyase. Spore photoproduct lyase removes the UV damage spore photoproduct. The enzyme is an iron sulfur protein that uses AdoMet as a cofactor and repairs the damage through a radical mechanism.
Chemistry of DNA Damage Repair

Repair of O₆-methylguanine by methylguanine DNA methyltransferase. Methylguanine DNA methyltransferase has no cofactors and recognizes DNA cooperatively. It flips the O₆-methylguanine into its active site and transfers the methyl group to a nucleophilic cysteine in an SN₂ reaction.

The damage recognition mechanism of DNA glycosylases is similar to the mechanism of recognition of DNA photolyase. The initial recognition by DNA glycosylases is by diffusion. The distortions and H-bond changes that occur as a result of DNA damage are recognized with a low affinity/specificity. This recognition is followed by base-flipping, which forms a high affinity/specificity complex (16, 18). The glycosyl bond is then cleaved off by the Fen1 endonuclease and the repaired DNA is ligated by Ligase I (1, 16, 17) (Fig. 5).

The damage recognition mechanism of DNA glycosylases is similar to the mechanism of recognition of DNA photolyase. The initial recognition by DNA glycosylases is by diffusion. The distortions and H-bond changes that occur as a result of DNA damage are recognized with a low affinity/specificity. This recognition is followed by base-flipping, which forms a high affinity/specificity complex (16, 18). The glycosyl bond is then cleaved off by the Fen1 endonuclease and the repaired DNA is ligated by Ligase I (1, 16, 17) (Fig. 5).

Nucleotide Excision Repair

Nucleotide excision repair is the primary repair system for the removal of DNA damage caused by chemicals that produce bulky adducts or by the UV component of sunlight, which produces cyclobutane pyrimidine dimers (Py4<->Pyr) and (6-4) photoproducts in DNA (1). These lesions, as well as other bulky lesions induced by chemical carcinogens, are removed by the universally conserved repair mechanism called nucleotide excision repair. Nucleotide excision repair can be used to repair almost all single base lesions in DNA. As a result, it recognizes an extremely diverse substrate range, and rather than recognizing the specific damage itself, the excision nuclease tends to recognize the backbone conformations created by the damage (1).

The basic steps of nucleotide excision repair are as follows: 1) damage recognition, 2) dual incisions, 3) release of the excised oligonucleotide, 4) repair synthesis to fill in the gap, and 5) ligation. In E. coli, nucleotide excision repair is carried out by three proteins, UvrA, UvrB, and UvrC, which excise a 12-13 nucleotide-long oligomer containing the damaged base. A (UvrA)₂(UvrB) complex traces along the DNA backbone until UvrA recognizes a damaged base. The complex then binds to the damaged base. This binding activates the helicase function of UvrB, which unwinds the DNA around the damage (about 5 bp on either side of the damage). The resulting single-stranded DNA leads to a tighter binding of UvrB to the single-stranded DNA and a dissociation of the UvrA from the tight UvrB-DNA complex. Therefore, UvrA is referred to as the "Molecular Matchmaker" (1). The resulting UvrB-damaged DNA complex then recruits UvrC, which initiates the dual incisions of the DNA. The 3' incision is performed first at the fourth or fifth phosphodiester bond to the lesion by the GIY-YIC endonuclease domain of UvrC. In a concerted but asynchronous reaction, the second active site in the C-terminal half of UvrC...
Chemistry of DNA Damage Repair

Figure 4. Direct repair through oxidative demethylation by the AlkB protein. AlkB uses molecular oxygen, 2-oxoglutarate, and Fe(II) to oxidize the methyl group and then releases it as formaldehyde.

makes the 5′ incision. Helicase II (UvrD) releases the 12–13 nucleotide-long oligomer carrying the lesion along with UvrC. The dissociation of UvrC leaves a free hydroxyl group at the 5′ incision. DNA Pol I binds to this hydroxyl group and fills in the gap while releasing UvrB. DNA ligase then completes the repair reaction (Fig. 6) (20).

In humans, excision repair is carried out by six factors: RPA, XPA, XPC, TFIIH, XPG, and XPF-ERCC1, which are comprised of 15 proteins. These act in a coordinated manner to make dual incisions bracketing the lesion and remove the damage in 24–32 nucleotide-long oligomers (13). Damage is recognized by RPA, XPA, and XPC acting in a cooperative manner, followed by recruitment of TFIIH by XPA and XPC. The resulting complex is called pre-incision complex 1 (PIC1). Within this complex, the DNA is unwound by about 25 bp around the damage site by the XPB and XPD helicase subunits of TFIIH in a reaction that requires ATP hydrolysis. Then, XPC is displaced concurrent with binding of the XPG nuclease to form PIC2; finally XPF–ERCC1 is recruited to form PIC3. In PIC3, XPG makes the 3′ incision at the 6th ± 3 phosphodiester bond and the excised 24–32 nucleotide-oligomer is released and the complex dissociates. The excision gap is filled in by DNA polymerases and then ligated (21, 22) (Fig. 6).

Both the prokaryotic and the eukaryotic excision nucleases employ thermodynamic discrimination and kinetic proofreading to achieve high specificity. However, despite these elaborate discriminatory mechanisms, excision nucleases do not have absolute specificity and “gratuitous repair” of undamaged DNA occurs at a low but potentially deleterious level (23-25).

Finally, in addition to nucleotide excision repair and base excision repair, an alternative repair pathway exists to remove Pyr→Pyr and (6–4) photoproducts in some organisms, including Schizosaccharomyces pombe. In this alternative pathway, an enzyme called UV dimer endonuclease cleaves DNA immediately 5′ to the damaged lesion. The damage is then removed by a 5′ to 3′ exonuclease (26, 27). Additionally, AP endonucleases, like Nfo in E. coli and APE1 in humans, have the ability to incise immediately 5′ to nonbulky lesions caused by oxidative damage and initiate excision of the lesion by a 5′ to 3′ exonuclease (28).

Double-Strand Break Repair and Recombinational Repair

Double-strand breaks occur naturally in V(D)J recombination to generate antibodies and during the immunoglobulin class switching reaction, but they also occur improperly during replication as a result of the replication fork stalling and collapsing. Double-strand breaks may also be caused by ionizing radiation, chemicals, and reactive oxygen species. Homologous recombination (HR) and non-homologous end joining (NHEJ) are the two mechanisms used to repair double-strand breaks (1).

Homologous Recombination (HR)

In homologous recombination, an intact homologous duplex is used to retrieve information and fix the duplex with strand breaks. The three basic steps in homologous recombination...
Chemistry of DNA Damage Repair

**Figure 5.** Base excision repair in mammalian cells. In base excision repair, the damaged nucleotide is removed in a two-step process. A glycosylase cleaves the glycosidic bond of the damaged base, releasing it, and then AP lyase and AP endonuclease cleave the phosphodiester bonds on either side of the abasic site, which releases the deoxyribose.

are strand invasion, branch migration, and Holliday junction formation (Fig. 7). Strand invasion and branch migration are carried out by RecA in prokaryotes (29) and Rad51 in eukaryotes (30). The Rad51 paralogs (XRCC2, XRCC3, Rad51B, Rad51C, and Rad51D) along with BRCA2 are considered the mediators during homologous recombination in mammalian cells. Before the initiation of strand invasion by Rad51, the Mre11/Rad50/Nbs1 (MRN) complex processes the termini of double-strand break to facilitate strand invasion. After strand invasion and branch migration, the two duplexes are linked through the Holliday junction, which is the key intermediate of homologous recombination. In the Holliday junction, the two strands that are being recombined are joined together covalently by single-strand crossovers. A heterodimeric resolvase, Mus81/Mms4, then cleaves the Holliday junction to separate the two duplexes (31) (Fig. 7). A key feature of HR is that any information that is lost by double-strand break is retrieved using the information from a sister chromatid or homologous chromosome.

**Nonhomologous End Joining**

In NHEJ in eukaryotes, the Ku heterodimer binds to both ends of the double-strand break (32) and recruits DNA–PKcs and the ligase IV–XRCC4 heterodimer. These then ligate both ends of the double-strand break regardless of whether they actually come from the same chromosome (33). In NHEJ, Artemis is the end-processing nuclease (34) (Fig. 8).

HR is important for the rescue of collapsed replication forks. NHEJ is the physiological recombination mechanism in V(D)J recombination and seems to be the major pathway of the repair of double-stranded breaks induced by ionizing radiation and radiomimetic agents in mammals.
Cross-Link Repair

Chemotherapeutic drugs such as cisplatin, psoralen, and mitomycin C and other DNA-damaging agents cause interstrand DNA cross-links. These lesions pose unique problems for the cellular machinery because both base excision repair and nucleotide excision repair rely on the redundant information in the undamaged strand to correctly fill in the gap resulting from the damage removal. However, in cross-linked DNA, both strands are damaged and the genetic information is therefore lost. The cell overcomes this difficulty by using multiple repair systems to repair the damage. Nucleotide excision repair and homologous recombination work either coordinately or in tandem to remove these lesions in E. coli and Saccharomyces cerevisiae (35, 36), and the mechanism of removal of cross-links in mammalian cells is not well understood.

E. coli overcomes the dilemma of cross-link repair by combining nucleotide excision repair with homologous recombination. First, dual incisions are made in one strand by the UvrABC excision nuclease. Second, through homologous pairing of the incised duplex with the intact sister chromosome, a segment of the intact duplex is paired with the gap, generating a three-strand noncovalent intermediate. The transferred strand is ligated to form a three-stranded covalent intermediate. Third, the “excised oligomer” remaining linked to the duplex through the cross-link is recognized as a bulky adduct and is excised by dual incisions. The gaps generated by repair and recombination in the damaged and sister duplex, respectively, are filled in by DNA polymerases and the newly synthesized DNA is ligated (35) (Fig. 9). Genetic evidence indicates that the yeast S. cerevisiae also employs combinations of nucleotide excision repair and recombination to eliminate cross-links (37, 38).
Chemistry of DNA Damage Repair

**Figure 7** Repair of the double-strand break by homologous recombination in mammalian cells. In homologous recombination, an intact homologous chromosome is used to retrieve information and repair double-strand breaks in the duplex. The three basic steps of homologous recombination are strand invasion, branch migration, and Holliday junction formation.

**Figure 8** Repair of double-strand breaks by nonhomologous end joining in mammalian cells. In nonhomologous end joining, the ends of the double-strand break are brought together by DNA–PKcs (light blue rectangle) in conjunction with the Ku70/Ku80 (dark blue circle) heterodimer. After end processing by the Artemis nuclease and some DNA polymerases, the ends are ligated by the XRCC4–Ligase IV complex.

The role of nucleotide excision repair in cross-link repair in humans is not currently clear. Replication fork collapse due to cross-links can induce double-strand breaks during replication both in vivo and in vitro (39). Mutations in XPF and ERCC1 render mammalian cells extremely vulnerable to DNA cross-linking agents (40). This has led to the conclusion that XPF•ERCC1, which is a structure-specific endonuclease, plays a significant role in cross-link repair that is distinct from its role in nucleotide excision repair. At the double-strand break induced by replication, XPF•ERCC1 degrades one of the cross-linked strands 3′ to 5′ though the cross-link (38). Recombination proteins, which include Rad51, Rad52, XRCC2, XRCC3, and RPA then act to generate a Holliday junction intermediate, which would be resolved as described previously (40).

It has been found that patients with Fanconi’s anemia (FA) are extremely susceptible to cross-links. Evidence has been found that suggests that the FA complex (which is composed of FANCA, C, E, and F, 4 of the 12 Fanconi anemia proteins) along with the M/R/N complex and BRAC1/BRAC2 use homologous recombination to participate in both cross-link repair and double-strand break repair (41).
Mismatch Repair

Mismatch repair removes mismatched bases from newly synthesized DNA. The general strategy of mismatch repair follows. The repair reaction is usually coupled to DNA replication and corrects errors introduced by DNA polymerases in the newly synthesized strand. The mismatch is recognized by a mismatch-specific protein, and the mismatched nucleotide in the newly synthesized strand is removed by an exonuclease that initiates hydrolyses of the newly synthesized strand at a nick that may be located up to 1000 nucleotides away from the mismatch and that may be 3' or 5' to the mismatch. Exonucleolytic degradation of DNA past the mismatch is coupled to DNA synthesis to replace the excised DNA and substitute the correct nucleotide in place of the mismatched one. The mismatch repair system is evolutionarily conserved among organisms ranging from E. coli to humans. The mismatch is recognized by M uS in E. coli and by M uS6 in humans (heterodimers of M uS paralogs M SH2 and M SH6 and M SH2 and M SH3, respectively). The strand-discriminating nick is introduced by M uH in E. coli. This enzyme recognizes the transiently hemimethylated GATC sequence in the newly synthesized strand. This sequence is methylated by deoxycytidine methylase (Dam) at the A residue, and shortly after synthesis, the newly synthesized DNA is transiently unmethylated. In a hemimethylated GATC site, M uH makes a nick at the 5' side of G in the unmethylated GATC. In humans, there is no methyl-directed mismatch repair and the enzyme responsible for introducing nicks is not known (42, 43). It is thought that the terminus of the elongating leading strand and the terminus of the Okazaki fragments of the lagging strand may constitute sites of entry of the mismatch removing exonuclease. In both E. coli and in humans, a protein called M uL provides communication between the nick and the mismatch and helps direct the exonuclease degradation from the nick to the mismatch. M uL is a homodimer in E. coli, and there are two heterodimers of three M uL homologs in humans (M uL u = M LH1 plus P MS2 and M uL β = M LH1 plus P MS1). The DNA located between the nick and the mismatch is removed by 3' to 5' or 5' to 3' exonucleases depending on whether the nick is 3' or 5' to the mismatch. The gap is then filled by DNA Pol III in E. coli (Fig. 10) by DNA Pol I in humans and is ligated. Mismatch repair increases the spontaneous mutation rate in E. coli and yeast. Humans defective in mismatch repair exhibit a cancer-prone syndrome called hereditary nonpolyposis colon cancer (HNPCC), and sporadic mutations of mismatch repair genes are also associated with up to 50% of all human cancers (42–44).

References

Chemistry of DNA Damage Repair

Figure 10  Methyl-directed mismatch repair in E. coli. The mismatch repair reaction is usually coupled to DNA replication and serves to correct errors in the newly synthesized strand that are introduced by DNA polymerase.

Sensing of DNA Damage
Xiaohong H. Yang, Massachusetts General Hospital Cancer Center and Harvard Medical School, Boston, Massachusetts
Lee Zou, Massachusetts General Hospital Cancer Center and Department of Pathology, Harvard Medical School, Boston, Massachusetts
doi: 10.1002/9780470048672.wecb132

Maintenance of genomic stability is essential for the survival of cells, organisms, and species. Genomic stability relies on the complete and accurate transmission of genetic materials from mother cells to daughter cells and from one generation to the next. This task is daunting, however, because the genome is constantly challenged by numerous intrinsic and extrinsic stresses that damage deoxyribonucleic acid (DNA). If left untreated, such damage can destabilize the genome by introducing gene mutations, duplications, and chromosomal rearrangements such as deletions and inversions, which fortunately does not normally happen as cells evolve a complex damage sensing and signaling mechanism named the DNA damage and replication checkpoint. Checkpoint activation effectively pauses the progression of the cell cycle, allowing more time for the removal of DNA lesion. Moreover, activated checkpoint also regulates and coordinates a number of cellular processes including DNA repair, DNA replication, and chromatin remodeling to alleviate the stress on the genome.

The Biology of Checkpoint
Deoxyribonucleic acid (DNA) damage poses a constant and serious threat to the stability and integrity of the genome. DNA damage can develop from exposure to external sources such as chemical carcinogens, ultraviolet (UV) light, or X ray. It can also be caused by sources from within: metabolic byproducts, free radicals, and interference with duplication and segregation of genomic DNA. To maintain genomic stability in the face of DNA damage, an intricate and elaborate signaling pathway called the DNA damage and replication checkpoint is evolved. Remarkably, essential components of checkpoint are largely conserved throughout evolution, underlying the importance of this pathway and making the study of checkpoint simultaneously in different model systems rational. In eukaryotes ranging from human to yeast, two signaling pathways, which are initiated respectively by the ATM (ataxia telangiectasia mutated) kinase and the ATR (ATM- and RAD3-related) kinase, are the major guardians of genome stability (1). Although ATM primarily responds to double-stranded DNA breaks (DSBs), ATR regulates the response to a wide spectrum of DNA damage, especially those interfering with DNA replication (2). Through phosphorylation of signal transducer and effector proteins, ATM and ATR relay the DNA damage signals to downstream cellular processes. The effectors of ATM and ATR include proteins involved in cell-cycle transitions, DNA replication, DNA repair, chromatin remodeling, telomere maintenance, transcription control, and apoptosis. Collectively, the phosphorylation of these effectors enhances the ability of cells to repair and to overcome the encountered DNA damage. When the extent of damage reaches an intolerable level, activated checkpoint can also lead to programmed cell death.

What will happen if cells do not have a functional checkpoint? The ATR checkpoint is essential for the survival of cells and the embryonic development in mammals (3, 4), which is likely a result of the important function of ATR in coping with the intrinsic stresses during DNA replication. On the other hand,
Sensing of DNA Damage

the ATM checkpoint is not essential, but its defects result in phenotypes associated with genomic instability, such as high cancer risk, premature aging, and neuron degeneration (2). Mutations in several other genes involved in the ATM or ATR checkpoint response, such as p53, Brcal, and Nbs1, have also been found to associate with cancer. Thus, it is clear that the ATM and the ATR checkpoint are crucial for survival at both cell and organism levels.

The Chemistry of DNA Damage Sensing

In eukaryotic cells, a prototypical checkpoint cascade consists of DNA damage sensors, signal initiators, signal transducers, and effectors. Similar to many other signal transduction cascades, the signaling of checkpoint pathways is primarily achieved by protein phosphorylation. ATM and ATR are believed to be the signal-initiating kinases in these pathways. Albeit related in sequence, ATM and ATR differ in their DNA damage specificity. ATM responds primarily to DSBs, a rare but acute threat to genomic stability. ATR, on the other hand, is more pleiotropic and indispensable for checkpoint responses to damage caused by disturbance during normal DNA replication. With the help of distinct sets of DNA damage sensors that recognize specific DNA damage-induced structures (see below), ATM and ATR are localized to sites of DNA damage in cells and phosphorylate their substrates at these sites.

Activated ATM and ATR phosphorylate proteins of various functions, two of which are the Chk1 and Chk2 kinases. Chk1 and Chk2 are pivotal signal transducers in the checkpoint pathways (5, 6). Phosphorylation and activation of Chk1 and Chk2 not only allow the checkpoint signals to reach additional effectors, but also provide an additional layer of signal regulation. A group of signal transducers, including Brcal, TopBP1, Claspin, 53BP1, and Mdc1, mediate the signaling from ATM and ATR to Chk1 and Chk2 (7–15). These proteins are often referred to as the mediators. Many of the mediators are substrates of ATM or ATR, and they contain phosphopeptide-binding domains such as the breast cancer C-terminal (BRCT) and the forkhead-associated (FHA) domains. These properties of the mediators enable them to interact with other proteins in a phosphorylation-dependent and damage-regulated manner, thereby allowing the checkpoint signals to be relayed to downstream effectors. As a result of checkpoint activation, the cell cycle can be temporarily halted at the G1/S or the G2/M transitions, or even within the S phase. Furthermore, activated ATM, ATR, and the Chk kinases also phosphorylate proteins involved in DNA repair, DNA replication, chromatin remodeling, and gene expression to facilitate the removal of DNA lesions and to alleviate the stress on the genome.

DNA Damage Sensing by the ATM Pathway

In response to ionizing radiation or certain types of DNA-damaging agents (e.g., cisplatin), the activity of ATM is rapidly stimulated by DSBs (16). In undamaged cells, ATM molecules exist as inactive oligomers. Upon DNA damage, autophosphorylation of ATM on a Serine residue (Ser1981) transforms the latent oligomers into active monomers (Fig. 1) (17), which appears to be an essential step for ATM-mediated checkpoint activation as mutation of this residue renders the kinase functionless. It should be noted that ATM autophosphorylation can be triggered by very low doses of ionizing radiation or disturbance of chromatin structure in the absence of detectable DSBs (17). It was hypothesized that alterations of chromatin structure might partially activate ATM. However, the exact mechanism that leads to ATM autophosphorylation remains unclear.

Activated ATM is capable of phosphorylating non-DNA-bound substrates like p53, but the phosphorylation of other substrates at sites of DNA damage requires actions of the Mre11-Rad50-Nbs1 (MRN) complex (Fig. 1). In cells lacking Mre11 or Nbs1, ATM does not associate with chromatin nor does it phosphorylate its substrates such as Brcal and Smc1 (18). The MRN complex can directly associate with double-stranded DNA ends in vitro (19, 20), and it possesses a 3’ to 5’ exonuclease activity (21). Mre11, the nuclease of the complex, stably associates with Rad50 to form a globular DNA-binding head with two flexible arms comprising the coiled coils of Rad50 (19, 20). The Mre11-Rad50 complex can bind DNA, and this association stabilizes the nuclease and increases its activity (22). The MRN complex interacts with ATM to form the MRN-ATM complex (Fig. 1), which is necessary for ATM activation in response to DNA damage. The MRN-ATM complex stabilizes ATM, allowing it to autophosphorylate and become activated.
Sensing of DNA Damage

to DNA ends and tether DNA fragments through the arms of Rad50 (20, 22). Nbs1 interacts with both Mre11 and ATM, and its ATM-binding domain in the C terminus is critical for the recruitment of ATM to DNA damage (23). Interestingly, a recent study using purified proteins demonstrated that the MRN complex is able to stimulate the phosphorylation of substrates by ATM in the presence of double-stranded DNA ends (24). These findings suggest that the MRN complex, which not only recruits ATM to DSBs but also directly stimulates its kinase activity, is an important DNA damage sensor of the ATM pathway.

The recognition of double-stranded DNA ends is only the first step of the DNA damage sensing by ATM. As soon as ATM is activated at the DNA ends, it phosphorylates the histone H2AX in the adjacent nucleosomes (Fig. 1) (25). Phosphorylated histone H2AX recruits the mediator Mdc1 through an interaction with its BRCT domain (12, 26, 27), allowing Mdc1 to recruit additional ATM molecules with its FHA domain (26). This Mdc1-mediated recruitment mechanism enables ATM to phosphorylate distant histone H2AX molecules farther away from the DNA breaks, and eventually generate a large chromatin region with phosphorylated histone H2AX. Phosphorylated histone H2AX may then bring in additional ATM substrates involved in DNA repair, chromatin remodeling, and checkpoint signaling to facilitate the functions of these proteins at sites of DNA damage (26).

In addition to phosphorylated histone H2AX, methylated histones H3 and H4 are also implicated in checkpoint signaling (28–30). It was shown that the mediator 53 BP1 recognizes methylated histone H3 through its tandem Tudor domains (28). Interestingly, the methylation of histone H3 is not induced by DNA damage, which led to the hypothesis that a change of chromatin structure, which exposes the preexisting methylated site on histone H3, is induced at sites of DNA damage.

It has become clear that the sensing of DNA damage by the ATM pathway is a multistep process involving several independent sensing mechanisms. The ATM pathway recognizes not only DNA breaks but also protein modifications at sites of DNA damage. These distinct sensing mechanisms ensure that the signaling of the ATM pathway is tightly regulated at multiple levels.

Recruitment of ATR to DNA Damage

Unlike ATM, which strongly prefers DSBs, ATR is a broad-spectrum signal initiator. Various types of replication interference, such as those induced by UV irradiation or ribonucleotide reductase inhibitor Hydroxyurea (HU), strongly elicit the ATR pathway. This versatility and the pivotal role of ATR in cell viability and genomic stability has prompted an intensive investigation into the mechanisms by which ATR senses different types of DNA damage and activates the checkpoint.

In human cells, ATR exists in a stable complex with another essential checkpoint protein called ATRIP (3). In budding yeast Saccaromyces cerevisiae, ATR’s homolog Mec1 also functions in a complex with Ddc2; the budding yeast ATRIP homolog (31–33). It has been shown that Mec1 can be activated in the cdc13 temperature-sensitive mutant as well as by the HO endonuclease (34, 35). Cdc13 is a telomere-binding protein. When mutated, telomeres are unprotected and stretches of single-stranded DNA (ssDNA) are exposed because of cleavage by exonucleases (36). In the other scenario, ssDNA is also generated at the exonuclease-processed DSBs induced by HO (37). Increasing amounts of ssDNA and ATR activation were also observed when yeast replication forks are blocked by HU and UV treatments (Fig. 2) (38, 39). All these findings suggested a link between ssDNA and the activation of the ATR pathway.

Replication protein A (RPA), a protein complex with affinity to ssDNA, is a key player in many types of DNA metabolisms such as DNA replication and DNA repair (40). In G2 yeast cells, depletion of RPA abolished the localization of Ddc2 to HU-induced DSBs (41), suggesting that RPA is important for the recognition of ssDNA and ssDNA processing in the activation of the ATR pathway.
the recruitment of Mec1 to sites of DNA damage. The distinct functions of RPA in DNA replication and checkpoint activation were demonstrated in rfa-1L1, an RPA mutant strain proficient in DNA replication yet partially defective in checkpoint activation (41). In the African frog Xenopus egg extract, another model system used in the study of DNA replication and checkpoint, ATR associates with chromatin in a replication-dependent manner (42). Depletion of RPA, however, ceased the association of ATR with chromatin (43). In human cells, RPA is required for the localization of ATR to DNA damage-induced nuclear foci and the efficient phosphorylation of Chk1 (44). Collectively, these findings suggest that ssDNA coated with RPA might be a common structure recognized by ATR and ATRIP.

In a series of in vitro biochemical experiments aimed at recapitulating the initial steps of DNA damage sensing in human cells, the roles of ssDNA and RPA were directly analyzed (41). In these experiments, single- or double-stranded DNA of various lengths was biotinylated and immobilized on streptavidin-tagged magnetic beads. The association of purified ATR and ATRIP with immobilized DNA was analyzed in the presence or absence of RPA. It was found that 1) purified ATRIP protein was efficiently recruited to ssDNA only in the presence of purified RPA; 2) RPA confers ATRIP higher affinity to ssDNA than dsDNA (double-stranded DNA); and 3) the ATR-ATRIP complex, but not ATR alone, binds to ssDNA efficiently in the presence of RPA. Together, these experiments demonstrated that ssDNA coated with RPA is a structure that efficiently recruits ATR-ATRIP (Fig. 2). As RPA-coated ssDNA is commonly generated during different types of DNA repair and when DNA replication was interrupted, it is highly possible that RPA-coated ssDNA is the key structure that enables ATR-ATRIP to respond to a broad spectrum of DNA damage (Fig. 2). These findings, however, do not rule out the possibility that ATR-ATRIP can localize to specific types of DNA damage through alternative protein-protein or protein-DNA interactions.

Roles of Replication Factor C (RFC)- and Proliferating Cell Nuclear Antigen (PCNA)-like Complexes in Damage Sensing

Although RPA-coated ssDNA is critical for the recruitment of ATR-ATRIP to sites of DNA damage, RPA-stained DNA is not sufficient for ATR-ATRIP to elicit a robust checkpoint response. The function of ATR-ATRIP requires additional regulatory proteins including the RFC-like Rad17 complex and the PCNA-like Rad1-Rad11-Hus1 (9-1-1) complex. During DNA replication, the RFC complex recognizes the primer-template junctions (the junctions of ssDNA and dsDNA) at replication forks and recruits ring-shaped PCNA complexes onto DNA in an ATP-dependent manner (44). Once loaded onto DNA, PCNA functions as a sliding clamp allowing DNA polymerases to stably associate with their template. Likewise, the PCNA-like 9-1-1 complex is recruited onto DNA by the RFC-like Rad17 complex (45). However, the recruitment of 9-1-1 occurs only after DNA damage, suggesting that the Rad17 complex can specifically recognize certain DNA structures induced by DNA damage.

Single-strand DNA generated at sites of DNA damage or stressed replication fork is always juxtaposed with junctions of ssDNA and dsDNA (Fig. 2). To investigate whether the Rad17 and 9-1-1 complexes recognize such DNA junctions, these complexes were expressed in engineered insect cells and purified using affinity chromatography. In several in vitro biochemical assays, the Rad17 complex recruited 9-1-1 complexes onto DNA structures with both ssDNA and dsDNA regions (46-49), and this recruitment was enhanced by RPA (48, 49). Interestingly, unlike RFC, which uses only the 3′ double/single-strand DNA junctions to PCNA, the Rad17 complex can use the 5′ double/single-strand DNA junctions to recruit 9-1-1 complexes (48, 49). This finding provides a possible explanation as to how Rad17 and 9-1-1 complexes are recruited to resected DSBs and telomeres that possess only 5′ double/single-strand DNA junctions. As ssDNA gaps are observed at stressed replication forks, both 5′ and 3′ double/single-strand DNA junctions are induced by replication interference (Fig. 2). Together, these experiments suggest that the Rad17 and 9-1-1 complexes are sensors of the double/single-strand DNA junctions at sites of DNA damage and stressed DNA replication forks.

Similar to the DNA damage sensing by the ATM pathway, the sensing by the ATR pathway is also a multistep process. ATR-ATRIP directly recognizes RPA-coated ssDNA, whereas the Rad17 and 9-1-1 complexes recognize the junctions of ssDNA and dsDNA. Additional checkpoint regulators may also contribute to the sensing of different DNA structures at sites of DNA damage. The co-localization of ATR-ATRIP and its regulators on damaged DNA may enable the ATR to be activated by these regulators and phosphorylated by its kinases. Consistent with this idea, TopBP1, a protein that interacts with the 9-1-1 complex, stimulates the kinase activity of ATR-ATRIP in vitro (50). In summary, the ability to recognize certain DNA structures at sites of DNA damage is most likely the key for the versatility of the ATR pathway. Furthermore, the involvement of multiple sensors in this pathway may enable it to distinguish different types of DNA damage and generate a signal accordingly.

Processing of DNA Damage

Many types of DNA damage that efficiently elicit the ATR checkpoint interfere with DNA replication, suggesting that DNA replication forks may play a particularly important role in the processing of DNA damage to structures recognizable to the ATR checkpoint (e.g., ssDNA and junctions of ssDNA and dsDNA). A accumulating evidence has suggested that the uncoupling of DNA helicases and DNA polymerases at progressing replication forks may lead to increased amounts of ssDNA at the forks (51). The stalling of DNA polymerases can directly result from DNA lesions themselves or from those recognized or processed by specific repair proteins. For example, xeroderma pigmentosum group A (XPA), a protein involved in the recognition of UV-induced DNA damage, is required for the activation of ATR checkpoint by UV during S phase (52).

In addition to the DNA damage that interferes with replication, the ATR checkpoint is also elicited by DSBs. These DSBs,
However, need to be first recognized and processed by specific factors. It was recently shown that the activation of ATR by ionizing radiation, but not that by replication stress, requires ATM and the MRN complex, RPA can no longer localize to DSBs, indicating that the formation of ssDNA at DSBs is compromised. The processing of DSBs also require CDK activity and is restricted to the S and G2 phases of the cell cycle. Thus, even with distinct DNA structure specificities, the ATM and the ATR checkpoints may function together to mediate a coordinated response to DNA breaks in the replicating or replicated genome.

**Tools and Techniques**

The sensing of DNA damage by the checkpoint has been characterized both in cells and in cell-free systems by using biochemical approaches. Several methods are commonly used to introduce different types of DNA damage in cells (56). For instance, to generate DSBs in cultured cells, cells can be treated with IR (ionizing radiation), laser beam, or therapeutic compounds such as cisplatin and bleomycin. To introduce replication stress to cells, cells are treated with UV or replication inhibitors such as HU and aphidicolin. It should be noted that the actual DNA damage induced by these methods are likely heterogeneous. For example, IR induces not only DSBs but also single-strand DNA breaks induced by these methods are likely heterogeneous. For example, IR induces not only DSBs but also single-strand DNA breaks and other types of DNA lesion.

Many checkpoint-mediated protein phosphorylation events can be detected by Western blotting with phospho-specific antibodies or mobility shift on protein gels due to the modification. As a result, the phosphorylation status of these proteins has been exploited as markers for checkpoint signaling in cells. Moreover, cell-cycle stage-specific checkpoint responses can now be monitored by stage-specific assays such as the G1, intra-S, and G2/M checkpoint assays (57–59). Checkpoint kinase activities can be measured by using immunoprecipitation and in vitro kinase assays. Immunofluorescence, chromatin fractionation, and chromatin immunoprecipitation are commonly used methodologies to analyze the recruitment of checkpoint proteins to sites of DNA damage.

Several cell-free assay systems have been used to dissect the mechanisms of DNA damage sensing in vitro. The most extensively applied system is the one using Xenopus extracts. As DNA replication can occur efficiently in Xenopus extracts, this system has been used to analyze how replication interference is sensed by checkpoint sensors (60, 61) and how activated checkpoint regulates DNA replication. The checkpoint can also be elicited in a replication-independent manner by various synthetic DNA structures in Xenopus extracts (9, 62). The checkpoint response in Xenopus extracts is often monitored by the phosphorylation of various checkpoint proteins. The association of checkpoint proteins with damaged chromatin or DNA is used to monitor the recognition of DNA damage in Xenopus extracts.

In vitro biochemical assays with purified proteins are developed to demonstrate the direct recognition of damage-induced DNA structures by checkpoint sensors (41). Various checkpoint proteins have been successfully expressed in bacteria, yeast, insect, and mammalian cells and have been purified bioclorically. The association of checkpoint proteins with various DNA structures can be captured and visualized by using biotinylated DNA or DNA labeled with radioactive isotopes. Finally, in vitro kinase assay is performed to measure the effects of DNA on the activity of checkpoint kinases and the phosphorylation of their substrates.

**Future Perspectives**

Since Hartwell and Weinert first observed that DNA damage led to cell-cycle arrest in budding yeast and proposed the concept of the DNA damage checkpoint (63), the checkpoint field has evolved quickly into a leading area of both basic biological research and cancer and disease-oriented research. DNA damage and replication checkpoint, as we now know, is an indispensable and evolutionarily conserved cellular process. In all species, checkpoint is an integral part of cell survival and organism development by maintaining genomic stability against continuous insults from both outside and inside. Instability of the genome in human cells is often both a prelude and a hallmark of cancer, a debilitating disease affecting millions and the leading cause of death in many countries including the United States. In fact, mutations (p53, Brca1, Chk2) or abnormalities of the checkpoint pathway have been found repeatedly to associate with cancer or other cancer-prone genetic disorders. A thorough understanding of the chemical and biochemical steps that lead to checkpoint activation and its impact on other cellular processes will surely generate much insight into tumorigenesis at the molecular, cellular, and organ level. The study of the mechanism of action and specificity of the checkpoint kinases will also help us design safer and better drugs to fight cancer. Indeed, several inhibitors of the checkpoint kinases are already in various stages of preclinical and clinical studies. From a more fundamental standpoint, we still know very little about the actual events leading up to the sensing of DNA damage and the activation of ATM and ATR. The biochemical and functional roles of a number of checkpoint proteins remain virtually unknown. Furthermore, from examining the disease symptoms and phenotypes of knockout mice, it is obvious that defects of checkpoint interfere with development. Yet, how checkpoint influences development and the difference among specific organs and systems is not clear at all. In summary, the DNA damage and replication checkpoint is a fundamental and elaborate cellular process. It protects cells from being genetically compromised by DNA damage. Uncovering the hidden secrets of the checkpoint mechanism will lead us closer to cracking the codes of cancer and other genetic and developmental diseases.

**References**

Sensing of DNA Damage


Further Reading


See Also

Chemistry of DNA Damage Repair
Cell Cycle Checkpoint
Signal Transduction by Posttranslational Modifications DNA Replication
Chemical Biology of Cancer

Sensing of DNA Damage

Further Reading
All DNA polymerases share a common two-metal ion catalyzed chemistry of nucleotide incorporation. Structure analysis, however, suggests that DNA polymerases share one of two different ancestors, which converged to employ the same mechanism. *Escherichia coli*, the prototypical bacterium, encodes five different DNA polymerases. The chromosomal replicase functions closely with clamps, clamp-loaders, and other proteins. Oxidative damage to DNA during normal cell growth requires interplay among the several distinct DNA polymerases, which enable the replicase to circumvent these obstacles and complete chromosomal replication. Additional processes involving DNA polymerases are brought into action during heightened levels of DNA damage. We review here DNA polymerase structure, catalytic mechanism, and several pathways in which various bacterial DNA polymerases act.

**Introduction**

*Escherichia coli* DNA polymerase (Pol I) is so named because it was the first DNA polymerase to be isolated and characterized (1). Pol I remains the most intensively studied DNA polymerase, and its general structure and properties generalize to other DNA polymerases. Sequence comparisons of DNA polymerases derived from many organisms indicate the presence of at least five different classes of DNA polymerase (Table 1) (2). For simplicity, we will focus in this review on the DNA polymerases of *Escherichia coli*. *E. coli* contains five different DNA polymerases that assort into four polymerase families (Table 1). Pol I is a member of the A-family; it is ubiquitous among bacteria and plays a major role in DNA repair. Pol II is a B-family DNA polymerase that seems to be involved in repair. Interestingly, eukaryotic and archael chromosomal replicases are all members of the B-family. The C-family of DNA polymerase is specific to bacteria and functions as the chromosomal replicase. *E. coli* contains one C-family polymerase, which is referred to as Pol III. Some bacteria contain two C-family DNA polymerases, which are referred to as Pol C (the replicase) and DnaE polymerase (after the gene encoding it) (3). *E. coli* contains two Y-family DNA polymerases, Pol IV and Pol V. Y-family DNA polymerases have relatively low fidelity and thus differ from typical DNA polymerases, which usually have very high fidelity in DNA synthesis (2, 4). For example, DNA polymerases I, II, and III make mistakes only once every $10^4$–$10^6$ nucleotide additions, and fidelity is assisted in these enzymes by the presence of a 3′–5′ exonuclease, which is referred to as a proofreader that removes most of the mistakes made by the polymerase (i.e., misincorporated nucleotides) (2, 4). In contrast, Y-family polymerases have intrinsically high error rates ($10^{-1}$–$10^{-3}$) and are the only *E. coli* DNA polymerases that lack 3′–5′ exonuclease activity (5, 6). These unique properties allow Pol IV and Pol V to pass over DNA lesions and thus enable chromosome duplication in the face of DNA damage, but they do so at the expense of creating mutations (5, 7). Pols II, IV, and V are induced by DNA damage and are proposed to function collectively and to enable replication over DNA lesions (4, 8–10).

In this review we summarize the general architecture of DNA polymerases and the chemistry of the DNA polymerase and the chemistry of DNA synthesis and 3′–5′ exonuclease activities (see also DNA replication). The ultimate function of DNA polymerases is the duplication of genetic material, and therefore, we also describe how Pol III functions at a replication fork. Lastly, we present a brief overview of the different repair reactions in which the remaining DNA polymerases act.

**Chemistry of DNA synthesis and 3′–5′ exonucleolytic proofreading**

The substrates for DNA synthesis are a 3′ primed site and deoxyribonucleoside triphosphate (dNTP) (see Fig. 2). DNA polymerase catalyzes a phosphoryl transfer reaction that adds a dNMP moiety to a 3′ terminus of an existing DNA strand, releasing pyrophosphate (Fig. 3a). The reaction is catalyzed exclusively by two metal ions (e.g., Mg$^{2+}$) (11, 12). Metal ion A extracts a proton from the DNA primer 3′ terminus using a hydroxyl group to produce an oxyanion nucleophile, which attacks...
the γ-phosphate of the incoming dNTP that is base-paired to the template. Metals A and B stabilize the resulting penta-coordinate transition state, and metal B also stabilizes the pyrophosphate leaving group.

The two metal ions are held in place by three conserved aspartic acid residues. The fact that catalysis is mediated exclusively by metal ions with no direct participation of amino acid side chains suggests that nucleotide polymerization may have originated before the evolution of proteins. Specifically, the two metals may have been chelated by RNA in the primordial "RNA World" that is thought to have operated before the evolution of proteins.

The 3′–5′ proofreading exonuclease is called into action when an incorrect dNMP is added to the 3′ terminus. The 3′–5′ exonuclease is located in a separate domain with a distinct active site. The chemical reaction of the exonuclease proceeds by hydrolysis (see Fig. 2c), but it is remarkably similar to the polymerase reaction. Specifically, two metal ions catalyze the reaction: metal ion A activates water to form a hydroxynucleophile that attacks the phosphodiester bond of the 3′ terminal mismatched nucleotide. Metal A stabilizes the developing charge on the dNMP leaving group.

Structure of DNA polymerases

The crystal structures of multiple representatives of each class of DNA polymerase have been solved and compared (12). In all cases, the overall shape is that of a right hand, and it contains a minimum of three subdomains, which are referred to as the palm, fingers, and thumb. These subdomains are indicated for the structure of Pol I shown in Fig. 2a. Although the chain folding patterns of the fingers and thumb are different among the different polymerase families, the core architecture of the palm domain contains the catalytic site for DNA polymerization and is structurally conserved among the A, B, and Y families. The chain topology of the palm, and the relative location of the three Asp residues that hold the catalytic metal ions in polymerases of the A, B, and Y families, are shown in Fig. 2b. This conservation of structure in the palm suggests that the members of the A, B, and Y family share a common evolutionary ancestor.

Interestingly, C-family DNA polymerases (bacterial replicases) have a different chain topology in the palm, and the location of the Asp residues is also unique (13, 14). The palm architecture is shared by the X-family of DNA polymerases that include eukaryotic Pol β and certain nucleotidyl transferases (Table 1). Presumably X-family polymerases share a common ancestor with DNA polymerases of the C-family.

Table 1. DNA polymerase families

<table>
<thead>
<tr>
<th>Family</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A family</td>
<td>Bacterial Pol I, phage T5, T7</td>
</tr>
<tr>
<td>B Family</td>
<td>Bacterial Pol II, Eukaryotic Pol α, Pol γ, Pol μ</td>
</tr>
<tr>
<td>C Family</td>
<td>Bacterial Replicases, Pol III, Pol C, DnaB</td>
</tr>
<tr>
<td>Y Family</td>
<td>E. coli Pol IV, Pol V</td>
</tr>
<tr>
<td>X Family</td>
<td>Eukaryotic Pol β, Nucleotidyl Transferases</td>
</tr>
</tbody>
</table>

Figure 1. Chemical mechanism of DNA polymerase and 5′→3′ exonuclease. (a) DNA polymerase reaction. The enzyme chelates two metal ions using three aspartic acid residues (only two are shown). Metal ion A abstracts the 3′ hydroxyl proton of the primer terminus to generate a nucleophile that attacks the γ-phosphate of an incoming dNTP substrate. The phosphorolysis results in production of a pyrophosphate leaving group, which is stabilized by metal ion B. (b) The 3′→5′ exonuclease proofreading activity is located in a site that is distinct from the polymerase site; yet it uses two-metal-ion chemistry similar to DNA synthesis. The reaction type is hydrolysis in which metal ion A activates water to form the hydroxynucleophile. Nucleophile attack on the phosphate of the mismatched nucleotide releases it as dNMP (5′UMP in the case shown).

In all DNA polymerases, the DNA lies on the palm of the polymerase and also interacts with the thumb (see Fig. 2c). The finger domain binds the incoming dNTP for incorporation into the DNA chain. The A–T and G–C base pairs have very similar geometry, and during binding of a dNTP, the fingers domain closes over the palm to match the dNTP to the template strand forming a tight enclosure into which only the geometry of a correct base pair can fit (15). Incorrect base pairs do not have the correct geometry to fit into this tight enclosure and are usually released instead of incorporated. In the rare instance of incorrect nucleotide incorporation, the primed site usually enters the 3′→5′ exonuclease site for excision of the incorrect terminal nucleotide (see Fig. 2c).
kinetic scheme shown in Fig. 3. Most often, the DNA polymerase selects and incorporates the correct dNTP (pathway 1 in Fig. 3). Correct incorporation events are rapid. Incorporation of an incorrect dNTP (pathway 2 in Fig. 3) is slow, several orders of magnitude slower than incorporation of a correct dNTP (16, 17). Misincorporation is slow because the incorrect dNTP does not form a normal base pair with the template strand, and thus, the geometry of the incorrect base pair does not induce the correct fit in the enzyme active site needed to bring the two active site metal atoms into the proper juxtaposition for the chemical step. Furthermore, if the DNA polymerase incorporates the wrong nucleotide, the resulting mismatched 3' terminus presents yet another slow kinetic barrier to the DNA polymerase. The polymerase is very slow to extend the mismatched product by another nucleotide, because the correct geometry of the active site metals depends on the DNA substrate being correctly base paired. Therefore, the slow kinetics of two steps, misincorporation and extension of a mispaired 3' terminus, result in very few mistakes relative to the rapid incorporation of correct dNTP substrates. The kinetic pauses associated with incorporation of an incorrect dNTP give time for a mismatched primed template to melt and reposition into the separate active site of the 3'-5' exonuclease (Fig. 2c and Fig. 3). Once in the 3'-5' exonuclease active site, the proofreading 3'-5' exonuclease rapidly excises the incorrect 3' terminal nucleotide and thereby restores the substrate to one that is fully base paired (dashed arrow in the scheme of Fig. 3). Thus, the slow steps involved in misincorporation, combined with the rapid proofreading exonuclease removal of mismatched nucleotides, provide the DNA polymerase with a second try at chain extension. Most retrials have a positive outcome, which results in the incorporation of a correct nucleotide base.

DNA polymerase III holoenzyme

High-fidelity chromosomal replication in E. coli is executed by a multicomponent complex referred to as DNA polymerase III holoenzyme (see Fig. 4a (18-21)). Pol III holoenzyme consists of three main subcomponents: Pol III core, β-clamp, and γ-complex clamp-loader. Pol III core is the replicative DNA polymerase that consists of three subunits (α, ε, δ): α exhibits DNA polymerase activity, ε performs 3'-5' exonuclease activity necessary for proofreading, and the function of δ is currently unclear. Pol III core requires direct association with the β-clamp to perform processive DNA synthesis. β-tethers the polymerase to its respective template by binding the α subunit of Pol III core while encircling DNA immediately behind the polymerase (see Fig. 4a). Coupling of Pol III core to β results in an increase in both the catalytic rate (~10^7 s^{-1}) and processivity (~50 kb). The structure of the β-clamp resembles a symmetrical ring with sufficient space to accommodate double-stranded DNA (see Fig. 5a). β is composed of a homotrimer with each protomer consisting of three globular subdomains. The structure of the eukaryotic clamp, proliferating cell nuclear antigen (PCNA), is similar to β (23, 24). However, PCNA is composed of a homotrimer with each protomer consisting of three subdomains. The head-to-tail protomer organization of β results in two structurally distinct "faces": the C-terminal face is involved in intermolecular interactions (i.e., binding Pol III core α subunit) (see Fig. 5a and 5b). β also binds Pol I, II, IV, and V and is therefore thought to play a role in polymerase trafficking during transcription synthesis (which is discussed in more detail below). Lastly, recent studies indicate that β interacts with DNA repair proteins M5, M6, and DNA ligase (25). The function of δ in DNA repair is currently unclear.

The γ-complex clamp-loader is a heteropentameric ATPase responsible for assembling β onto DNA at primed sites (26-28). The clamp loading process is illustrated in Fig. 5b. The γ-complex binds and opens the ring-shaped β-clamp in the presence of ATP. The ATP-bound γ-complex selectively binds to a primed site that stimulates ATP hydrolysis and results in release of the clamp-loader and closure of β around DNA (20).
The proteins currently thought to act at the replication fork include two Pol III core molecules, two γ-clamps, γ-complex, DnaB helicase, primase, and single-stranded DNA binding protein (SSB) (see Fig. 4a) (20, 32). Each Pol III core is dedicated to either the leading or the lagging strand and is tethered to its respective template via binding to the γ-clamp. The unwound portion of the lagging strand is thought to form a "trombone" loop that enables bidirectional movement of the lagging strand polymerase and the replisome (33).

The clamp loader acting at the replication fork is thought to be composed of subunits γ, τ, h, l, χ, ψ (32, 34, 35). In this configuration, two γ subunits are substituted by two τ subunits. γ (47 kDa) and χ (71 kDa) share identical N-terminal clamp loading sequences as they are encoded by the same gene, dnaX. However, γ is truncated because of an efficient translational frameshift that produces approximately equal amounts of γ and χ. The χ subunit connects to SSB and is held in the structure by τ (36). The γ-to-SSB connection enhances processivity and enables efficient transfer of RNA primers from primase to the clamp loader (37). Each γ subunit binds to a single Pol III core thereby linking the leading and lagging strand polymerases together (see Fig. 4a) (14, 21, 38, 39).

DnaB is the replicative helicase that unwinds the parental duplex DNA ahead of the replication fork (see Fig. 4a) (40). DnaB encircles the lagging strand as a homohexamer and uses the energy of ATP hydrolysis to unwind DNA with 5'-3' polarity (see also DNA Helicases, Chemistry and Mechanisms of). DnaB is connected to Pol III holoenzyme via its interaction with the χ subunit of the clamp loader (41). This interaction greatly stimulates DnaB activity at the replication fork (42). DnaB also binds to and stimulates primase, which is a specialized RNA polymerase that synthesizes RNA primers approximately 12 nucleotides in length to initiate DNA synthesis (1, 43).

The mechanism of Pol III action at the replication fork is illustrated in Fig. 4. The lagging strand is synthesized as discontinuous sections of 1-3 kb called Okazaki fragments. An Okazaki fragment is initiated by primase action (see Fig. 4a). The clamp loader uses the energy of ATP hydrolysis to assemble β onto the newly synthesized RNA primer (see Fig. 4a). The Pol III core binds β and processively synthesizes the DNA portion of the Okazaki fragment, which creates a DNA loop (see Fig. 4c). The Okazaki fragment is completed when Pol III encounters the 5' end of the previous Okazaki fragment and dissociates from the DNA, which disassembles the loop. Pol I is required for Okazaki fragment maturation and is discussed below.

The leading strand polymerase performs DNA synthesis in a continuous fashion and therefore requires an initial priming event. However, DNA damage or other impediments along the leading strand may block progression of the replication fork and...
Bacterial DNA polymerases: Chemistry of Primase synthesizes a RNA primer(a) Leading strand

Lagging strand polymerase synthesizes the next Okazaki fragment (d) Lagging strand polymerase cycles to the next Okazaki fragment (c)

A sliding-clamp is assembled at the newly primed site (b)

Figure 4 Pol III function at the replication fork. Leading and lagging strand synthesis are performed in a continuous and discontinuous fashion, respectively. The lagging strand is synthesized in contiguous sections called Okazaki fragments. (a) An Okazaki fragment is initiated as primase synthesizes a short RNA primer along the lagging strand. (b) The γ-complex clamp loader uses the energy of ATP hydrolysis to assemble β at the newly primed site. (c) The lagging strand polymerase cycles to a newly synthesized primed site after completing the previous Okazaki fragment. (d) The lagging strand polymerase synthesizes the DNA portion of the next Okazaki fragment. (Adapted with permission from Reference 20.)

The repair of arrested replication forks involves two major pathways: recombinational repair and translesion synthesis. Recombinational repair uses RecA and a host of other recombination factors that facilitate strand exchange and non-mutagenic replication across damaged DNA (reviewed in References 48–50). Translesion synthesis employs low-fidelity DNA polymerases that often perform mutagenic replication opposite a damaged template (reviewed in References 5–7).

Translesion synthesis by low-fidelity DNA polymerases Chromosomal DNA is often damaged by exposure to exogenous (i.e., UV irradiation, chemical agents, etc.) and endogenous (i.e.,
Figure 5. Structure and function of β-clamp and γ-complex. (a) The structure of β resembles a ring with sufficient space to accommodate duplex DNA. It is a homodimer in which each protomer contains three globular subdomains. The protomers are arranged in a head-to-tail fashion resulting in two structurally distinct "faces." The C-terminal face is involved in intermolecular interactions. Protomers are indicated in light and dark shades (PDB ID 2POL). (b) γ-complex uses the energy of ATP hydrolysis to assemble β onto DNA at primed sites. γ-complex binds to and opens β in the presence of ATP. The clamp loader selectively binds to primer-template junctions, which stimulates ATP hydrolysis and results in dissociation of γ-complex and closure of β around the DNA. DNA polymerase binds the C-terminal face of β and initiates DNA synthesis at the primed site. (c) Subunit organization of γ-complex clamp loader. The γ-complex is composed of five subunits (γ3, δ′, δ) that are arranged in a spiral-like fashion. γ-complex contains three ATP sites that lie at the subunit interfaces and include an "arginine finger" within a conserved SRC motif. The ATP site organization of the γ-complex is proposed to confer cooperativity between subunits. (Adapted with permission from Reference 20.)

oxidative damage) insults (see also DNA Damage, Chemical Biology of Diseases Related to (51). Spontaneous mutagenesis of nucleotide bases by cytosine deamination to yield uracil also occurs with high frequency in the cell. DNA repair mechanisms have evolved to excise and repair various types of DNA damage (discussed below) (see also DNA Damage Repair, Chemistry of). However, the replication fork sometimes encounters DNA damage that has evaded the repair machinery. Pol III and other high-fidelity DNA polymerases cannot replicate damaged DNA. Thus, the replication fork becomes arrested at the site of DNA damage. Low-fidelity lesion bypass DNA polymerases can extend DNA across various types of damaged DNA, but this often results in a mutation. Thus, switching between high-fidelity and low-fidelity DNA polymerases is thought to facilitate replication through sites of DNA damage (4, 8-10, 52). Pol I and Pol III exhibit high-fidelity DNA synthesis and make mistakes on average of only 1 in 10^10 nucleotides (5, 16, 17). Y-family Pol IV and Pol V misincorporate nucleotides 10 to 1000 times more frequently and are error-prone lesion bypass polymerases that are associated with the mutagenic response of cells to DNA damage (5, 6, 16, 52). Structures of Y-family polymerases suggest that replication of past aberrant nucleotides is likely caused by a highly solvent exposed DNA binding cleft (53, 54). Thus, bulky lesions and nucleotide mismatches are more permissible among this class of enzyme.
Replication fork arrest by DNA damage leads to RecA-mediated induction of the SOS response pathway. The SOS response involves upregulation of over 40 genes that facilitate increased viability during stress (i.e., chemical agents, UV light, and nutrient deficiency) (reviewed in References 5 and 6). The DNA polymerases that are upregulated during the SOS response include Pol II, Pol IV, and Pol V. Pol II is the first DNA polymerase to be induced (∼1 minute). Interestingly, Pol II has been shown to perform high-fidelity DNA synthesis (error rate = 10⁻³ to 10⁻⁴ per base pair) on undamaged templates, but despite this, Pol II seems to be involved in error-prone DNA synthesis in vivo (∼1 minute). Pol II is somewhat of an enigmatic enzyme because it has been implicated in translesion synthesis, adaptive mutation, and recombinational repair (which is discussed below) (∼4–6, 56). Additional study is required to understand the relevant functions of Pol II in the cell.

Pol V is upregulated from ∼15 to ∼200 copies per cell approximately 45 minutes after SOS induction. Pol V is a heterotrimer with a subunit composition of UmuD′C, (see Fig. 6a). UmuD contains the catalytic domain of Pol V. UmuD′ is the product of RecA-modulated proteolytic cleavage of UmuD. Importantly, Pol V function requires RecA, and the mechanism by which RecA stimulates Pol V activity has been under investigation for several years (57). Recent data indicate that RecA-nucleoprotein filaments act in trans to stimulate Pol V (see Fig. 6a) (57, 58). Interestingly, RecA filaments in cis (immediately 5′ to Pol V on the DNA template) do not stimulate Pol V (59). Pol V and SSB may cooperate to displace RecA from DNA (57, 60). Alternatively, recent data indicate that RecA may be removed by UvrD helicase, as UvrD is also induced during the SOS response, but whether UvrD plays a direct role in translesion synthesis is currently an open question (61, 62).

Pol IV and Pol V use the η-clamp (63-66) and are capable of synthesizing DNA opposite various types of lesions (i.e., abasic site, thymine dimer, and benzo(a)pyrene) (see Table 2) (4, 8, 9, 67). Pol IV is expressed constitutively at ∼250 copies per cell, suggesting a role during normal growth, and consistent with the finding that Pol IV is required for adaptive mutation under nonlethal conditions (6). Pol IV is increased 10-fold after SOS induction. Pol IV is less than 200 copies per cell (4, 8, 9). Pol IV is involved in translesion synthesis during translesion synthesis based on structural and biochemical studies is illustrated in Fig. 6b. The dimeric η-clamp can bind both Pol IV and Pol III simultaneously, and it is proposed to hold Pol IV away from the DNA template while Pol III activity synthesizes DNA (68, 69). However, during a Pol III encounter with a lesion, Pol IV is allowed to gain access to the template in place of Pol III. After the lesion is bypassed by Pol IV, the high-fidelity Pol III regains control of the primed site. These events limit the action of the low-fidelity Pol IV to the region in which it is required to bypass a lesion.

DNA polymerase activity during repair

Oxidative damage that leads to replication fork arrest is common during bacterial growth and probably occurs at least once during each cell division even under the most favorable conditions (48, 51). Most lesions are not dealt with by mutagenic polymerases, but instead they are repaired by recombination, which is an error-free process. Recombinational repair can take many different paths and employs several different proteins, but ultimately, it involves high-fidelity DNA polymerases such as Pol I, Pol II, and/or Pol III (6, 49, 56). For additional information, we refer the reader to more thorough reviews of the various recombinational repair pathways (64, 48–50, 70).

Even high-fidelity DNA polymerases sometimes incorporate an incorrect nucleotide. All cells contain a specialized mismatch repair system that catches these errors and corrects them. Mis-match repair involves excision of DNA past the mismatch, and eventually, it allows Pol III to try again (reviewed in References 71 and 72). The process of mismatch repair requires proteins MutS and MutL, among others. Mutations in the human homologues of these proteins lead to a predisposition to development of tumors (73).

Nucleotide excision repair (NER) is responsible for removing a wide array of DNA lesions that predominately occur as a result of exposure to oxygen or UV light (reviewed in Reference 74) (see also DNA Damage, Chemical Biology of Diseases Related to). Unlike mismatch repair, NER employs Pol I to replace damaged DNA. NER also requires the action of Uvr (ultraviolet radiation) proteins that were originally discovered because of their ability to sustain cell growth after exposure to UV irradiation. Like MutS and MutL, mutations in the human homologues of the Uvr proteins predisposes individuals to cancer.

Base excision repair (BER), in which a single damaged base is excised and replaced, also uses Pol I to replace the modified nucleotide. BER is involved predominately in the repair of apurinic and apyrimidinic nucleotides (abasic (AP) sites) that frequently develop from spontaneous hydrolysis (reviewed in Reference 75). Importantly, BER is also responsible for removing uracil from DNA, which is the product of cytosine deamination. Bases modified by oxidation and alkylation are also repaired; however, these modifications occur less frequently.

A recent study suggests that Pol IV and Pol V may also be involved in BER because of their intrinsic lyase activity (76). Shen et al. demonstrate relatively weak 5′-phosphodiester bond cleavage of an abasic site (AP′-dRP activity) by Pol IV and Pol V. Interestingly, several DNA polymerases from the X, A, and Y families exhibit lyase activity. Human Pol η, for example, exhibits strong AP′-dRP activity and is thought to be involved in the excision and incorporation of nucleotides during BER (77, 78). Furthermore, eukaryotic pol ι, which is a Y-family homolog of Pol IV and Pol V, has also been shown to exhibit AP′-dRP activity (79). Additional studies are required to determine whether Pol IV and Pol V are involved in BER.

Lastly, Pol I is responsible for repairing DNA during Okazaki fragment maturation. Pol I replaces the RNA portion of Okazaki
Bacterial DNA polymerases: Chemistry of...

Figure 6. Mechanisms of translesion synthesis. (a) Activation mechanism of Pol V during translesion synthesis. Pol V is a heterotrimer composed of subunits UmuC and UmuD. UmuC is the catalytic domain, and UmuD is the product of RecA-mediated proteolysis. Translesion synthesis by Pol V is activated by the presence of a RecA filament in trans. (b) Model of DNA polymerase switching during translesion synthesis. Pol II* and Pol IV each bind to a β protomer at a conserved hydrophobic protein binding pocket (QL[S/D]LF). 1. Pol III* is arrested at the site of DNA damage, whereas Pol IV is held in an inactive state away from the DNA. 2. Pol IV gains hold of the primer terminus from Pol III* at the stall site; Pol III* is now held away from the DNA. 3. Pol IV extends the DNA past the lesion. 4. Pol III* regains hold of the primer terminus from Pol IV.

Table 2. Properties of translesion DNA polymerases

<table>
<thead>
<tr>
<th>Lesion</th>
<th>DNA polymerase</th>
<th>Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abasic site (2, 80)</td>
<td>Pol IV, Pol V</td>
<td>-1 frameshift</td>
</tr>
<tr>
<td>Benzo(a)pyrene (4, 81)</td>
<td>Pol IV, Pol V</td>
<td>Preferential incorporation of dATP and dGTP</td>
</tr>
<tr>
<td>N-2-acetylaminofluorene (4)</td>
<td>Pol IV, Pol II</td>
<td>-1 frameshift</td>
</tr>
<tr>
<td>8-oxo-guanine (9)</td>
<td>Pol IV, Pol V</td>
<td>-2 frameshift</td>
</tr>
<tr>
<td>6-4 Thymine-thymine photoproduct (2)</td>
<td>Pol IV, Pol V</td>
<td>-2 frameshift</td>
</tr>
<tr>
<td></td>
<td>Pol V</td>
<td>T → C</td>
</tr>
</tbody>
</table>

Adapted with permission from Reference 5.

fragments using its 5′→3′ exonuclease activity and synthesizes DNA in its place. The remaining nick between adjacent Okazaki fragments is repaired by DNA ligase.

Future perspectives

The work of many laboratories has provided significant insight into the detailed workings of Pol III function at the replication fork. The way in which the replisome interacts with other DNA processes (i.e., DNA compaction, transcription, and recombination, repair), however, is poorly understood. For example, how does Pol III holoenzyme replicate through high-affinity protein-nucleic acid complexes such as RNA polymerase or repressor proteins? What is the fate of the replication components after an encounter with a blocking lesion? Does the replisome directly recruit or interact with recombination factors that repair broken forks? β has been shown to bind mismatch repair proteins MutS and MutL. What is the significance of these interactions?

The presence of five DNA polymerases in E. coli suggests that access to primed sites might be regulated. Indeed, recent biochemical data indicate that β plays a major role in switching between high-fidelity and low-fidelity DNA polymerases during translesion synthesis. High-fidelity DNA polymerases Pol I and Pol II also bind to β and are expressed at relatively high levels relative to Pol III. Do high-fidelity DNA polymerases compete for primed sites during normal growth? Furthermore, we still do not understand the major role of Pol II in the cell.

Lastly, translesion synthesis by Pol V is stimulated by RecA filaments in trans. The presence of RecA filaments in cis (on the DNA template), however, inhibits the activity of Pol V.
Additional work is required to elucidate the exact mechanism and organization of RecA filaments in the cell responsible for Pol V activation. Future studies will likely reveal the answers to many of the questions addressed above.

Acknowledgment

The authors are grateful to their source of funding, NIH GM 38339

References

3. Bruck I, O’Donnell M. General features of chromosomal repli-
6. Goodman MF. Error-prone repair. DNA polymerases in prokary-
9. Wagner J, Etienne H, Janel-Bintz R, Fuchs RP. Genetics of muta-
genesis in E. coli: various combinations of translation polymerases (Pol II, IV and V) deal with lesion/sequence context diversity. DNA Repair (A) 2002;2:159–167.
34. Glover BP, McHenry CS. The DdD-binding subunits deltar and psi are bound to gamma and not tau in the DNA polymerase III holoenzyme. J. Biol. Chem. 2000;275:3017-3020.
35. Glover BP, McHenry CS. The chi psi subunits of DNA poly-
erase III holoenzyme bind to single-stranded DNA-binding pro-


DNA Helicases, Chemistry and Mechanisms of

Robert A. Pugh, Department of Biochemistry at the University of Illinois at Urbana-Champaign, Illinois

Maria Spies, Department of Biochemistry, Center for Biophysics and Computational Biology at the University of Illinois at Urbana-Champaign, Illinois

doi: 10.1002/9780470048672.wecb133

DNA helicases are members of a larger class of enzymes composed of molecular motors that couple the energy of NTP binding and hydrolysis to directional translocation along a nucleic acid lattice. Structurally and mechanistically related helicases perform drastically different roles in DNA metabolism by associating with various macromolecular machineries that orchestrate DNA processing events. Within these macromolecular ensembles, DNA helicase translocation is coupled to separation of duplex DNA into two single strands, which is necessary for DNA replication, repair, recombination and transcription. Translocation by DNA helicases can also result in the disassembly of protein–nucleic acid complexes formed during DNA metabolism. One major consequence of cellular dependence on DNA helicases is that defects in these enzymes result in a broad spectrum of disorders usually characterized by premature aging, susceptibility to cancer, and other diseases normally associated with aging.

We will start this article by discussing the structure–function relationship in the six major DNA helicase families as defined by the presence of distinct sets of helicase signature motifs and conserved structural features. We will then outline similarities and mechanistic differences between DNA helicases and other related motor proteins, as well as describe currently accepted mechanistic models for the mechano-chemical coupling of NTP hydrolysis with DNA translocation and unwinding. Both traditional and developing experimental techniques used to study DNA helicases along with the most important and challenging questions regarding the molecular mechanisms and cellular activities of these enzymes will be also discussed.

DNA helicases are ubiquitous enzymes found in all domains of life. Bona fide helicases use the energy derived from binding and hydrolysis of nucleoside triphosphates (NTPs) or deoxynucleoside triphosphates (dNTPs) to translocate along a nucleic acid lattice (reviewed in Reference 1). Unidirectional translocation provides the duplex separation required for all aspects of DNA metabolism, as well as the disassembly of protein–nucleic acid complexes formed during DNA-manipulation processes. Nucleic acid helicases and related proteins are classified into six superfamilies based on the presence of highly conserved amino acid sequence motifs that define the motor core of these enzymes (1, 2). Despite sharing the highly structurally conserved motor core, the cellular roles of the helicases within the same superfamily are highly specific and determined by unique structural features that are incorporated into the otherwise conserved core or through protein–protein interaction domains that lead to their incorporation into larger macromolecular machineries.

Biological Background

In all cellular organisms, double-stranded DNA (dsDNA) carries genetic information. To access, process, and interpret this information, the double-stranded helix must be separated into its two-component single-stranded DNA (ssDNA) strands. Once the duplex is unwound, DNA processing events such as DNA replication, repair, recombination and transcription may proceed. Both unidirectional translocation of a helicase along the DNA strand and helicase-mediated unwinding of double-stranded DNA by disruption of the hydrogen bonds formed between pairs of complementary bases are energetically unfavorable, and they depend on the energy derived from hydrolysis of the phosphoanhydride bonds of NTP.

Biomedical importance of DNA helicases

Cellular DNA suffers an estimated $1.9 \times 10^5$ modifications per day per cell as a result of environmental stress induced
by ultraviolet light exposure, toxic chemicals, and the byproduct of normal cellular metabolism (3). This constant attack necessitates existence of multiple DNA-repair mechanisms for maintaining genomic integrity. Many cellular DNA repair machineries rely on the strand-separation properties provided by helicases. Therefore, it is not surprising that defects in human helicases are linked to a broad spectrum of disorders, which are usually characterized by premature aging, susceptibility to cancers, and other diseases normally associated with aging, immunodeficiency, or mental retardation (4). Several diseases, for example, are associated with mutations in the nucleotide excision repair (NER) helicases: XPD (ERCC3) and XPF (ERCC2). These diseases include Xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy (5). These NER helicases are also involved in protecting cells from retroviral infections (6). Increased susceptibility to breast and ovarian cancer as well as the chromosome instability disorder Fanconi anemia has been linked to mutations in Bach1, which is a helicase that has been implicated in repair of double-stranded breaks in DNA and DNA interstrand crosslinks (7, 8). The precarious aging diseases, Werner’s syndrome and Rothmund-Thomson syndrome, as well as Bloom’s syndrome, have all been associated with mutations in three of the five human RecQ-like helicases [WRN(RecQ3), RecQ4, and BLM, respectively], which are involved in DNA repair through homologous genetic recombination and DNA replication (reviewed in References 4 and 9).

Classifications

Through the use of bioinformatics, helicases are grouped into six superfamilies (SF1–SF6, respectively) that are defined by the presence of so-called helicase signature motifs (for recent review see Reference 1) (Fig. 1). These motifs are the most conserved elements of the helicase motor cores and encode the amino acids involved in the binding of NTP and nucleic acid and are responsible for transduction of the conformational change on NTP binding and hydrolysis. It is important to note that not all proteins that possess the characteristic signature motifs are bona fide helicases. Some of these enzymes act as nucleic acid translocating motors (for example, motor subunits of type I restriction enzymes and viral DNA packaging motors), protein-nucleic acid remodeling machines (such as chromatin remodeling factors) (10), or ATP-induced conformational switches that provide local RNA strand separation (11, 12). Alternatively, we currently lack sufficient predictable power to use an amino acid sequence comparison to determine a correct translocating lattice for a given helicase-like protein. Related enzymes have been shown to translocate on DNA, RNA, or even a polypeptide lattice. Examples of peptide translocases include SccA (13), which is a component of the bacterial protein export machinery that contains motifs characteristic to SF2 helicases, and CtpA (14), which is a member of the AAA + ATPase family (15) that resembles SF3 and SF6 hexameric helicases. Thus, any other classifications of bona fide DNA helicases are possible only when extensive biochemical studies have been carried out. Because the phosphodiester backbone of nucleic acid strands have defined polarity, helicases can be classified based on direction of their translocation along a single strand of DNA. Helicases that move in 3' → 5' direction are designated as "A", whereas 5' → 3' helicases are categorized as "B". A final distinction can be made among related helicases based on their ability to translocate on ssDNA or dsDNA designated as "u" or "p", respectively (1).

Enzymes in the largest superfamilies I and II (SF1 and SF2, respectively) are structurally similar. The motor cores of these helicases are formed by two so-called RecA-like folds shown in (Fig. 1a), and they contain the signature motifs. Although they perform similar functions, the motifs shared among SF1 and SF2 helicases differ in composition and arrangement. High-resolution three-dimensional crystal structures of several members of SF1 and SF2 revealed that the conserved signature motifs are located at the interface between the two RecA-like folds [ternary complex of UvrD helicase with a nonhydrolyzable ATP analog and forked DNA substrate (16) is shown as a typical SF1 representative, whereas N33 dsDNA binary complex (17) represents SF2 helicases in (Fig. 1a)].

Helicases from superfamilies III, IV, V, and VI form ring structures that consist of six protomers (Fig. 1b). Each protomer within the ring contains an AAA + (ATPases associated with multiple cellular activities) fold or a fold similar to the RecA/F7-ATPase (1). These six subunits interact to form six asymmetric nucleotide-binding pockets at the interfaces between adjacent subunits. These enzymes use the sequential ATP hydrolysis mechanism (18), which will be described below. The crystal structure of the E1 helicase from bovine papillomavirus bound to DNA in the central channel and the asymmetric ATP binding sites (19), which exemplifies a hexameric replicative DNA helicase, is shown in (Fig. 2).

Unique features and incorporation into macromolecular machineries

Each cell contains numerous DNA helicases along with a multitude of RNA helicases, nucleic acid translocating motors, and helicase-like molecular switches. Therefore, it is crucial for cellular DNA metabolism that a specific enzyme will interact with the designated DNA-processing intermediate at the appropriate time. Many helicases contain unique (or group-specific) auxiliary domains that enable enzymes with additional functions or alter their mechanistic properties, which makes them suitable for performing a diverse set of activities in various DNA processing pathways (Fig. 2). The unique features of helicases confer on these enzymes the ability to: 1) interact with a particular cognate DNA processing intermediate, which could be either a DNA molecule or a protein-nucleic acid complex; 2) unwind just right amount of DNA; and 3) “hand-out” the transiently separated DNA strands to the next step in the DNA processing event (Fig. 2). Specificity can be achieved in dsDNA through use of auxiliary domains that enable a helicase to bind to its cognate DNA substrates. For example, the archael XPB (Rad25) helicase possesses a domain responsible for the recognition of bulky DNA damage during initiation of NER (20). Another NER helicase, XPD (Rad3), preferentially binds to the ssDNA-dsDNA junction of a forked DNA substrate (21). Similarly to XPD, other members of the Rad3 family (such as Ding (22) and Bach1 (23)) are also structure-specific helicases.
DNA Helicases, Chemistry and Mechanisms of

Figure 1 Classification of helicases and conserved motifs. (a) Nonhexameric SF1 and SF2 helicases. The crystal structures of SF1 UvrD helicase (PDB: 2IS6) bound to DNA and AMPNPP and SF2 HCV NS3 helicase (PDB: 1A1V) bound to ssDNA are shown. The conserved motifs comprising the motor core of the enzymes are highlighted in colors coordinated to the labels bearing the letter or numeral identification of the respective motif. The two RecA-like folds indicated as 2A and 1A (SF1) and 2 and 1 (SF2). The conserved motifs are located at the interface between the folds and are positioned near the ATP binding site as indicated in each structure. (b) Conserved motifs of the hexameric helicases of SF3-SF6. The crystal structure of an SF3 Papillomavirus E1 hexamer helicase (PDB: 2GXA) in complex with ssDNA and ADP. Each protomer within the hexameric ring is depicted in a different color. The signature motifs located at the interface between each pair of protomers with motifs A, B, and C as well as the ssDNA binding loop contributed by one protomer, whereas motifs B' and switch 2 contributed by the adjacent protomer, are shown for one active site. Although the E1 structure was solved in the presence of ADP, the two nucleotide-binding sites within the ring were found in a conformation expected for the ATP-bound state. These sites are indicated as ATP.

Prototypical representatives for each superfamily are listed in the colored circles that schematically depict the relative sizes of the respective superfamilies. The signature motifs for each superfamily are shown by their respective circles.

In bacteria, RuvA specifically recognizes the Holliday junction formed during homologous recombination (HR) and promotes its migration (24). Additionally, several members of the RecQ helicase family also display specificity for various recombinational intermediates (reviewed in Reference 25).

Helicases are often characterized biochemically isolated from their interacting partners. However, in the cells, helicases usually function as integral parts of macromolecular machineries responsible for carrying out DNA metabolism. The presence of the associating proteins can greatly affect the unwinding properties of helicases and can aid in the identification of cognate DNA substrates, which represent the DNA processing intermediates in respective pathways.

An important characteristic of a helicase commonly affected by interaction with accessory proteins is processivity. Processivity is defined as the probability that the helicase will take the next step forward and separate another base pair as opposed to dissociating from DNA. Some helicases are extremely processive for example, the bacterial recombinational helicase UvrB can unwind 30,000 base pairs per binding event (26). Although RecBCD contains two helicase motors and a "built-in" processivity factor (27), other processive helicases require interaction with external processivity factors to unwind long stretches of DNA at a speed comparable with that of the RecBCD. Examples of such enzymes include the Escherichia coli replication helicase, DnaB, which interacts with the \( \tau \) subunit of DNA polymerase. This interaction between DnaB and DNA polymerase increases the rate of movement of the replication fork to over 1000 base pairs per second (28). Interaction with the origin recognition protein RepD makes PcrA a more processive helicase that can sustain rolling circle replication of large plasmids (29).

Although DNA unwinding by a helicase stems from its directional translocation along DNA, many helicases actively separate duplex DNA. These helicases usually have a structure that is driven through the middle of the duplex and serves as a molecular ploughshare that physically separates the two strands. RecG and UvrB have such wedge structures positioned strategically at the ssDNA-dsDNA junction (30, 31). This structure also allows for maintaining separation of the two strands as the displaced strand is directed through a second DNA-binding channel within the enzyme. RecBCD also has a pin structure...
DNA Helicases, Chemistry and Mechanisms of

Figure 2  Accessory elements that allow DNA helicases to perform unique functions in the cell. Co- and trans-acting factors enabling helicases with unique properties are schematically depicted around a conserved motor core, which is involved in coupling of ATP binding and hydrolysis to directional translocation along DNA. These elements include processivity factors, structural features involved in duplex destabilization, factors providing the substrate specificity, stabilization of the nascent ssDNA strands, as well as additional enzymatic activities.

located within the RecC subunit that after binding of the enzyme to duplex DNA splits the two strands into two separate channels within the helicase (27).

A final property of helicases is product stabilization and hand-out of the single DNA strands for the next DNA processing step. Renaturing of nascent single strands after unwinding is often prevented by the ssDNA binding proteins, SSB in E. coli (32), and RPA in eukaryotes (33).

Cellular roles and functions

DNA helicases are integral parts of many cellular machineries that play critical roles in DNA replication, repair, recombination, and transcription. Every stage of DNA replication, which includes initiation, elongation, termination, and restarting of the replication fork after stalling or collapse, involves at least one DNA helicase. Most replicative helicases (for example bacterial DnaB, viral E1, and archaeal and eukaryotic MCM helicases) are ring-shaped hexamers that have been identified as SF3 or SF4 helicases. These enzymes are highly processive and very fast; this attribute is imperative for complete and rapid replication of the genome. Although the closed ring shape is an important determinant of processivity, ring assembly around one strand of the circular bacterial DNA or at the replication origins internal to long eukaryotic chromosomes requires assistance by other proteins. During replication initiation, these helicases are recruited to the replication origin, which results in local helix destabilization and strand separation (reviewed in Reference 34). The helicase rings then translocate away from the origin and unwind the double helix ahead of DNA polymerase and the rest of the replication machinery.

When the replication machinery encounters a site of DNA damage that requires repair, progression of the replication fork stalls or collapses. One of the best-studied helicases involved at this stage is bacterial RecG. This enzyme specifically interacts with and catalyzes regression of the stalled fork, which provides a second chance for cellular DNA-repair machineries to fix the damage (reviewed in Reference 35). Bacterial replication fork restart may also be achieved through a recombination-dependent process that involves the RecBCD helicase/nuclease and the PriA helicase, which is specific for D-loop structures produced by recombinatorial proteins (see Reference 28 for review). WRN, which is a member of the RecQ family of helicases in humans, has also been implicated in progression and restart of the replication after DNA damage or replication arrest (36).

NER and transcription that involves transcription factor IH (TFIIH) in eukaryotes and archaea provides an interesting case of a process that requires the function of two DNA helicases of opposite polarity cooperating in local strand separation at the sites of DNA damage or during promoter opening. TFIIH contains two helicases: a 3′→5′ enzyme XPB (ERCC3 or Rad25) and a 5′→3′ helicase XPD (ERCC2 or Rad3). Cooperating XPB and XPD enzymes separate duplex DNA at the site of DNA damage (37, 38) and hand over the resulting repair bubble to the structure-specific endonucleases that cleave both upstream and downstream of the damaged region, which allows it to be removed and a new strand to be synthesized (39). Notably, interactions with other proteins of the TFIIH complex significantly increase the processivity and ATPase activity of XPB and XPD helicases, and they are likely to be involved in regulating activity of these helicases so that just enough of the duplex is unwound around the damaged site for the repair process to proceed (See References 37 and 39, and the references within). Like many aspects of cellular DNA metabolism, HR requires transient formation of the ssDNA intermediates and therefore the effector DNA helicases (reviewed in Reference 40). Additionally, some steps in HR, which do not require unwinding of a DNA duplex, benefit from the translocate activity of the helicase-like enzymes. The main recombination pathway in E. coli is initiated by the RecBCD complex that, similar to TFIIH, contains two helicase subunits of opposite polarity, a 3′→5′ helicase RecE and a 5′→3′ helicase RecD. In contrast to TFIIH, however, RecE and RecD subunits move along two opposite DNA strands and in conjunction with the third subunit RecC
form an exceptionally efficient machine for processing DNA ends. RecBCD is equipped with all the necessary accessory factors required for recognition of a blunt ssDNA end, fast and processive unwinding of the DNA duplex, recognition of the recombination hot spot, imbedded into the DNA on which the enzyme translocates, \( \gamma \)-regulated nuclease and helicase activities, and the ability to facilitate formation of the active species in HR—which is a nucleoprotein filament formed by RecA recombination on ssDNA.

RecBCD has only a few homologs in bacteria, and no eukaryotic analog of this helicase/nuclease has been identified. In some genetic backgrounds, most RecBCD activities can be performed by the RecQ helicase acting together with RecF nuclease to process the ssDNA ends and the RecFOR proteins, which facilitate loading of RecA recombination on to ssDNA produced by RecQ/RecF activities (reviewed in Reference 41). The RecQ helicases have also been implicated at subsequent steps in HR, which require caretaker functions that involve disruption of the nucleoprotein filaments as well as stabilization of the invading strand, and that facilitate branch migration and resolution of Holliday junctions (reviewed in Reference 9). Notably, whereas all of these activities can be carried out by a single bacterial RecQ helicase, its five mammalian counterparts evolved to perform more specialized tasks.

**Nucleotide Hydrolysis and Translocation**

**ATP (NTP) Hydrolysis**

Most helicases display preference for ATP as an energy source. However, some of these enzymes display a broad specificity for nucleotides or deoxynucleotides, whereas some prefer different nucleotides, such as for example phage T7 helicase, which selectively hydrolyzes dTTPs (42). Although the precise mechanisms for individual helicases differ, all these enzymes generate force through conformational changes within nucleotide binding folds, which allow them to perform active work. Not surprisingly, the highest structural conservation among helicases is observed between the nucleotide-binding and hydrolysis cores. This structure uses Walker A and B motifs (Walker P-loop) observed between the nucleotide-binding and hydrolysis cores. Notably, whereas all of these activities can be carried out by a single bacterial RecQ helicase, its five mammalian counterparts evolved to perform more specialized tasks.

**ATP binding/hydrolysis fold of RecA (see Fig. 1)**

ATP synthase (43) provides an environment where Mg\(^{2+}\) ion is chelated between the phosphate of ATP through its amino group while the threonine (or serine) coordinates the Mg\(^{2+}\) ion. The Walker B box (motif II) contains a characteristic DEAH, DExH, or DExx sequence. The aspartic acid from this motif coordinates the Mg\(^{2+}\) ion through its carboxyl group, and the glutamate serves to activate the water molecule for nucleophilic attack on the \( \gamma \)-phosphate. Additionally, residues in all eight SF1 signature motifs form important contacts with the ATP-Mg complex, and residues in motifs Ia, III, V, and VI form contacts with DNA. All the motifs work in a concerted manner to transduce a conformational change between the nucleotide and DNA binding sites (reviewed in Reference 1).

Hexameric helicases in SF 3-6 also employ the Walker A and B boxes for NTP (or dNTP) hydrolysis (Reviewed in Reference 18). The ATP binding sites are formed with one subunit that provides the Walker A and B motifs and the neighboring subunit providing an arginine finger (Fig. 2). Crystal structures of the E1 helicase from bovine papilloma virus (SF3) (19), and of the gp4 helicase from the T7 bacteriophage (SF4) (44), show that these sites are asymmetric, and they have different affinities for the nucleotide. These sites have been described as ATP binding, ADP + Pi, and empty (apo). The structural data support a sequential ATP hydrolysis mechanism in which ATP binding to the empty site induces a conformational change in the helicase. ATP is then hydrolyzed to create the ADP + Pi state, and then the nucleotide and inorganic phosphate is released, which results once again in an empty state that permits a new molecule of ATP to bind.

**Translocation**

The mechanism(s) employed by helicases to translocate along nucleic acid lattices were the subject of a heated debate over the past decade. In the last few years, however, the models converged, which resulted in the general acceptance of a so-called “inchworm” mechanism (reviewed in Reference 45) (Fig. 3). Evidence to support this mechanism has been found for both SF1 and SF2 helicases and the related DNA translocases of the Swi2/Snf2 family, with the major difference being that SF1 helicases bind to ssDNA through stacking interactions between DNA bases and aromatic residues of the helicase (46) (Fig. 3a). However, SF2 helicases bind the phosphate backbone of a DNA molecule and translocate along the backbone, avoiding significant base interactions (47) (Fig. 3b). While “inning” forward along ssDNA through the use of NTP-induced conformational changes, the helicase unwinds duplex DNA by displacing the complementary strand with the help of one of the structural tools described above. The inchworm mechanism requires that the helicase maintains at least two DNA binding sites that can alternate between two states, one being tightly bound and other a loosely bound state. The alternating DNA-binding sites may belong to different subunits of a helicase ring (as most profoundly illustrated for BPV E1 helicase (19)) or to a single monomer as is the case with SF1 and SF2 helicases. The crystal structures of prototypical SF1 helicases, PcrA (46), and UvD (16) combined with vast biochemical data support this model of translocation and show a single strand of DNA on which the helicase translocates, spanning the region between the RecA-like folds and bound simultaneously to both folds (46). The helicase binds to ssDNA, and after binding to ATP, it undergoes a major conformational change resulting in closing of the cleft between the two RecA-like folds, with the leading fold binding tightly to DNA, and the trailing fold easing its grip. The helicase takes a step forward on the hydrolysis of ATP or on the release of the
DNA Helicases, Chemistry and Mechanisms of

(a) "Inchworm" translocation by the SF1 helicases whereby the helicase binds to and interacts with the bases of ssDNA. In the absence of ATP, the cleft located between the RecA-like folds is apart, and the folds are separated by a single DNA base. After binding to ATP, the cleft closes, and the trailing RecA fold moves one nucleotide closer to the leading fold, which results in displacement of a single base.

(b) SF2 helicases translocate along a nucleic acid lattice through extensive contacts with the DNA backbone also using the inchworm mechanism. Here ATP binding and hydrolysis induces a conformational change that involves closing of the cleft between the RecA-like folds.

(c) Hexameric helicases are believed to translocate by an escort mechanism in which the DNA binding loops located within the center channels of the ring structure for each subunit are linked through the set of helicase signature motifs to the ATP binding pockets located on the outside of the structure. In this staircase mechanism, the loops associated with ATP are located at the top of the channel, bind to the next incoming base, and escort that base through the channel. The loops located in the intermediate positions along the staircase are bound to ADP, and the empty pocket is located at the exit position, where after binding to ATP the loop can return to the top and bind the next entering base.

The mechanism of translocation employed by the SF3, hexameric helicase E1 uses the asymmetry of the ATP-binding sites to escort DNA through the central channel of the ring structure (19) (Fig. 3c). Individual subunits within the hexamer have DNA binding loops that protrude into the central channel. The ATP-bound subunits place their loops at the top of the channel and bind to the DNA backbone of the next entering nucleotide. These subunits maintain contact with this nucleotide and escort the base through the central channel, as ATP is hydrolyzed to ADP + P and ADP and P are released. The loops associated with the empty site are located at the bottom of the channel (exit) and on binding a new ATP, return to the top of the channel and bind a new DNA base for escorting.

The step size of a helicase is one of the more contentious characteristics because of the various ways it can be defined. It may be determined in terms of efficiency, which refers to the number of base pairs unwound or translocated per binding and hydrolysis of 1 NTP molecule. The physical step size of a helicase can be determined based on its ability to traverse discontinuities in the DNA lattice, whereas "kinetic step size" refers to a frequency of kinetically rate-limiting step that occurs during DNA unwinding (48). The three step sizes are rarely identical for a given enzyme. Although SF1 helicases hydrolyze one ATP molecule per one nucleotide they translocate (49), or per one base pair of dsDNA unwound (16), their kinetic step size is much larger (50). Integration into larger macromolecular ensembles may alter the perceived step size even more. For example, when a typical SF1 helicase RecB is incorporated into RecBC enzyme, it displays a physical step size of 23 nucleotides (51).

Depending on the method of analysis, the kinetic step size of SF2 helicase NS3 was estimated as 18 base pairs (kinetic and physical step size on RNA) (52), 11 base pairs (also correlated to a measured physical step size and binding site size) (53), 9 base pairs on DNA (54), and 3–4 base pairs (55). The difference in these step sizes may reflect the two nucleic-acid-binding domains found in SF2 helicases. One RecA-like fold reaches out and binds to duplex nucleic acids about 11 base pairs ahead while the trailing domain follows unwinding the duplex behind the leading domain (53) in discrete steps. The larger apparent 18 base pair step size may reflect either the max step size or the low processivity of NS3 helicase. Recent single-molecule analysis of NS3 helicase connected its kinetic step size of about 3–4 base pairs to smaller substeps associated with ATP hydrolysis. In these experiments, NS3 unwound duplex DNA in three base pair steps by a spring-loaded mechanism, in which the lead binding fold of NS3 moves forward one base pair per ATP hydrolyzed. After three successive steps, enough tension has built up to cause the trailing fold to snap forward, which leads to unwinding and release of three nucleotides (55).

Although SF1 and SF2 enzymes can translocate and some can unwind DNA as monomers, these enzymes show evidence of increased processivity and helicase activity when they act as cooperating monomers, dimers, or oligomers. For example,
the SF1 helicase UvD and the SF2 helicase NS3 both display low processivity at low concentrations; however, at high concentrations or under conditions when more than one helicase can bind in succession to a single strand of DNA, their processivity is dramatically increased (54, 56). The exact mechanisms of duplex unwinding and therefore the nature of oligomerization requirement for SF1 and SF2 helicases remains disputed. According to one model, cooperation between the independent monomers achieves two goals: 1) it allows for exertion of greater forces and 2) it increases the probability of the helicase to take a step forward versus falling off DNA and hence its processivity. An alternative theory is that these helicases form specific higher-order complexes with each other resulting in dimers (UvD) (57) or oligomers (NS3) (58). These structures are necessary for efficient unwinding and may result in activation of latent helicase activity (reviewed in Reference 59). This activation may also be achieved through the protein-protein interactions within respective macromolecular machines. Although evidence supports both models, together these results indicate that protein-protein interactions, and assembly into higher-order structures play an integral role in modulating helicase activity, and they serve to stabilize the helicase when bound to a nucleic acid lattice.

Tools and Techniques Used to Characterize Helicases

The biomedical importance of DNA helicases and their intriguing molecular mechanisms prompted continuous development of assays to monitor interactions between these enzymes and their nucleic acid substrates, translocation, NTP hydrolysis, duplex separation, and remodeling of nucleoprotein complexes. DNA binding and duplex unwinding by helicases is conventionally monitored by gel electrophoresis-based assays. In these assays, DNA substrates are produced by annealing two partially complementary oligonucleotides, one of which is radioactively labeled at the 5' end (Fig. 4a). Several DNA substrates that mimic cellular DNA repair intermediates can be produced in this manner. These substrates can be used to determine the polarity exhibited by a particular helicase and to identify its preferred or cognate substrate. In the binding assays, protein-DNA complexes are separated from the free radioactively labeled substrate taking advantage of the difference in electrophoretic mobility through polyacrylamide gels under native conditions (Electrophoretic Mobility Shift Assays). In the unwinding assay, the DNA substrate is incubated with the helicase, and DNA unwinding is initiated by the addition of ATP. After quenching the reaction, the displaced oligonucleotide can be separated from the intact substrate by electrophoresis through nondenaturing polyacrylamide gel. The relative quantities of the radioactive label in the substrate and product bands can be used to quantify binding and unwinding.

These traditional assays have been adapted to monitor duplex separation by following fluorescence resonance energy transfer (FRET), which is a powerful spectroscopic technique for measuring distances in the 0.7-75 Å range (see Fluorescence Resonance Energy Transfer (FRET) for Proteins for more information). In this assay, two fluorophores, a donor, and an acceptor are incorporated into complementary strands of a DNA duplex (Fig 4b). When the duplex is intact and the donor and acceptor are in close proximity to each other, the excitation energy of the donor fluorophore can be transferred to the acceptor fluorophore (whose emission frequency should correspond to the detection of the donor) via an induced dipole-dipole interaction. As the helicase separates labeled DNA strands, there is a decrease in fluorescence by the acceptor fluorophore and a simultaneous increase in fluorescence by the donor fluorophore.

Both, gel- and FRET-based techniques are discontinuous methods, which means that only complete unwinding of each substrate molecule results in a measurable signal. For the processive enzymes, it is important to monitor duplex unwinding as it occurs. Continuous monitoring of a helicase activity can be achieved through the use of a fluorescence-based helicase assay that exploits the use of several DNA-intercalating dyes that display a measurable increase in fluorescence intensity when bound to ssDNA relative to the unbound state or to the dsDNA-bound form (60) (Fig. 4d). As the DNA molecule is unwound, fluorescence decreases as ssDNA products form and the dye is released.

Translocation assays

Duplex separation and disassembly of the nucleoprotein complexes directly stems from unidirectional translocation of a helicase along DNA. Most commonly, directional translocation is monitored by following displacement of streptavidin from a radiolabeled DNA molecule biotinylated near one of the termini (61) (Fig. 4c). Translocation by DNA helicase results in accelerated dissociation of the streptavidin from biotin. The products of such reaction are then resolved using an electrophoretic mobility shift assay that separates the free oligonucleotide from streptavidin-bound DNA. More quantitatively, translocation can be monitored by following a change in the fluorescence of a dye incorporated into a DNA substrate (62) (Fig. 4c).

Single-molecule techniques

The complex nature of DNA helicase reactions and mechanisms necessitated the development of multiple techniques that allow direct observation of individual protein-DNA complexes. Most groundbreaking insights into helicase mechanisms from single-molecule techniques have come from the use of total internal reflection microscopy (TIRM), magnetic tweezers, and laser tweezers (optical trap) (63). In these approaches, DNA is immobilized either on a surface of a microscope slide, a glass capillary, and/or a polystyrene or magnetic bead using biotin-streptavidin or antibody-antigen interaction.

TIRM allows detection of hundreds of single molecules simultaneously with high sensitivity (Fig. 4d). These single-molecule assays are usually an extension of the FRET-based methods described above. The advantage of this approach over the bulk studies is that single events can be captured, recorded, and visualized as they occur, in contrast to the bulk assays that monitor the ensemble average for all molecules in solution and can obscure key steps in DNA manipulation processes.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Methods for the helicase analyses. (a) Observation of strand separation through the use of discontinuous helicase assays. A radioactive label ($^{32}\text{P}$) can be used to label one strand and monitor unwinding activity caused by the difference in electrophoretic mobility of the duplex substrate compared with the ssDNA product. The same process can be monitored using a FRET pair of fluorophores, donor (D), and acceptor (A). (b) Continuous unwinding assays can be monitored fluorescently through the use of an intercalating dye that has increased fluorescence when intercalated into duplex DNA. (c) Deposition of transcription DNA substrates and their products. Translocation can be observed by the ability of the helicase to displace streptavidin from a biotinylated oligonucleotide. Other methods include observing the change in fluorescence of a fluorophore incorporated into the DNA lattice or observing the displacement of a triplex forming oligonucleotide by a dsDNA translocating enzyme. (d) For TIRM, single-DNA substrates are attached to a slide through a streptavidin/biotin interaction. The substrates are labeled with a FRET pair of fluorophores. Single unwinding events can be observed through monitoring fluorescence decrease of the acceptor fluorophore and simultaneous increase in the donor fluorophore. (e) Magnetic tweezers can be used to observe helicase binding and unwinding properties through changes in the force-extension properties of DNA. DNA is attached to surface at one end, and tethered to a magnetic bead at the other. The magnetic bead can be manipulated through the application of a magnetic field, and changes in DNA can be monitored. (f) Polystyrene beads attached to one or both ends of a DNA molecule can be trapped by the laser tweezers. Unwinding can be monitored through displacement of an intercalating dye. This technique can also be used to monitor translocation of a fluorescently labeled enzyme.

In magnetic tweezers, one end of the DNA molecule is attached to the surface of the glass capillary and the other end is attached to the magnetic bead (Fig. 4e). The bead can be "lifted" from the surface by applying the electromagnetic field. The magnetic field exerts a constant force, so the resulting extension of the molecule can be measured as a function of the applied force. Using this approach, one can monitor the displacement of the bead in all three dimensions (displacement in the x and y dimensions is monitored directly, whereas the z direction can be estimated from the refraction pattern). DNA helicase activity can be observed from the difference in the force-extension properties between ssDNA and dsDNA.

A rather informative single-molecule technique combines optical trapping and fluorescence microscopy. Here, a long DNA molecule is attached to a polystyrene bead, which can be captured by the laser tweezers (a device that allows trapping of a small semi-transparent object in the focus of a laser beam) (Fig. 4f). Translocation of a helicase along the DNA molecule extended by a flow of buffer can be monitored after displacement of the intercalating fluorescence dye or directly if the fluorescence label is introduced in the helicase. Chemically modified nucleic acids can be used in each technique described above to deduce the mechanisms of helicase translocation and unwinding. Discontinuities can be introduced into the nucleic backbone through nicks and polyglycol linkers to determine whether the continuity of the phosphate backbone is necessary for translocation by the helicase (64). By varying the size of the discontinuity, one can also determine the physical step size of the helicase by testing how large of a gap the enzyme can tolerate (51). Additionally, the nucleic acid moieties can be removed resulting in abasic sites, which may impede the translocation (65). The electrostatics can be interrupted by incorporating a peptide nucleic acid moiety into the sequence or by introducing a methylphosphonate into the backbone (66).
Future Research and Remaining Questions

In addition to identifying new helicases and biochemically characterizing their properties, a major point of emphasis in the future is how these helicases are incorporated into larger complexes and macromolecular machineries involved in DNA metabolism. The protein–protein interactions within these assemblies likely play integral roles in modulating helicase properties. One of the more intriguing questions that remains to be addressed involves the structural characteristics responsible for defining a helicases polarity and translocation lattice.

References


© 2008, John Wiley & Sons, Inc.


Further Reading


See Also

DNA-binding Molecular Motors, Chemistry of
DNA Damage Repair, Chemistry of
Protein-nucleic Acid Interactions
Nucleic Acid Metabolism
Fluorescence Resonance Energy Transfer (FRET) for Proteins
Genetic recombination is classically manifest as the reassortment of genes during meiosis, where extensive pairing of homologous chromosomes is associated with “crossover” between them to create new derivative hybrid chromosomes. The associated enzymatic pathway, homologous recombination (HR), is now known to execute such exchange not only during meiosis but also between homologous sequences in mitotic cells, for example between sister chromatids. Nonhomologous recombination (NHR), also called illegitimate recombination, was recognized initially and defined as genetic recombination that failed to show this canonical feature of HR, i.e., hybrid derivative chromosomes that lacked extended homology at the junction between parents. Instead, junctions often showed 1-10 bp segments of “microhomology” too short to be used by HR enzymes. Such events seemed to defy the logic and benefits of HR in a chemically distinct alternative reaction that was, on its face, detrimental. In this review, the processes that underlie NHR are placed in context as DNA repair pathways that are in fact associated with genome preservation, but that can secondarily give rise to chromosome rearrangements and other mutations. The biochemical requirements of these reactions are discussed as well as current understanding of how the responsible enzymes have been adapted to meet these requirements.

Biological Background

As might be expected from the fact that both HR and NHR represent new joints between previously separate DNA duplexes, an intimate relationship exists between recombination and DNA double-strand breaks (DSBs) (Fig. 2a). Processes that give rise to DSBs not only underlie meiotic and mitotic HR, but also they predispose cells to NHR. Conversely, deficiency in cellular components required for both HR and NHR causes hypersensitivity to DSBs. Indeed, it may be most accurate from an evolutionary perspective to consider HR and NHR foremost as alternative mechanisms for silent repair of DSBs, with recombination as a secondary consequence. It is thus critical to consider the cell cycle-dependent interplay of homologous and nonhomologous mechanisms at a DSB as the stage on which NHR plays out.

Interplay between homologous and nonhomologous DSB repair

Cells constantly suffer DNA lesions that degrade the genome. Some DSBs are created deliberately and by specific mechanisms (see below), but many result from oxidative, radiation, or physical stresses that occur more randomly (Fig. 2b). When acting as a DSB repair pathway, HR uses a homologous donor chromosome as a template to copy the genetic information required to bridge the broken ends without altering the template itself (Fig. 2c). Such a second intact copy of a chromosome is present in diploid cells, although in large, complex, and repetitive genomes, finding this potential donor is not trivial because chromosome pairs are not associated. A nonideal donor exists when one of a pair of sister chromatids is broken, where
Nonhomologous Recombination

Figure 1  Microhomology use in NHR. (a) Demonstration of microhomology, which is evident as a sequence common to both chromosome substrates (boxed). Note that the position of initiating DSBs cannot be inferred from the product. (b) Possible mechanisms for generating microhomology. If DSB overhangs encompass a microhomology, it can be rejoined directly by NHEJ. Internal microhomologies must be exposed, either by unwinding, as suggested for DNA-PKcs (1), or by 5′ resection, as suggested for deletion MMEJ (2).

Figure 2  Interplay between HR and NHR. (a) When a DSB occurs (top), it can be resected and used to copy a donor chromosome (red) during HR (left) or rejoined directly (right), sometimes inaccurately (jagged line). HR and NHR lead to recombination by crossover resolution and engagement of a different DSB (green), respectively (bottom). (b) A timeline shows a period during which NHEJ may be iteratively attempted before DSB resection; after which, HR is attempted. Alternative pathways (below) might lead to rejoining, but these are often less efficient and uncommon when the above pathways are functioning normally.

the other sister is both physically associated and strictly homologous. This special relationship is strengthened by understanding that replication itself can lead to DSB formation. It is thus both predictable and observed that HR is a principal repair pathway in S or G2/M cell cycle stages after replication but before cell division (6, 7). Given that nonhomologous DSB repair processes are, by definition, not HR, it is in turn predictable and observed that such repair predominates at DSBs occurring randomly outside of replication (6, 7). But what is DSB repair in the absence of a homologous donor? Stated simply, nonhomologous repair is the direct rejoining of DSB ends (Fig. 2a) (5). The details of these mechanisms are addressed below, but simply recognizing that they are executed without an information template correctly predicts that they have a higher error rate. However, when HR is impractical or impossible because of inaccessibility of a homologous template, even error-prone rejoining of DSB ends will yield a relative preservation of the genome as compared with continuing cell division with a broken chromosome. Also, repair of simple DSBs by direct rejoining can be >99% accurate (8, 9), so that nonhomologous repair is often genetically silent.

A major difference between homologous and nonhomologous DSB repair is that HR begins with extensive resection of the 5′ DSB strands, whereas NHR is comparatively duplex preserving. Mounting evidence indicates that the 5′ resection is the regulated step in DSB repair pathway “choice” because it is an essentially irreversible commitment to HR (6, 8). Resection is under direct cell-cycle control, being activated by cyclin-dependent kinases, at least in yeast, and therefore largely inactive in G1 cells (6). Even in replicating cells, resection is delayed, approximately an hour in yeast, before it begins in earnest (10). The basis of these effects is poorly understood, mainly because the mechanism(s) of resection remain enigmatic. Regardless, their net effect is to provide a period during which nonhomologous repair of a DSB might be attempted, with transition to homologous repair of persistent breaks at a pace that is cell-stage dependent.
Nonhomologous repair (15, 16).

NHR events. Potential complexity of the processes giving rise to observed stable resolution (13). These cycles underscore the potential complexity of the processes giving rise to observed NHR events.

Telomeres and breakage–fusion–bridge cycles

The ends of linear chromosomes are a type of “DSB” worth specific consideration. McClintock first recognized that specialized structures called telomeres protect chromosome ends (13, 14). A detailed description of this structure is beyond the scope of this article, but it involves a repeating DNA sequence and multiple proteins that bind to it, including DSB repair proteins (14). The key importance is that when this structure fails, because of loss of telomere proteins or degradation of the DNA repeat, chromosome ends become a DSB target for NHR mechanisms that can fuse sister chromatids by direct joining. The resulting dicentric chromosome is destined to be broken during mitosis if pulled toward opposite spindle poles, which leads to SSBs that might again lead to chromosome fusion, ongoing “breakage–fusion–bridge” cycles, and NHR before ultimate stable resolution (Fig. 3a) (13). These cycles underscore the potential complexity of the processes giving rise to observed NHR events.

V(D)J recombination

Another important context for understanding NHR is a series of genome manipulations in the developing vertebrate immune system where nonhomologous repair mechanisms are co-opted to create deliberate mitotic chromosome rearrangements. In V(D)J recombination, specific variable, diversity, and joining sequences from a collection of such segments in the germline genome are juxtaposed in recombinant chromosomes to create functional immunoglobulin and T-cell receptor genes (Fig. 3a) (15, 16). Both by the reassortment of different coding segments and by maximizing the error-prone nature of nonhomologous junctions, a large repertoire of antigen receptor genes results. V(D)J recombination is initiated by a mechanism-specific recombinase, the RAG1/2 proteins, but it is completed by the same nonhomologous repair mechanisms used in genome preservation, which highlights the potential duplicity of these mechanisms. Although less well defined, immunoglobulin class switch recombination also seem to require nonhomologous repair (15, 16).

Chromosomal instability in cancer

Nearly all cancer cells show deranged chromosomes manifest as translocations, deletions, and amplifications. Some rearrangements likely develop as a consequence of aberrant cell growth, but others are unquestionably causative. In one pattern, which is typical of lymphoid malignancies, specific cancers are associated with recurring, usually reciprocal, translocations between nonhomologous chromosomes. Translocation t(9;22) in chronic myeloid leukemia is a classic example that allows treatment with inhibitors that target the product of the resulting BCR–ABL fusion gene (17). The other main pattern, which is typical of epithelial malignancies, shows massively deranged genomes with multiple, usually nonreciprocal, rearrangements. Extensive efforts continue to identify oncogenes and tumor suppressor genes whose function is affected by these rearrangements. Junctions of rearranged chromosomes in cancer frequently show microhomology in the absence of extended homology and are therefore inferred to originate by direct rejoining/NHR. The lesions and mechanisms that yield these rearrangements are undoubtedly multiple, including both telomere-dependent and telomere-independent events. Some rearrangements represent errors of the V(D)J process because they occur in cells and loci undergoing this targeted recombination (Fig. 3b). This issue provides another caution regarding the multiplicity of NHR mechanisms, because even these errors could represent use of inappropriate loci in the V(D)J reaction, joining of RAG1/2-generated and spontaneous DSBs, or even DSB-independent transposition between chromosomes (15). Other NHR presumably originates from two independent DSBs (Fig. 2a), but the final sequence might have little to do with the configuration of the inciting damage, as elegantly demonstrated in a yeast model (18).

Chemistry of DNA End Joining

The chemical transactions of DNA end joining are predictably similar to other DNA repair pathways and even to replication in that all of these processes manipulate base-pairing and phosphodiester bonds to create intact duplex molecules. However, the direct repair of DSBs presents biochemical challenges unique to these substrates. In this section, I consider both the general and the specific chemical requirements of nonhomologous DSB repair, how they are manifested in the two main pathways of repair, and the sometimes unusual properties of key enzymes that determine their function in these pathways.

Chemical steps in strand break repair

It is important to recognize that although NHR and HR share a substrate, they are mechanistically very different. It is more useful to compare NHR with the direct repair of single-strand breaks (SSBs), which is a subset of the base-excision repair (BER) pathway (Fig. 4a) (19). Three reaction classes might be required for such repair, depending on the specific nature of the break. All DNA repair culminates in ligation of a strand nick bearing 3' OH and 5' P termini. To arrive at that point, it might first be necessary to resynthesize bases missing as a result of the DNA damage. Both polymerization and ligation might also depend on the removal or reversal of blocking nucleotide fragments residual on the break termini. It might include lyase-dependent removal of 5' deoxyribosephosphate deoxyribosephosphate.
Nonhomologous Recombination

Figure 3. Special mechanisms of NHR. (a) Telomere failure unmasks chromosome ends for fusion by NHEJ, which leads to potentially iterative cycles of breakage and fusion before ultimate stable telomere resolution. Oval, centromere; rectangle, telomere. (b) Immune receptor loci contain multiple V, J, and sometimes D segments with signal sequences (triangles) that are cleaved by RAG1/2 and subjected to NHEJ to generate recombined products. Engagement of an inappropriate chromosome can lead to RAG1/2-dependent chromosomal rearrangement.

Nonhomologous end joining

The term “nonhomologous end joining” (NHEJ) initially referred generically to direct rejoining of DSBs, but it is now most commonly used to refer to a specific and dominant pathway of nonhomologous repair defined by both DNA features and protein components (5). Regarding the DNA, NHEJ typically entails little or no nucleotide loss from DSB ends. In the strictest manifestation, which is typified by the behavior of budding yeast NHEJ acts almost exclusively by annealing DSB overhangs (9, 20). Because such overhangs were often paired in the original duplex, overhang joining tends toward accurate repair directly analogous to BER (Fig. 4a). This restriction to overhangs is apparent relaxed in other species, but current evidence nonetheless distinguishes NHEJ and more extensive nucleotide loss (see below). Regarding proteins, NHEJ requires both specific structural proteins to bind and bridge ends and enzymes to catalyze reactions similar to BER (Fig. 4a) (5). Unlike HR, whose genes were first identified by screens in yeast and bacteria, many proteins that catalyze NHEJ were discovered in mammals by observations that their deficiency causes sensitivity to DSBs and V(D)J impairment (4, 5, 16). This observation led to the common notion that mammals and single-cell organisms predominantly use NHEJ and HR, respectively. This notion is misleading, because all eukaryotes maintain complex HR and NHEJ pathways that are highly active in the correct cell states (6, 7).

NHEJ structural proteins include Ku, which is a protein preserved in all organisms that possess NHEJ and whose utilization provides the most specific current definition of this pathway (5). Ku is a heterodimer of the Ku70 and Ku80 proteins in eukaryotes and a homodimer in prokaryotes but in all organisms forms a ring that provides initial lesion recognition by passage of the DSB end through the ring (21). Interestingly, although Ku likely binds all DSBs, it is only required for NHEJ, not for HR. The specific roles of Ku in NHEJ in higher eukaryotes center on its identity as the principal DNA binding component of the DNA-dependent protein kinase (DNA-PK), whose catalytic subunit DNA-PKcs is also required for efficient NHEJ (5). As a kinase activated by DSB ends, DNA-PKcs contributes to NHEJ in part by phosphorylating NHEJ proteins and regulating the ensuing cascade of events. However, in vitro evidence also supports a primary structural role of DNA-PKcs as part of an end bridging complex (22, 23), although cellular evidence that supports this function is needed. Indeed, one key action of activated DNA-PKcs is autophosphorylation, which in some fashion seems to release the DSB ends for access by downstream enzymes (24). Despite this central role, neither lower eukaryotes nor bacteria are known to require a DSB-dependent kinase for NHEJ, which suggests that such regulation was imposed on...
Nonhomologous Recombination

Figure 4  Chemistry of strand break repair. (a) The steps of SSB and DSB rejoining are highly analogous, with an added requirement for end bridging in NHEJ. For simplicity, only select mammalian proteins are indicated. Curved line, terminal damage; red arrow, polymerization. (b) The end-bridging step is realized in unusual ways by NHEJ polymerases (Pol) and ligase (Lig). At 3′ overhangs, polymerases tolerate an incomplete primer–template pair, in part using special protein motifs (cyan loop). Similarly, DNA ligase IV can ligate (carat) across from strand discontinuities. In some yeast, an additional NHEJ requirement exists for the DSB-binding Mre11–Rad50–Xrs2 (MRX) complex (9), which may account for the bridging function in place of DNA-PKcs (26). Consistently, current best evidence does not suggest a role for the analogous Mre11–Rad50–Nbs1 (MRN) complex in vertebrate NHEJ (27), but this is poorly explored because MRN is essential for viability.

NHEJ enzymes include foremost the requisite DNA ligase. In eukaryotes, DNA ligase IV (Lig4) is both strictly required for Ku-dependent NHEJ and largely if not entirely dedicated to this pathway (5). In prokaryotes, the principal NHEJ ligase, called LigD in mycobacteria, is present in an operon with Ku, which again demonstrates a coupling of Ku with a specific NHEJ ligase (28). However, the ligase presents a point of divergence between kingdoms in that LigD is no more related to Lig4 than to other ATP-dependent ligases. LigD does not act alone in support of NHEJ but it interacts via tandem C-terminal BRCT domains with a coiled-coil domain of a protein called XRCC4 (29). XRCC4 itself binds DNA (30) and so is arguably a structural protein that acts as a scaffold between the ligase and the DNA but because of the strength of its association is accurately viewed as a subunit of DNA ligase IV. Strikingly, yet another protein associates with Lig4::XRCC4 called XLF/Cernunnos that is predicted to be structurally similar and evolutionarily related to XRCC4 (31).

Other NHEJ enzymes execute end processing before ligation and are generally less well defined and less strictly required. Nucleases include the protein Artemis in higher eukaryotes, which executes unique NHEJ functions in V(D)J recombination (32) and the slower repair of approximately 10% of irradiation-induced DSBs that correlate with cytotoxicity (33). However, Artemis is absent from most lower eukaryotes, perhaps because of its intimate association with DNA-PKcs and V(D)J recombination (32), and even in organisms that possess it, Artemis is unlikely to account for all end trimming. Thus, some NHEJ nucleases are yet to be described. NHEJ polymerases include select members of the Pol X family of DNA polymerases typified by the BER enzyme Pol β, specifically Pol4 in yeast (20) and terminal deoxynucleotidyl transferase (TdT, used only in V(D)J recombination), Pol λ, and Pol µ in mammals (34). However, like nucleases, these proteins cannot account for all NHEJ polymerization, as demonstrated clearly in yeast (20). Also, deficiency of even both Pol λ and Pol µ leads to only a mild phenotype manifested as altered V(D)J joint spectra but not frank sensitivity to DSBs (35). In prokaryotes, NHEJ end processing remarkably includes two protein domains often fused to the ligase, although these can function independently (28, 36). LigD in particular is a polyfunctional protein that contains polymerase, nuclease, and ligase domains in tandem, so that mycobacterial NHEJ can be transferred with just two polypeptides (37). As with the ligase, the prokaryotic nuclease and polymerase are not particularly related to Artemis or Pol X polymerases (28, 36).
Microhomology-mediated end joining

In early experiments it was observed that DSBs were sometimes repaired by an erroneous mechanism associated with nucleotide loss from the ends (38). In much literature this repair was attributed to NHEJ in its generic sense, but more recent work has continued to refine both NHEJ and this alternative form of repair. A major clue that the latter was a distinct mode of repair was the general observation that joints formed in NHEJ mutants showed more extensive nucleotide loss. This phenomenon has been carefully explored, and data support the notion that Ku-dependent NHEJ is primarily responsible for accurate rejoining (39, 40). One name given to the alternative rejoining pathway is “NHEJ” (41), but this is confusing given the more common convention of using NHEJ to describe Ku-dependent joining. The designation microhomology-mediated end joining (MMEJ) was applied to Ku-independent repair in budding yeast (42). Similar joining, which is typified by microhomology use and substantial nucleotide loss, can also be observed in insects and plants (43, 44), and until additional information suggests a better classification, MMEJ seems a suitable name for all. MMEJ is poorly explored in bacteria.

The mechanisms of MMEJ are not well established and are likely multiple. A unifying feature in all systems is that MMEJ is repressed by Ku in mammals, developing data suggest that poly-ADP-ribose polymerase and DNA ligase III might participate in one subpathway (41). Because these factors are typically associated with BER, the analogy drawn above between the direct repair of SSBs and DSBs might in fact be realized in competing repair pathways. More detailed explorations of MMEJ have been performed in yeast, where a different type of repair was revealed that is not yet possible to describe as a simple linear pathway (2, 42). In Saccharomyces cerevisiae, MMEJ shows only a partial dependence on any factor (42). Interestingly, two of these factors are the NHEJ enzymes Dnl4 and Pol4, even in the absence of Ku. Another dependence is on the 3′→5′ flap nuclease Rad1-Rad10, which is not implicated in NHEJ but does function in HR and, more relevantly, in single-strand annealing (SSA) (3, 4). SSA is a form of repair in which larger internal direct repeats are annealed. SSA is not typically described as NHEJ because these repeats are too big to be considered microhomologies and because it uses the HR protein Rad52 to facilitate annealing. Although Rad52 is not obviously required for MMEJ in S. cerevisiae (42), its homologue in Schizosaccharomyces pombe is, which leads to the designation of MMEJ as “micro-SSA” (2). This finding is compelling as it invokes the most consistent apparent requirement of deletion MMEJ—namely, that strand resection must expose microhomologous sequences for annealing, which is probably common with HR (Fig. 2b, 2, 45). Importantly, NHEJ might sometimes use internal microhomologies close to the DSB ends through the action of proteins such as DNA-PKcs, which are distinct from 5′→3′ resection and MMEJ (Fig. 2b, 1). A fundamental question remains whether MMEJ is a distinct mechanism of DNA repair or simply a different manifestation of previously described pathways. Yeast MMEJ is markedly inefficient and largely irrelevant to basal DNA repair (40, 43), in contrast to HR and NHEJ (6, 8). However, repair consistent with MMEJ can be very efficient in higher eukaryotes, so that although the quality changes, rates of end joining do not necessarily decrease markedly with NHEJ mutation (40, 42). Moreover, even an inefficient pathway might be of relevance to NHR, so it is important to consider which mechanisms actually catalyze rejoining of inappropriate DSB pairs. NHEJ can efficiently mediate nonhomologous integration of transformed fragments in some fungi (46). A low, it was found that some spontaneous yeast nonreciprocal rearrangements depend on NHEJ (47) and that yeast NHEJ can mediate reciprocal translocations between two induced DSBs (48). It is now known how to disable MMEJ specifically, but residual chromosome rearrangement observed in NHEJ mutants indicates that at least one other mechanism, presumably MMEJ, is also at play (47, 48). It is not yet clear how often mammalian NHEJ might execute spontaneous NHR, but it is clearly a net caretaker of the genome whose absence leads to increased chromosome rearrangements by alternative repair mechanisms (49).

DNA ligase IV

Clearly, Lig4 is specifically suited to NHEJ, but why? In part, it is because the DNA ligase IV complex can engage the NHEJ structural machinery by specific protein contacts (25). However, the enzymatic function of Lig4 might also be optimized for NHEJ, because transferring its BRCT domains to a different ligase did not transfer NHEJ capacity (50). This function is not manifested by a notably greater ability of Lig4 to ligate compatible overhangs. Instead, Lig4 has a unique ability to ligate ends that do not align to a simple nick but are incompatible or have a gap (51). Unlike most ligases, Lig4 apparently can bring the 3′→5′ P′ and 5′→3′ P′ into position for catalysis anyway, which requires tolerance for non-duplex DNA (Fig. 4b). This activity would have the tendency to promote rejoining of complex ends and therefore chromosome stability at the expense of potentially less deleterious local mutations. Concerns are that others have not observed markedly different ligation fidelity of Lig4, but importantly this was measured nicks, not DSBs (52). Also, my laboratory has examined many NHEJ substrates in yeast genetic assays and has rarely observed outcomes consistent with promiscuous ligation (8, 20). Other NHEJ proteins might also influence the different types of outcomes. XLF/Cernunnos depletion from extract-mediated NHEJ reactions had a much greater effect on joining of complex as opposed to simple DSB substrates (53). This effect might imply a role for XLF/Cernunnos in promoting either end processing or possibly promiscuous ligation by Lig4 (again, though, XNL), the yeast XLF homologue, is equally required for both simple and complex end configurations (our unpublished results). It is still the early days for these studies, and it is of great interest to determine how important the catalytic properties of the DNA ligase IV complex are beyond its recruitment by NHEJ structural proteins.

Pol X NHEJ polymerases

Strong evidence points to the acquisition of special enzymatic properties in the Pol X family of NHEJ polymerases that allow them to deal with the limiting substrates inherent to DSBs.
Nonhomologous Recombination

The comparison with BER is particularly helpful here. Pol β is similar to Pol4, Pol λ, and Pol µ in that it fills gaps by extending a primer strand, bringing them to ligation readiness (19). However, BER and NHEJ differ markedly in the template strand, which is continuous in BER but discontinuous in NHEJ (Fig. 4). Criticality, the point of template discontinuity in NHEJ depends on the polarity of the DSB (Fig. 4b). At 5' overhangs, the discontinuity is ahead of the primer-template pair, past the gap, and outside of the active site of the enzyme. At 3' overhangs, the discontinuity is behind the point of synthesis, within the active site in the position of primer-template pairing. Numerous findings indicate that Pol X NHEJ polymerases tolerate this latter substrate perturbation, whereas other polymerases cannot. Yeast Pol4 is required for synthesis only at 3', but not at 5', overhangs (20). Pol X NHEJ polymerases also show an unusual propensity toward frameshift mutations that suggest a greater flexibility in use of primer-template pairings (54). Extensive structural data also reveal that contacts with the distal 5' side of a gap are paradoxically more numerous than with the 3' (i.e., synthesis) side (55). Moreover, two protein loops unique to Pol X NHEJ polymerases make variable contacts to, or occupy the position of, the template strand (56, 57). These loops account in part for the stabilization of extrahelical bases during frameshift synthesis, and they might also contribute to stabilization of template discontinuities (Fig. 4b). One loop, in conjunction with a Pol µCFD histidine that bridges the dNTP and primer terminus, also facilitates template-independent synthesis and possibly template-dependent synthesis in the absence of initial primer-template base-pairing, which thus adds base-pairing potential to initially incompatible overhangs (Fig. 4b) (34, 51, 58). A picture emerges in which extended and atypical polymerase-DNA contacts overcome the instability of DSB joints caused by template strand discontinuity.

Tools and Techniques for Studying End Joining

The temporal sequence imposed in vivo on repair of single, often complex, chromatin DSB lesions is difficult if not impossible to recapitulate in vitro. It is thus advisable to take genetic data as a primary reference point for understanding NHR mechanisms. In vitro analyses nonetheless allow focused study of enzymes properties, and structural biology will continue to be irreplaceable in elucidating enzyme-substrate relationships.

Genetic analysis

A first approach to studying DSB repair is to treat cells with DSB-inducing agents such as ionizing radiation. A change in either survival or mutagenesis caused by a gene defect can be taken as evidence for its involvement in DSB repair, especially if supported by epistasis analysis to ascertain which repair pathway is affected. An alternative approach is to express in cells mega-endonucleases that create DSBs specifically at native or engineered chromosomal sites (3). In this way, repair can be tracked not only by survival but also by the status of genetic markers at the cut site. A limitation is that most joints restore the cut site and will simply be recleaved, so that repair outcomes are influenced by the method. Linking DSB formation to termination of endonuclease expression limits this issue (8), but the problem is never completely avoidable. A final approach introduces plasmids that were linearized in vitro into cells, where recircularization is required for plasmid maintenance (9, 20). Although limited by the use of naked DNA, the simplicity of this approach and the tremendous variety of available DSB ends make it common and powerful.

In vitro reconstitution

Two main approaches have been used for studying NHEJ in vitro. In the first approach, crude extracts are prepared from cells and are spiked with DNA fragments whose joining is followed by electrophoretic monitoring or PCR amplification (27, 53). Extracts have a full complement of cellular proteins in appropriate ratios and might therefore provide the best recapitulation of the in vivo situation while allowing manipulation of conditions, DSB ends, and so on. More reports reconstituted the NHEJ reaction from purified mammalian, yeast, or bacterial components (26, 33, 51). Some rejoining can be catalyzed by nothing more than a ligase, but focusing on incremental activity with combined components and complex DSBs gives confidence that results are meaningful to cellular NHEJ. Another concern is that NHEJ reconstitution has always used mass action chemistry with sometimes high concentrations of proteins and DNA, which is in marked contrast to the single pair of ends presented in vivo.

Structural analysis

Crystal structures of many NHEJ proteins have tremendously revealed the overall shape of the large and complex DNA-PKcs structure (23). For most other domains, structural inferences can be drawn by mapping NHEJ proteins onto crystallized enzymes of the same class. However, no published structures show two DSB ends engaged together, which I argue defines the unique functions of NHEJ proteins.1 Similarly, poor insight exists into the interconnected assembly of the full NHEJ repair complex. Furthermore, crystal structures are static, whereas NHEJ is dynamic with multiple different reactions.

Future Directions

Nonhomologous DSB repair has a complex relationship with genome preservation in that it can yield both silent repair and NHR. More understanding of this relationship will require improved genomic systems for monitoring the influence of nonhomologous repair pathways on genome integrity. It will also

1 A recent paper has now reported a structure of ends being bridged by the polymerase domain of the bacterial Lpg enzyme (Brisette NC, Pitcher RS, Juarez R, Pitcher AJ, Green AJ, Dafforn TR, Fox GC, Blanco L, Doherty AJ. Structure of a NHEJ polymerase-mediated DNA synaptic complex. Science 2007;318:856-8).
require delineation of the cell-cycle, checkpoint, and chromatin mechanisms that regulate S’ resection and the interplay of NHR and HR. Better molecular definition of the various nonhomologous repair pathways is needed, especially with regard to NHEJ, in which clear mechanisms are still pending. Pol X polymerases provide a conceptual framework for exploring how individual proteins engage the unstable substrates inherent to a DSB, which could be extended to the ligase and beyond. More interaction mapping and biochemical and structural analyses will be required to understand the assembly of a full NHEJ repair complex and the function of currently enigmatic contributors to it. Specific goals include single-molecule NHEJ reconstitution in vitro, structures with two-sided DSBs, and delineation of dynamic changes and regulatory switches that occur during reaction progression.

References

15. Lieder MR, Yu K, Raphaelian SC. Roles of nonhomologous DNA end joining, V(D)J recombination, and class switch recombination in chromosomal translocations. DNA Repair 2006;5:1234–1245.
33. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
34. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
35. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
36. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
37. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
38. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
40. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.
41. Afligon P, Smith P, Jackson SP. XLF interacts with the Xrc4-DNA ligase IV complex to promote DNA nonhomologous end joining. DNA Repair 2006;5:362–368.


61. Teo SH, Jackson SP. Ligase IV can ligate incompatible DNA ends and can ligate across gaps. EMBO J. 2007;26:1010–1023.


Further Reading


This special issue of DNA Repair contains a series of review articles on factors that affect chromosome maintenance. Several articles are referenced herein, but the entire volume is an excellent single source for beginning a more in-depth exploration of this field.

See Also

DNA Damage
DNA Damage Repair, Chemistry of Eukaryotic DNA Polymers, Chemistry of Homologous Recombination
Nucleic Acids, Enzymes that Modify DNA
Ca\textsuperscript{2+}-\textit{ATPase}/Ca\textsuperscript{2+} Pump, Chemistry of

Anne Nyholm Anthonisen and Jens Peter Andersen,
Institute of Physiology and Biophysics and PUMPKIN Center, University of Aarhus, Denmark

doi: 10.1002/9780470048672.wecb047

Sarco(endo)plasmic reticulum Ca\textsuperscript{2+}-\textit{ATPase} (SERCA) transports Ca\textsuperscript{2+} ions against a concentration gradient using energy derived by the hydrolysis of ATP. SERCA forms an aspartyl phosphorylated intermediate, which is acid stable, thus enabling studies of the reaction mechanism. Methods have been established for a detailed functional analysis of mutants, allowing definition of the roles of the individual amino acid residues in the partial reaction steps, and the mutational studies can now be correlated with the positions of the residues in crystal structures of SERCA in various conformational states. Here, we use the glutamine mutant of Glutamate 183 as an example of the methods and analyses. This mutant exhibits a block of the dephosphorylation of the E2P intermediate, but it is little affected in the E1 state, which is consistent with a key role for Glu183 in positioning the attacking water molecule during the hydrolysis of the acyl phosphate bond in the E2P phosphoenzyme.

Sarco(endo)plasmic reticulum Ca\textsuperscript{2+}-\textit{ATPase} (SERCA) is a prime example of a membrane-bound ion pump. This energy transducer or "nanomotor" transports Ca\textsuperscript{2+} against a steep concentration gradient across the sarco(endo)plasmic reticulum membrane, which uses the energy derived from ATP hydrolysis. Like the Na\textsuperscript{+}, K\textsuperscript{+}-ATPase and several other membrane-bound transport ATPases, SERCA belongs to the family of P-type ATPases that form an acylphosphate intermediate during the reaction cycle. This intermediate is acid stable, thus enabling kinetic studies of its formation and decay by acid-quenching techniques. Examination of the partial reaction steps using isotopically labeled ATP, phosphate, and Ca\textsuperscript{2+} have led to a detailed description of the reaction cycle (see References 1 and 2, and more references in Further Reading). SERCA functions as a single monomeric 110-kDa polypeptide chain; thus, the ability to hydrolyze ATP coupled with the protein conformational changes involved in translocation of bound Ca\textsuperscript{2+} is contained within a single molecule (3). The amino acid sequence was determined by cDNA cloning (4), and the development of a technique for efficient expression of the cDNA paved the way for studies of the structure-function relationship using site-directed mutagenesis (5). The first successful task accomplished by this methodology was the identification of almost all the residues binding the Ca\textsuperscript{2+} ions (6-8), which later was confirmed by the crystal structure (9). To date, several hundred mutant forms have been expressed and analyzed functionally, leading to identification of residues involved in essential steps of the Ca\textsuperscript{2+} transport mechanism, binding of ATP, phosphorylation and dephosphorylation, as well as the transport associated conformational changes (6). A database on all SERCA mutations with functional consequences is available by request to the corresponding author.

Because of the ease with which the SERCA1a isoform of the Ca\textsuperscript{2+}-ATPase can be prepared in almost pure form and high yield from rabbit skeletal muscle, this enzyme has been amenable to a large variety of biochemical and biophysical experiments. In 2000, the atomic structure of SERCA1a with bound Ca\textsuperscript{2+} determined by X-ray crystallography was published as the first high-resolution structure of a P-type ATPase (9). This achievement was followed by crystal structures of the enzyme in several functionally relevant conformations (see References 10-13). Many structural details correlate well with the functional analysis of mutants, thus providing a sound basis for the use of site-directed mutagenesis to obtain additional insight into the structure-function relationship. With detailed structural information now available, mutagenesis of the Ca\textsuperscript{2+}-ATPase entered a new age, in which the task is to explain the functional relevance of the structural features observed rather than to predict structure.
Ca2+/ATPase: Ca2+/Ca2+ Pump, Chemistry of

**Background**

**Reaction Cycle**

Figure 1 shows the reaction cycle of SERCA as currently understood. Two Ca2+ ions are translocated in each cycle of ATP hydrolysis in exchange for 1 to 3 protons (2 being illustrated). Four major conformational states can be defined, of which two are phosphorylated and the other two are not. The E1 forms display specificity for reaction with ATP/ADP ("kinase activity") and the binding of Ca2+ with high affinity, whereas the E2 forms react with water/oxygenophosphate (P) instead of nucleotide ("phosphatase activity") and cannot bind Ca2+ with high affinity. During ion translocation, the ions become occluded (i.e., cannot associate with the surrounding media on either side of the membrane) as indicated by the closed boxes (2, 3, 10–12). The E1 form exposes its two high-affinity Ca2+ binding sites to the cytoplasm, and these sites have to be occupied before phosphorylation from ATP can occur. As long as the enzyme has not yet become phosphorylated, the Ca2+ ions can dissociate back toward the cytoplasmic side, although they may be temporarily occluded (the gate closes but is not yet locked, flickering open from time to time). The phosphorylation locks the cytoplasmic gate, and the phosphorylated E1P form, now with occluded Ca2+ ions, subsequently undergoes a conformational transition to E2P—a state where the Ca2+ sites are exposed toward the lumen (the luminal gate opens) and exhibit low affinity. After the luminal release of Ca2+, protons bind at the transport sites from the luminal side, and E2P dephosphorylates by hydrolysis of the aspartylphosphate bond, which leads to occlusion of the protons in E2. As a result of the transition to E1, the translocated protons are released on the cytoplasmic side, which reestablishes the cytoplasmically facing high-affinity Ca2+ binding sites.

Acting as a catalyst at the phosphorylation site, Mg2+ ions are part of the reaction cycle as it coordinates ATP/Pi. Alkali metal ions like Na+ and K+ are also important as they accelerate the dephosphorylation of E2P by binding to a regulatory site in the cytoplasmic part of the protein (14). The whole transport cycle is reversible. Thus, the pump can be phosphorylated by Pi in the E2 form, which predominates at acidic pH in the absence of Ca2+ on the cytoplasmic side. If luminal Ca2+ is present, then it can bind to the luminal facing ion binding sites of the E2P form, and thereby allow backward formation of E1P from E2P. Transfer of the phosphoryl group from E1P to added ADP results in ATP synthesis accompanied by movement of Ca2+ from the lumen to the cytoplasm (1). The two phosphoenzyme conformations E1P and E2P are denoted ADP-sensitive and ADP-insensitive phosphoenzymes, because they are dephosphorylated by reaction with ADP (yielding ATP) and water (yielding Pi), respectively (1, 2).

The E2 form of the enzyme has been shown to be target of the SERCA specific inhibitors thapsigargin, CPA (cyclopiazonic acid), and BHQ (2,5-di-tert-butyl-1,4-dihydroxybenzene), as well as vanadate, which inhibits a large variety of enzymes that interacts with phosphate compounds.

---

**Correlation with Structure**

SERCA consists of a membrane-spanning domain of 10 mostly alpha-helical segments (M1 through M10, numbered from the N-terminal end) and a large cytoplasmic head-piece made up by three distinct domains, named N (nucleotide binding), P (phosphorylation), and A (actuator) (9). Figure 2 shows two examples of SERCA structures determined by X-ray crystallography. The transition state occurring immediately before formation of the E1P state was mimicked by adding AlF4− as phosphoryl analog together with ADP to Ca2+-bound enzyme (left) (10). The E2 state with noncovalently bound Pi, E2P, occurring between E2P and E2 in the reaction cycle, was mimicked using MgF42− as phosphate analog together with the inhibitor thapsigargin that binds in the transmembrane region (right) (11). Three key residues in the Ca2+-transport mechanism are indicated in the figure: Glu309, A3sp351, and Glu318. In the E1 form, two Ca2+ ions are coordinated by residues in the transmembrane segments M4, M5, M6, and M8 (6–10, 15, 16), of which M4 and M6 are unwind in the middle to make space for the Ca2+ ions and allow the side chains to interact with the ions. The glutamic acid residue in M4, Glu309, not only coordinates one of the two Ca2+ ions but also seems to function as a gate at the cytoplasmic entrance of the binding cavity (16). The M1 transmembrane segment is bended in the occluded form, and the part of M1 near the bend approaches Glu309, which thereby contributes to stabilization of the closed state of the cytoplasmic gate (17). A sp351 in the P-domain is the residue that undergos phosphorylation and dephosphorylation during the functional cycle, and no phosphorylation occurs in mutants with alteration to A3sp351 (5). In the E1 and E2P forms, the catalytic site is made up exclusively by residues in the P- and N-domains. During the E1P to E2P transition, the A-domain rotates more than 90° horizontally around an axis perpendicular to the membrane, such that Glu183 of the conserved TGES motif in the A-domain
Functional Analysis of Mutants

For functional studies of mutant enzyme forms, mutated cDNA must be expressed in a suitable system. The mammalian COS cell system was introduced originally for expression of Ca\(^{2+}\)-ATPase mutants by Maruyama and MacLennan (5). This system is advantageous because the microsomal fraction that contains the expressed mutant or wild-type Ca\(^{2+}\)-ATPase exhibits only a very low background of endogenous Ca\(^{2+}\)-ATPase and other enzymes that hydrolyze ATP or form an acid-stable phosphoenzyme. For phosphorylation studies with radioactive (gamma-32P)ATP or 32P, or studies of Ca\(^{2+}\) occlusion (7, 8), conventional transfection methods such as calcium phosphate precipitation and lipofection suffice to produce enough material, whereas, for more demanding measurements of equilibrium binding of Ca\(^{2+}\), adenovirus vectors can be used to increase the transfection efficiency in the COS cells (15). For biophysical studies, yeast expression, which yields more enzyme (but with higher background of other ATPases), is recommended (see Further Reading).

Two basically different types of phosphorylation experiments are employed.

1. Ligand dependency of steady-state or equilibrium phosphorylation—the enzyme is incubated with various concentrations of (gamma-32P)ATP or 32P, or the concentration of Ca\(^{2+}\) is varied in the presence of the phosphorylating substrate, to determine the apparent affinity for the ligands. Likewise, the concentration dependence of inhibition of phosphorylation by vanadate, thapsigargin, or a fluoride phosphor analog such as AlF\(_4^\text{-}\), can be determined by addition of the inhibitor at various concentrations followed by measurement of phosphorylation in the presence of (gamma-32P)ATP and Ca\(^{2+}\) (18, 19).

2. Transient state kinetics—the enzyme is incubated for sequential time intervals with (gamma-32P)ATP or 32P, to allow measurement of the rate of phosphorylation, or the dephosphorylation rate is determined by chasing the phosphoenzyme in conditions that prevent additional phosphorylation. It is possible to design conditions that lead to selective accumulation of each of the four main intermediates, E1, E1P, E2P, and E2, at steady state or equilibrium (18, 20). The transformation of one state to the next in the reaction cycle can then be followed on addition or removal of a ligand. These experiments are carried out either by hand mixing at 0°C or with the use of quench-flow instrumentation that allows measurements of rates under more physiological conditions (25°C, ms to seconds) (20).

Because of the acid stability of the aspartylphosphate bond of the phosphoenzyme, acid is used to quench (i.e., "freeze") the phosphorylated state for quantification of the protein-bound radioactive phosphate. After acid quenching, the protein is washed by centrifugation and isolated by sodium dodecyl sulfate polyacrylamide gel electrophoresis under acidic buffer conditions. Quantification of the phosphatase activity associated with the Ca\(^{2+}\)-ATPase is performed conveniently by quantitative autoradiography ("imaging") of the dried gels.

The Glu183Gln Mutant as Example

Here we illustrate some techniques applied in the functional analysis of mutants using as an example the mutant having Glu183 of the conserved TGES loop of the A-domain replaced by glutamine (Reference 10; see also Reference 19 where the alanine mutant was analyzed with similar results).

Ca\(^{2+}\) affinity

The Glu183Gln mutant is well expressed in the COS cell system, but unlike the wild type it does not possess any ATPase...
activity or ability to transport 
and accumulate calcium into the microsomal vesicles. In panel (a) of Fig. 3, it is demonstrated that Ca\textsuperscript{2+} does not readily bind to the enzyme and activates the phosphorylation process from ATP with an apparent affinity slightly higher than that of the wild type (K\textsubscript{a} for Ca\textsuperscript{2+} activation slightly reduced). Hence, there is no disruption of the function of the Ca\textsuperscript{2+} sites in the E\textsubscript{1} form in this mutant. The slight increase in apparent Ca\textsuperscript{2+} affinity at steady state may be explained by the reduced phosphoenzyme turnover rate (Fig. 3, panels d, e, and f), and because lower Ca\textsuperscript{2+} concentrations are required for the phosphoenzyme to accumulate when the turnover rate is reduced.

**Formation of the E1P phosphoenzyme from E1**

Panel (b) of Fig. 3 depicts the results of a transient kinetic experiment performed using quench-flow instrumentation to determine the time course of phosphorylation from (gamma-32P)ATP of the Ca\textsuperscript{2+}-bound E1 form (step 2 in Fig. 3). The rate constant determined for the mutant by fitting a monoexponential function is 22 s\textsuperscript{-1} versus 35 s\textsuperscript{-1} for the wild type (i.e., a rather moderate difference).

**The Ca\textsuperscript{2+} binding E2 to CaE1 transition**

Taking advantage of the fact that only the Ca\textsuperscript{2+} state can form a phosphoenzyme by reaction with ATP, the rate of the E2 to CaE1 transition (steps 6 and 1 in Fig. 1) can be determined as illustrated in panel (c) of Fig. 3. Ca\textsuperscript{2+}-deprived enzyme (excessing in a Ca\textsuperscript{2+} affinity equilibrium between E1 and E2), see Fig. 1) is incubated with Ca\textsuperscript{2+} for sequential time intervals (t), and the amount of phosphophosphorylated CaE1 accumulated during the Ca\textsuperscript{2+} incubation step is determined in each case by another 34 ms incubation with (gamma-32P)ATP, followed by acid quenching. Because the phosphorylation from ATP (step 2 in Fig. 1) requires the enzyme to be in the CaE1 form and is rapid, as documented in panel (b) of Fig. 3, the amount of phosphoenzyme measured at t + 34 ms follows the same function of t as the appearance of CaE1.

---

**Figure 3**

Functional analysis of mutant Gln185Gln (squares) and the wild-type Ca\textsuperscript{2+}-ATPase (circles). Panel (a): Ca\textsuperscript{2+} dependence of steady-state phosphorylation with ATP. Phosphorylation was performed for 10 s at 0°C by manual mixing in a buffer that contained 40 mM MOPS/Tris pH 7.0, 80 mM KCl, 5 mM MgCl\textsubscript{2}, 1 mM EGTA, 5 (gamma-32P)ATP, and various additions of CaCl\textsubscript{2} to obtain the free Ca\textsuperscript{2+} concentrations indicated on the abscissa. The lines show the best fits of the Hill equation to the data. Panel (b): Quench-flow measurement at 25°C of the time course of phosphorylation with ATP of Ca\textsuperscript{2+}-bound enzyme. The enzyme, preincubated in 40 mM MOPS/Tris pH 7.0, 80 mM KCl, 5 mM MgCl\textsubscript{2}, and 100 nM CaCl\textsubscript{2}, was mixed with an equal volume of the same medium that contained in addition 10 nM (gamma-32P)ATP (final concentration 50 nM), and acid quenching was performed at the indicated time intervals after mixing. Panel (c): Quench-flow measurement at 25°C of the time course of phosphorylation with ATP of enzyme preincubated without Ca\textsuperscript{2+}. The enzyme, which was kept in 40 mM MOPS/Tris pH 7.0, 80 mM KCl, and 2.2 mM EGTA, was mixed with an equal volume of medium that contained 40 mM MOPS/Tris pH 7.0, 80 mM KCl, 10 mM MgCl\textsubscript{2}, 1 mM EGTA, and 1.1 mM CaCl\textsubscript{2} for 34 ms, prior to acid quenching. Panel (d): Turn-over of phosphoenzyme formed from ATP and distribution of E1P and E2P. The enzyme was phosphorylated in the presence of 5 (gamma-32P)ATP for 10 s at 0°C in 40 mM MOPS/Tris pH 7.0, 80 mM KCl, 5 mM MgCl\textsubscript{2}, 2 mM calcium chelator EGTA, 5 mM MgCl\textsubscript{2}, and 0.55 mM CaCl\textsubscript{2} (yielding a free Ca\textsuperscript{2+} concentration during phosphorylation of 10 nM on both sides of the membrane). Dephosphorylation was followed by acid quenching at the indicated time intervals after chase with 6.7 mM EGTA (open symbols, total phosphoenzyme). To determine the fraction of the phosphoenzyme which is ADP-insensitive E2P at the indicated time intervals, ADP-sensitive E1P was removed by inclusion of 0.9 mM ADP for 2 s before acid quenching (closed symbols). Monoexponential decay functions were fitted to the data. Panel (e): Decay of E1P as a result of the E1P to E2P transition calculated as the difference between the two curves corresponding to the open and closed symbols in panel (d). Panel (f): Dephosphorylation of E2P. The E2P phosphoenzyme was formed by incubation with 0.5 mM ATP for 10 minutes at 25°C in 100 mM MES/Tris pH 6.0, 10 mM MgCl\textsubscript{2}, 2 mM EGTA, and 30% v/v DMSO. The reaction mixture was cooled for 2 minutes on ice, prior to a 19-fold dilution into ice-cold dephosphorylation medium that contained 40 mM MOPS/Tris pH 7.0, 2 mM EGTA, 1 mM MgCl\textsubscript{2}, 80 mM KCl, and 0.5 mM nonradioactive Pi, followed by acid quenching at the indicated time intervals.
hydrolysis of E2P to E2 + Pi. The rate of the conversion of E1P during the first 34 ms reaction of Ca^{2+} (type) corresponds to the amount of phosphoenzyme formed during the first 34 ms reaction of Ca^{2+}-deprived enzyme with (gamma-32P)ATP together with Ca^{2+}. This amount depends on the rate of the E2 to Ca^{2+}E1 transition as well as on the relative amount of enzyme that initially is present as E1 and, therefore, reaches the phosphorlylatable Ca^{2+}E1 state faster than E2.

The E1P to E2P transition

Panels (d) and (e) of Fig. 3 show how the rate of the Ca^{2+}-translocating E1P to E2P transition of the phospho-enzyme is determined. The enzyme is phosphorylated with (gamma-32P)ATP under conditions where E1P accumulates as the major steady-state intermediate in the wild-type enzyme [0°C, neutral pH; K+ present to activate dephosphorylation of E2P (14)]. Phosphoenzyme decay in the forward direction of the pump cycle is examined by addition of excess EGTA, which removes Ca^{2+} and thereby terminates phosphorylation from ATP, followed by acid quenching at sequential time intervals (open symbols in panel (d) of Fig. 3). It is observed that the dephosphorylation is much slower in the mutant than in the wild type. The phosphoenzyme determined here is the sum of the ADP-sensitive E1P and the ADP-insensitive E2P (see Fig. 1). To quantify the amount of E2P, ADP is added to remove E1P 2 s before quenching (closed symbols). It is revealed that approximately 60% of the mutant phosphoenzyme is ADP-insensitive E2P at the beginning of the chase, in contrast to the wild type that is 100% ADP-sensitive E1P. The dephosphorylation occurs through the conversion of E1P to E2P followed by hydrolysis of the latter to E2 + Pi (Fig. 3), and the retention of the ADP sensitivity during phosphoenzyme decay in the wild type demonstrates that the E1P to E2P conversion is rate limiting for the dephosphorylation. In the mutant, dephosphorylation is very slow, and the ADP-insensitive fraction of the phosphoenzyme increases with time (panel (d), closed triangles, in Fig. 3), which indicates that the rate-limiting step for the dephosphorylation is not the conversion of E1P to E2P, but the hydrolysis of E2P to E2 + Pi. The rate of the conversion of E1P to E2P is calculated by subtracting the data that corresponds to the ADP-insensitive E2P (closed symbols in panel (d)) from the data that represents the total amount of phosphoenzyme [open symbols in panel (d)]. The results of the calculation shown in panel (e) indicate that the E1P to E2P transition is about threefold slower in the mutant, as compared with the wild type, even though E2P is observed to accumulate in the mutant.

The dephosphorylation of E2P

The reason for the accumulation of E2P in the mutant is clear from the results depicted in panel (f) of Fig. 3, which reveal a block of the hydrolysis of E2P to E2 + Pi. To measure this reaction separately, the E2P phospho-enzyme is formed backward by reacting E2 (accumulated in the absence of Ca^{2+}) with 32P, in the presence of 30% (v/v) dimethyl sulfoxide, which increases the affinity for Pi. The 32P-labeled E2P phospho-enzyme is chased by dilution into a dephosphorylation medium that contains an excess of nonradioactive Pi, and the phospho-enzyme decay is followed by acid quenching at sequential time intervals. The difference in the rates of E2P dephosphorylation of wild type and mutant is conspicuous, amounting to ~100 fold.

Conclusion for Glu183Gln based on the mutational analysis in conjunction with the crystal structures

The results in Fig. 3 demonstrate that the Glu183Gln mutant behaves almost normally in the E1 state, whereas the dephosphorylation of E2P is blocked in the mutant. This finding is consistent with the hypothesis that the catalytic site fluctuates between two major configurations, E1/E1P with “kinase activity” and E2/E2P with “phosphatase activity,” because of the movement of Glu183 into the catalytic site in E2/E2P. The position of Glu183 in the crystal structure of E2 with AlF4− bound as phosphoryl transition state analog (12) suggests that the role of Glu183 in the dephosphorylation of E2P is to position the attacking water molecule, see Fig. 4. The explanation of the distinct reactivities of E1P and E2P toward ADP and water is simply that in E2P the TGES loop with Glu183 occupies the position that ADP has in E1P (10, 11).

Furthermore, the rates of the conformational changes E2 to E1 and E1P to E2P are reduced about threefold in the mutant, which indicates that the rate-limiting step for the dephosphorylation is not the conversion of E1P to E2P, but the hydrolysis of E2P to E2 + Pi. The rate of the conversion of E1P to E2P is calculated by subtracting the data that corresponds to the ADP-insensitive E2P (closed symbols in panel (d)) from the data that represents the total amount of phospho-enzyme [open symbols in panel (d)]. The results of the calculation shown in panel (e) indicate that the E1P to E2P transition is about threefold slower in the mutant, as compared with the wild type, even though E2P is observed to accumulate in the mutant.

The dephosphorylation occurs through the conversion of E1P to E2P followed by hydrolysis of the latter to E2 + Pi (Fig. 3), and the retention of the ADP sensitivity during phosphoenzyme decay in the wild type demonstrates that the E1P to E2P conversion is rate limiting for the dephosphorylation. In the mutant, dephosphorylation is very slow, and the ADP-insensitive fraction of the phosphoenzyme increases with time (panel (d), closed triangles, in Fig. 3), which indicates that the rate-limiting step for the dephosphorylation is not the conversion of E1P to E2P, but the hydrolysis of E2P to E2 + Pi. The rate of the conversion of E1P to E2P is calculated by subtracting the data that corresponds to the ADP-insensitive E2P (closed symbols in panel (d)) from the data that represents the total amount of phospho-enzyme [open symbols in panel (d)]. The results of the calculation shown in panel (e) indicate that the E1P to E2P transition is about threefold slower in the mutant, as compared with the wild type, even though E2P is observed to accumulate in the mutant.
which suggests that Glu183 is involved in interactions that facilitate these conformational changes. In some E2 crystal structures (e.g., Protein Data Bank ID 1XP5 and 1WPG), Glu183 is within hydrogen bond distance of Thr352 in the P-domain, and such interaction might facilitate both the insertion of the TGES loop in the catalytic site during the E1P to E2P transition and the disengagement during the E2 to E1 transition (19).

Final Remarks on the E1-E2 Model for The Reaction Cycle

The "E1-E2" two-state nomenclature ("E-E" in the earlier Ca2+-ATPase literature) was introduced solely on the basis of functional measurements (1), and its relevance has been questioned, although a similar "Post-Albers scheme" has also been useful in the functional analyses of the Na+/K+-ATPase (see Further Reading). With the determination of atomic structures of the various conformational states and the supporting functional evidence from mutagenesis, the two-state nomenclature has turned out to have a sound structural basis, even though at least 4 different conformations exist. Hence, in all forms referred to as E1 (E1, E1P, and transition states between them) the TGES loop with Glu183 is separated from the catalytic site (Fig. 2), and the Ca2+-bound residues are in position to bind the ions with high affinity. The same set of residues is involved in Ca2+-binding in all E1 states, irrespective of whether the sites are occupied or not, as the occlusion depends on the position of M1. On the other hand, all E2 forms (E2, E2P, and transition states between them) have the A-domain rotated with Glu183 approaching the catalytic site, and these structures show the Ca2+ binding sites to be disrupted (either facing the lumen, as in E2P, or rearranged to occlude protons as in E2 and E2P).

References


Further Reading


Vilsen B, Andersen JP. Interdependence of Ca2+-oocclusion sites in the unphosphorylated sarcoplasmic reticulum Ca2+-ATPase complex with Ca2+-ATPase. J. Biol. Chem. 1992;267:3359-3350.

Yu X, Han L, Inesi G. All change of acidic residues contributes to cation countertransport in the Ca2+-ATPase of sarcoplasmic reticulum. Role of H+ in Ca2+-ATPase countertransport. J. Biol. Chem. 1994;269:16606-16611.

Additional crystal structures of Ca2+-ATPase and related P-type ATPases: Jensen AM, Sørensen TL, Olsen C, Møller JV, Nissen P. Modulatory and catalytic modes of ATP binding by the calcium pump. EMBO J. 2006;25:2305-2314.


More examples of mutagenesis of the Ca2+-ATPase


Andersen JP, Vilsen B. Amino acids Asn796 and Thr799 of the Ca2+-ATPase of sarcoplasmic reticulum bind Ca2+ at different sites. J. Biol. Chem. 1994;269:13931-13936.


Clausen JD, McIntosh DB, Woolley DG, Andersen AN, Vilsen B, Andersen JP. Asparagine 706 and glutamate 183 at the catalytic site of sarcoplasmic reticulum Ca2+-ATPase play critical but distinct roles in E2 states. J. Biol. Chem. 2006;281:9471-9481.
Chemistry of Bacteriorhodopsin

Janos K. Lanyi, University of California, Los Angeles, California
doi: 10.1002/9780470048672.wecb035

Thirty years of research with bacteriorhodopsin has provided answers to many questions about how protons are transported by transmembrane pumps. In this small seven-transmembrane protein, absorption of light by the retinal chromophore initiates a reaction cycle in which the initial state recovers through multiple conformational changes of the retinal and the protein, and a proton is translocated stepwise from one side of the membrane to the other. Spectroscopy, extensive use of site-specific mutations, and crystallography have defined the photocycle reactions in atomic detail and provide a step-by-step description of the proton transfers, the transient local and global perturbations in the protein and how they arise, and the energy flow through the system, which add up to the mechanism of the pump.

Membrane proteins perform some of the most topologically complex and important functions in biology, such as translating small and large molecules into and out of cells and organelles, coupling energetic functions through the generation and the discharge of transmembrane ion gradients, initiating signaling cascades in cells, and inserting and folding of polypeptides into the lipid bilayer. Bacteriorhodopsin belongs to the heptahelical membrane-protein family that also includes a wide variety of receptors for biological signals. As a light-driven pump for protons in halophilic archaea (1, 2), it generates the inwardly oriented transmembrane electrochemical proton gradient used for synthesis of adenosine triphosphate (ATP) and the uptake of nutrients by the cells. Illumination causes isomerization of the purple retinal chromophore from all-trans to 13-cis,15-anti, and the ensuing sequence of thermal reactions regenerates the initial state in ca. 10 ms, through protein conformational changes and internal proton transfers that translocate a proton across the membrane. Interest in bacteriorhodopsin has been renewed in the last few years because many bacteriorhodopsin-like photoreactive proteins are now found also in eubacteria and eukaryotes, and thus, it is becoming clear that the use of retinal for light-sensing and energetics is more widespread in biology than ever imagined. In addition to the already well-known chloride pump of the archaea, halorhodopsin, these newly discovered proteins include the proton pumps proteorhodopsin and xanthorhodopsin, and various sensory rhodopsins.

Background

As an active transporter, bacteriorhodopsin functions as a biological energy transducer. Proteins of this kind move ions against their electrochemical potential across the membrane barrier, using energy generated by a chemical reaction or from light. Such ion pumps are necessary for muscle contraction (Ca\(^{2+}\) transport), nerve conduction (Na\(^{+}\) and K\(^{+}\) transport), as well as ATP synthesis in mitochondria and photosynthetic systems (H\(^{+}\) transport). Because of their large size and in most cases complex multi-subunit structure, progress in understanding the mechanism of the transport and how it is coupled to the driving reaction in the pumps has been slow.

The conceptual and methodological advantages of bacteriorhodopsin have made it possible to gain insights into its structure, reaction cycle, and function as an ion pump in far more detail than for other membrane proteins. The outstanding questions concern the interhelical interactions that fold the protein and confer stability to the seven-helical bundle immersed in the lipid bilayer, the identity and location of functional groups in the protein interior, the nature of the perturbation of the retinal chromophore upon photoisomerization, the propagation of conformational changes from retinal to near and far regions of the protein, the kinetics and energetics of the reaction cycle in which a proton is transferred from group to group inside the protein, the proton release and uptake at the membrane surfaces, and the nature of the protonation switch that confers directionality on the pump. Answers to such questions are sought for many ion pumps, but in bacteriorhodopsin, the large amount of information now available gives a vivid picture of the molecular events that accomplish the transport.
Structure and Mechanism

The protein (26 kDa) is a bundle of seven helices, A through G, with three helices normal to the membrane plane and the rest inclined at small angles to the normal. The retinal is covalently bound through a protonated Schiff base to Lys-216 at the middle of helix G, and it lies nearly parallel to the membrane in the space surrounded by the seven helices. The protein forms in vivo a homotrimer, and the trimers are assembled into an extended hexagonal lattice (P3 symmetry), the "purple membrane." Protein-protein contact in these patches is through a continuous boundary layer of lipids no more than one lipid wide, at the monomer periphery. The interhelical loops as well as the N- and C-termini are short, with the exception of the connection of helices B and C through a structured as the N- and C-termini are short, with the exception of the wide, at the monomer periphery. The interhelical loops as well as the N- and C-termini are short, with the exception of the connected by a continuous boundary layer of lipids no more than one lipid wide, at the monomer periphery.

At the active site, the positively charged protonated retinal Schiff base donates a hydrogen bond to a water molecule (wat402) that donates in turn to the extracellular side, and the E–F interhelical loop with a twist, on the cytoplasmic side.

In the L-state, the retinal is still somewhat twisted but the Schiff base regains a hydrogen bond. There is disagreement over the partner in this hydrogen bond. In the highest resolution X-ray structure (9), it is wat402 (still to the extracellular direction), but in others, it is wat402 moved to the cytoplasmic side (10). In the latter case, rotation of the C15=N–H bond disconnects the Schiff base N–H bond from the extracellular side, and the E–F interhelical loop with a twist, on the cytoplasmic side.

Chemistry of Bacteriorhodopsin

Absorption of a photon by the purple all-trans retinal chromophore (with a single broad absorption band with a maximum at 568 nm) initiates the reaction sequence BR \( \rightarrow \) K \( \leftrightarrow \) L \( \leftrightarrow \) M \( \rightarrow \) N \( \rightarrow \) O \( \rightarrow \) BR (7, 8), where each state and substate is well defined by spectroscopic and crystallographic means. Although a kinetic scheme that rigorously fits all data into a linear sequence has not yet been produced, the proton transport mechanism can be understood by the molecular properties of the intermediate states and by their interconversions.

The first stable state, the K-intermediate, contains a highly twisted 13cis,15-trans retinal, with prominent hydrogen-out-of-plane (HOOP) bands, particularly at 320 nm. A nomalous Schiff base C=\( \leftrightarrow \)N stretch frequency, lesser \( \rightarrow \)N chemical shift, and the X-ray diffraction structures of K indicate that rotation of the C15=N bond disconnects the Schiff base N–H bond from wat402. At this stage, the overall contour of the all-trans retinal is maintained despite the isomerization. In the states that follow K, the retinal gradually relaxes, and deprotonation of the Schiff base as well as an increasingly accommodating binding site allow it to assume the bent configuration expected for a 13-cis,15-trans retinal.

In the L-state, the retinal is still somewhat twisted but the Schiff base regains a hydrogen bond. There is disagreement over the partner in this hydrogen bond. In the highest resolution X-ray structure (9), it is wat402 (still to the extracellular direction), but in others, it is wat402 moved to the cytoplasmic side (10). In the latter case, rotation of the C15=N–H bond disconnects the Schiff base N–H bond from the extracellular side, and the E–F interhelical loop with a twist, on the cytoplasmic side.

As the retinal relaxes, C13 and the 13-methyl group moves in the cytoplasmic direction toward Trp-182. The resulting
displacements of side chains between helices F and G, and the breaking of the connection between helices F and G through the hydrogen bonds of wat501 to Trp-182 and Ala-215 (Fig. 2), separate helices F and G. The link of Lys216 to wat502 and thereby to the peptide segment of Thr-46 propagates perturbation from the retinal to the Asp-96/Thr-46 pair. Entry of water and growth of water clusters creates a hydrogen-bonded chain of four water molecules between Asp-96 and deprotonated Schiff base (6). The increase of hydration at Asp-96 lowers its initially high pK_a and the aqueous chain conducts the proton to the Schiff base. The M-to-N reaction corresponds to this proton transfer, followed by reprotonation of Asp-96 from the cytoplasmic surface that produces N' from N. Crystallographic evidence and much non-crystallographic data indicate that these photocycle steps are accompanied by the outward and inward tilts of the cytoplasmic ends of helices F and G, respectively. The tilt of helix G arises from rearrangement of hydrogen bonds that eliminate, partly, the n-bulge in this helix at Lys-216. The tilt of helix F is a consequence of the separation of helices F and G and the repacking of side chains between them. There is reason to believe that it is the tilts that loosen the structure and allow the entry of water.

Reprotonation of Asp-96 allows the thermal re-isomerization of the retinal to all-trans, through another long-range coupling mechanism (12), which probably involves the hydrogen-bonded chain of water that bridges them. This produces the red-shifted O-state, where the retinal is, once again, twisted. Transition to the initial relaxed all-trans isomeric state occurs as the protonated Asp-96 loses its proton to the vacant proton release site at Glu-194/Glu-204, thereby recovering the initial electrostatic environment of the extracellular region and completing the cycle.

Key Experimental Methodology

The current understanding of how bacteriorhodopsin transports protons is the result of the use of an unprecedented variety of biophysical methods in many laboratories. Indeed, in many cases, the methods were developed with bacteriorhodopsin as the test subject. There is room to discuss only the most important of these methods.

Transient and low temperature spectroscopy in the ultraviolet/visible and infrared [Fourier transform infrared (FTIR) and resonance Raman] have yielded spectra of the intermediate states and provided many insights, even if not a fully satisfactory solution, to the kinetics of their interconversions. Transient spectroscopy in the visible (7) makes use of following absorption changes at selected single wavelengths, over as many as six orders of magnitude of time after flash photolysis, as well as an optical multichannel analyzer to obtain complete spectra at selected times after the flash. Alternatively, the intermediates can be trapped at cryogenic temperatures, after producing photostationary states (13). Both methods produce characteristic difference spectra and are well suited for identifying which of the various states is produced, as their single absorption maxima for the most part are well defined. Quantification of these states and calculation of a kinetic model from the observed succession of time-resolved spectra is more difficult, because the spectra are broad and several overlap one another. Furthermore, the reaction scheme is more complex than what is expected from a linear sequence in which each state arises by unidirectional transition from the preceding one. Multiple rise and decay components have been interpreted as evidence for back-reactions that lead to transient equilibration of states, and as multiple parallel photocycles. If present, the latter could arise from conformational heterogeneity, interactions within the bacteriorhodopsin trimers, and heterogeneous protonation states near the pK_a for the proton release group. Infrared and Raman spectroscopy is performed along the same lines, but the vibrational spectra (14–16) provide essential molecular level information about changes in the isomeric state of the retinal, the protonation states of dissociable residues, the strength of hydrogen bonds, and backbone conformation. An additional information on these spectra comes from the use of D_2O, H_2^18O, as well as 13C-, 18O-, and 15N-labeled amino acids, and polarized infrared beam for measuring linear dichroism, as well as single-site mutations to change the photolysis and to assign the positive and negative difference bands to specific amino acid residues.

Solid-state nuclear magnetic resonance (NMR) spectroscopy of bacteriorhodopsin is with the membranes either uniformly oriented relative to the magnetic field (17) or nonoriented but with "magic angle" spinning (18). In either case, the samples are enriched with 1H_2O or 12C isotopes in the examined parts of the molecule. The chemical shifts identify the protonation states of acidic groups, their hydrogen-bonding status, and the electrostatic environment. Dipolar interactions provide accurate...
interatomic distance and dihedral angle. Importantly, like FTIR, this method does not average structural information in the case of conformational heterogeneity. Difficulties include a requirement for large amounts of specifically labeled bacteriorhodopsin and problems of thoroughly illuminating a nearly opaque sample to trap the photocycle intermediates.

The purple membrane is isolated as an extended two-dimensional array of bacteriorhodopsin and lipids and, after some manipulation to increase the size of the patches, is an ideal material for cryo-electron microscopy. The maps produced to 3.5-7 Å resolution (19) identifies the positions of the seven-transmembrane helices, the retinal, and some bulky side chains. The latter helped to establish the rotational position of most helices in the seven-helical bundle, and thus, it yielded the first useful structural model. In the last decade, it has been possible to grow well-diffracting three-dimensional crystals. Bacteriorhodopsin crystals are usually thin (10-15 µm) hexagonal plates, 100-150 µm in diameter. X-ray diffraction of crystals produced reflections complete to 1.4-1.5 Å resolution, and the resulting model (3) defines the locations of water molecules bound in the protein interior and the hydrogen bonds relevant to the function and the stability of the structure. The highest crystallographic order is in crystals grown with the cubic phase method (20). The protein forms trimers in the same extended P3 sheet as in the native purple membrane, but the stacking of these sheets to produce the P63 symmetry is not with uniform orientation and the crystals usually show merohedral twinning. This problem is absent in crystals grown by other methods (10), but they are less ordered and the resolution is poorer. Many mutations of functional interest apparently disrupt the structure in subtle ways and make the protein unstable in detergent or do not allow the growth of crystals.

The cubic phase for the crystallization (20) is formed by a roughly 1:1 mixture of mono-olein and bacteriorhodopsin solubilized in octylglucoside. In a process not well understood, salts such as KCl or Na phosphate cause the conversion of the cubic phase with the incorporated bacteriorhodopsin into a lamellar phase and condensation of the protein into crystals. A related method, based on bicelles, adds phospholipids to the mixture. Conventional vapor phase crystallization in the presence of detergent yields bacteriorhodopsin crystals with a somewhat different packing arrangement. The crystals in general are suitable for the trapping of most (K-, L-, various M-, and N-) but not all (N and O), intermediate states (9). The M-state(s) can be accumulated with greatest ease because its absorption band is far removed from the initial state and its decay can be slowed by suitable mutations. Depending on their spectral overlap with the non-illuminated state, the quantum yields for the forward and back photo reactions, the wavelength of illumination, temperature, and mutations (if any), the other intermediates accumulate up to 60% in the photostationary states. Difficulties are measured at 100 K, where the intermediates are thermally stable and damage from the X rays is minimized.

Some movements at specific locations in the protein, for example, at the cytoplasmic surface in the M- and N-states, can be determined with site-specific spin-labels. This method is of particular value for intermediate states that do not accumulate in illuminated crystals. The paramagnetic probes are covalently linked to engineered cysteines, and the erythrocyte sedimentation rate (ESR) spectra of single or pairs of labels, determined statically in photostationary states or dynamically during the photocycle initiated by flash photoexcitation, reveal conformational shifts as changes in environmental restrictions on motion or changes in inter-spin distance with two labeled sites (21, 22). As in other proteins, a complementary method is to follow the reactivity of cysteine residues, placed one by one, along appropriate segments of the protein, such as interhelical loops.

Energetics of the Pump

The quantum yield of the photoisomerization from all-trans to 13-cis,15-trans is 0.60. From calorimetric measurements, the enthalpy gain in the first ground-state intermediate, K, is ca. 45 kJ/mol (23). Part of this will be used for proton transport, the rest is dissipated in the photocycle. One proton is translocated in each cycle across the membrane dielectric, against a proton gradient as high as 200 mV. The pump efficiency (the free energy of the transported proton vs. the energy gain in the K-state) is ca. 50%, and the overall efficiency of use of the energy input (the free energy of the transported proton vs. the energy of the absorbed photon) is ca. 10%.

In the K-state, the energy gain is conserved in bond torsions of the retinal from steric constraints, and electrostatic potential from loss of the Schiff base—counterion interaction and the resulting rearrangement of charges in the extracellular network. As the retinal Schiff base protonates Asp-85 in the L-to-M reaction so as to lose their electrostatic interaction, the binding site relaxes, stepwise, to accommodate the changed shape of the retinal, and free energy is passed from the chromophore to the protein. Most photocycle steps occur without a net free energy change, i.e., near equilibrium (8). The physiological pH (7.5) is well above the pKsa of the extracellular proton release complex (ca. 5.0), and the proton release constitutes one of the two steps where energy is dissipated, here as dilution entropy. In contrast, the pH where proton is taken up at the cytoplasmic surface is about the same as the pKsa of Asp-96 (ca. 7). The last step in the cycle is where the greatest part of the free energy gain is lost, resulting in a strongly unidirectional reaction that fully recovers the initial state from O. Loss of free energy in this step occurs most likely from proton transfer from Asp-85 to the proton release complex after these regain their low and high proton affinities, respectively (where the pKsa are 2.5 and ca. 9.5).

References


Further Reading


See Also

Membrane Proteins, Chemistry of
Photoreceptors, Chemistry of
Proton Translocation, Bioenergetics of
Crystalization of Proteins: Overview of Spectroscopic Techniques: Overview of

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc. 5 - 1190 -
Drug Transport in Living Systems

Yael Elbaz and Shimon Schuldiner, Alexander Silberman Institute of Life Sciences, Department of Biologic Chemistry, Hebrew University of Jerusalem, Jerusalem, Israel

doi: 10.1002/9780470048672.wecb143

Living organisms are constantly assailed by a host of harmful chemicals from the environment. Because of the diversity of these “xenobiotics”, cellular survival mechanisms must deal with an immense variety of molecules. Polyspecific drug transporters supply one such strategy. They recognize a wide range of dissimilar substrates that may differ in structure, size, or electrical charge, and they actively remove them from the cytoplasm. As such, these transporters provide an essential survival strategy for the organism. However, given that the substrates of these multispecific transporters include many antibiotics as well as antifungal and anticancer drugs, they are associated with the phenomenon of multidrug resistance (MDR) that poses serious problems in the treatment of cancers and infectious diseases, and some of them are coined multidrug transporters (MDTs). As expected from their central role in survival, these transporters are ubiquitous, and in many genomes, several genes coding for putative MDTs have been identified. MDTs are found in evolutionary unrelated membrane transport protein families, which suggests that they might have developed independently several times during the course of evolution. In this review, we discuss some basic concepts regarding drug transport from bacteria to humans.

Drug elimination in living systems is a multistep process that may involve metabolism, binding to proteins in the circulatory system and binding to specific receptors and excretion processes. A common way by which living organisms, from bacteria to humans, protect themselves against the harmful effect of a toxic compound is the removal of these molecules from the cell, which thus reduces the amount of drug that reaches its target site of action. Membrane transport proteins recognize these drug molecules and in an energy-dependent manner transport them across the plasma membrane and away from the target, either out of the cell or into subcellular organelles. The scope of this review encompasses numerous biologic networks and a variety of transport systems operating in concert. We will only discuss some major systems involved in transport of so-called drugs across biologic membranes in organisms from bacteria to humans. We chose to focus on polyspecific drug transporters, and by examination of some of the best characterized, we extract some general concepts.

Multidrug Transporters (MDTs)

The discovery of P-glycoprotein (P-gp/MDR1) in the mid 1970s introduced the concept of a single protein that can confer resistance to a relatively large number of structurally diverse drugs (1). In the following decades it has become clear that all living cells, be it a bacterial cell avoiding the deleterious effect of an antibiotic, a kidney cell eliminating an environmental toxin such as nicotine from the body, or a cancerous cell evading chemotherapeutic agents, express multidrug efflux transporters. Multidrug transporters (MDTs) recognize a wide range of dissimilar substrates that may differ in structure, size, or electrical charge, and they actively remove them from the cytoplasm. As such, these transporters provide an essential survival strategy for the organism. However, given that the substrates of these multispecific transporters include many antibiotic, antifungal, and anticancer drugs, they are associated with the phenomenon of multidrug resistance (MDR) that poses serious problems in the treatment of cancers and infectious diseases (2-4). The phenomenon of multidrug transport is an intriguing one because it seems to challenge the basic model of an enzyme
binding specifically to a single substrate in an optimized set of interactions as a prerequisite to efficient catalysis. MDTs, on the other hand, are polyspecific proteins that recognize a remarkably broad array of substrates. The occurrence of MDTs in evolutionary unrelated membrane transport proteins families indicates that they have originated independently several times during the course of evolution (5, 6). MDTs have been identified in several families, based on their amino-acid sequence similarities ([References (7) and (8); Table 1](9–16)].

### ATP binding cassette (ABC) superfamily

This superfamily contains primary active transporters that couple ATP binding and hydrolysis to substrate translocation across the lipid bilayer. This superfamily of transporters will be reviewed in a different chapter in this encyclopedia and will not be discussed here (see the article ATP binding cassette (ABC) Transporters in this encyclopedia). One of the many families included in the MFS is the SLC (solute carrier) superfamily. The SLC superfamily represents approximately 300 genes in the human genome that encode for either facilitated transporters or secondary active symporters or antiporters. Members of the SLC superfamily transport various ionic and nonionic endogenous compounds and xenobiotics. The SLC22 family includes anion and cation transporters (Organic Anion Transporters, OCTs; Organic Cation Transporters, OCTs). The SLC22 family shares a common theme with the MDTs discussed above. These polyspecific transporters handle an impressive broad substrate range. Many substrates are noxious to the organism, and these transporters actively remove them. The reasons they are not considered MDTs are therefore probably just historic.

### Resistance–nodulation–cell division (RND) superfamily

RND is a large ubiquitous superfamily of transporters with representations in all domains of life. Composed typically of about 1000 amino-acid residues, they are arranged as 12 transmembrane helices proteins with two large hydrophilic extracytoplasmic loops between helices 1 and 2 and helices 7 and 8. It has been postulated that these proteins developed from an internal gene duplication event. The members of the RND family are also secondary active transporters that catalyze the proton-motive-force driven transport of a range of substrates, including hydrophobic drugs, bile salts, fatty acids, heavy metals, and more ([12].

### Small multidrug resistance (SMR) family

One of the many families included in the MFS is the SLC (solute carrier) superfamily. The SLC superfamily represents approximately 300 genes in the human genome that encode for either facilitated transporters or secondary active symporters or antiporters. Members of the SLC superfamily transport various ionic and nonionic endogenous compounds and xenobiotics. The SLC22 family includes anion and cation transporters (Organic Anion Transporters, OCTs; Organic Cation Transporters, OCTs). The organic anion and cation transporters of the SLC22 gene family share a common theme with the MDTs discussed above. These polyspecific transporters handle an impressive broad substrate range. Many substrates are noxious to the organism, and these transporters actively remove them. The reasons they are not considered MDTs are therefore probably just historic.

### Multidrug and toxic compound extrusion (MATE) family

This newly identified family of MDTs called MATE includes representatives from all domains of life ([17]). These transporters contain 12 putative transmembrane helices. This family of MDTs was first identified with the characterization of NorM, which is a Na+/drug antporter from Vibrio parahaemolyticus ([18, 19]).

### Major facilitator superfamily (MFS)

The MFS is one of the largest and most widespread families of transporters and is composed of numerous subfamilies of diverse functions. This superfamily consists of single-polypeptide transporters from bacteria to higher eukaryotes involved in the symport, uniport, or antiport of a variety of small solutes. Among the various transporters of MFS transporters are sugars, drugs, neurotransmitters, Krebs cycle metabolites, amino acids, and many more ([20, 22]). Most MFS transporters are about 400–600 amino acids in length and possess either 12 or 14 transmembrane helices. Most proteins in this superfamily are secondary-active transporters that couple ion gradients to the translocation of a substrate. Evidence demonstrating that proteins within several MFS families consist of duplicated six transmembrane units supports the possibility that MFS permeases developed by an internal gene duplication event ([20].

### Table 1: MDTs families, examples, and selected references

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Examples</th>
<th>Recent reviews and additional reading</th>
</tr>
</thead>
</table>
| ABC | P-Glycoprotein | the article ATP binding cassette (ABC) Transporters in this encyclopedia ([9]
| MFS | MdrA, OATs (SCL22 family), OCTs (SCL22 family) | (10, 11, 12) |
| MATE | NorM | ([13]) |
| RND | AcrB | ([14]) |
| DMT | EmrE (SMR family) | ([15, 16]) |

---

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
systems that function as heterooligomers are encoded by a pair of genes that belong both to the SMR family and are located in close proximity or even in an overlapping region in the genome. The only heterodimer characterized thus far is the EmrE transport system from Bacillus subtilis (25).

Members of several RND, MFS, and ABC subfamilies have been proposed to function in combination with a periplasmic membrane fusion protein (MFP) and an outer membrane protein (OMP) to generate a single energy-coupled tripartite efflux system across both membranes of the gram-negative bacteria. The best characterized tripartite systems are MexAB-OprM from Pseudomonas aeruginosa and, by far, AcrAB-ToIC from Escherichia coli.

Insights Into Structures of MDTs

Much research is directed toward deciphering the mechanisms by which MDTs confer resistance against cytotoxic drugs. Such understandings hold promise for tremendous clinical implications. Our current knowledge of the multidrug transport mechanism is based on a wealth of data obtained from extensive research using biochemical and genetic approaches and on recently achieved structural data. In our attempts to understand how these transporters confer multidrug resistance, several questions should be addressed. The first and most obvious question regarding this class of transporters is the question of substrate recognition: What is the architecture of a binding site that enables the binding of numerous dissimilar drugs? Next to be answered are questions regarding the drug translocation pathway across the membrane, the coupling of ion/drug fluxes and conformational changes that must occur throughout the transport cycle. Acquisition of high-resolution structures of membrane proteins using X-ray crystallography is not an easy task, and so, to date, reliable high-resolution structural models were obtained only for a few ion-coupled transporters from the MFS and RND families. No structural information exists to date for transporters from the MATE family. A low-resolution (7 Å) projection map of 2-D crystals of EmrE is the closest to atomic resolution data that exists for the SMR class (26). Based on the combined evidence gathered up to date, several general perceptions can be drawn as to the function mechanism of MDTs.

The first structural evidence that approaches atomic resolution of an MFS transporter was supplied by the interpretation of the three-dimensional structure of the oxalate transporter, OxlT, obtained at the low resolution of 6.5 Å (27, 28). Later on, structures of two proteins from this superfamily were achieved at higher resolution: LacY, the lactose permease (29), and GlpT, the glycerol-3-phosphate transporter (30). While LacY and GlpT are both MFS members, they share a mere 21% sequence identity and differ in their mechanism of function. Although LacY acts as a symporter of pyridine nucleotides and protons across the membrane, GlpT is an organic phosphate/inorganic phosphate (Pi) antiporter. Surprisingly, despite these differences, both structures present a highly similar overall fold. In addition, also the general overall structure of OxlT seems to be very similar to the LacY and GlpT structures. The finding that the fold might be better conserved than the sequence suggests the possibility that a general architecture exists for the 12 transmembrane MFS proteins and that the specific function of each transporter is achieved by subtle changes in the substrate binding site and translocation pathway. Following this line of thought, various structural representations of MFS transporters have been generated based on structural homology to LacY and GlpT, among which we can find models of VMA T (31), TetAB (31), MdfA (32), and rOCT1 (33).

Thus, although neither one of the solved MFS structures is a multidrug transporter, it is reasonable to assume that general insights can still be extracted and applied.

Recently, another structure of an MFS multidrug transporter, the structure of EmrD from Escherichia coli, solved to a 3.5 Å resolution, was published (34). Regrettably, the structure was solved without any substrate bound, and, to date, no biochemical data exist to support the mechanism speculated by the authors. On the other hand, two newly solved structures of AcrB, which is a well-characterized RND multidrug transporter from Escherichia coli, with and without bound substrate, provide valuable information regarding transport mechanisms of tripartite multidrug efflux systems, alongside newly arising questions (35, 36). Functional implications arising from these publications will be discussed in more detail.

Substrate Recognition: Multiple Drug Binding Proteins

Substrate recognition by MDTs is the first and most obvious puzzle that arises when considering the remarkable broad-ranged substrate specificity of these proteins. How can a binding pocket accommodate so many structurally dissimilar substrates? This characteristic of MDTs has long been consid- ered paradoxical because of its ostensible contradiction to basic dogmas of enzyme biochemistry (37, 38).

Soluble regulatory proteins

The first structural information to shed light on the nature of the multi-specific binding pocket came from the solved 3-D structures of soluble multidrug-recognizing proteins. As membrane proteins are not easy to crystallize, the focus of structural analysis attempts has been directed also to the soluble drug-binding proteins that regulate the expression of MDR proteins. The underlying rationale is that the regulatory proteins and the transporters share some substrates, and so the information obtained from these structures would be applicable to the binding sites of the MDTs. BmrR, which is a transcription regulator from Bacillus subtilis, promotes the expression of Bmr, which is an MDR transporter of cationic lipophilic drugs. The structure of BmrR bound to TPP⁺, which is an analog of the coactiva- tor TPP⁺, reveals the drug-binding domain. The drug-binding pocket contains hydrophobic and aromatic residues that inter- act with the drug molecule. The phenyl rings of the substrate molecule interact via van der Waals and stacking interactions with aromatic residues, and an essential electrostatic in- teraction takes place between a negatively charged residue and the...
positively charged substrate. Complementory electrostatic interactions are found between TPSs and carbohydrate groups of several acidic residues positioned in close vicinity (39). Valuable information was also achieved from the solved structures of QacR, which is a Staphylococcus aureus multidrug binding protein that represents the transcription of the QacA multidrug transporter gene. Structures of QacR were obtained in complex with six different lipophilic drugs, both monovalent and bivalent cations and many of them substrates of QacA (40). The structure of QacR exhibits a large drug-binding site with several aromatic residues forming stacking interactions with the substrate’s rings, and several polar residues that mediate drug-specific interaction through hydrogen bonds. Four glutamate residues within the binding pocket maintain electrostatic interactions with the substrate’s positively charged moiety. The different structures of QacA–drug complexes reveal a large, multifaceted drug-binding pocket in which different drugs can bind to different and partially overlapping sets of residues.

**Transport proteins**

With MDTs the only detailed structural evidence is provided by AcrB. The recent structure of AcrB in the presence of substrates reveals a substrate binding pocket in each of the three protomers (35). Bound antibiotic–drug molecules, minocycline and doxorubicin, are observed only in one of the three protomers. The substrate binding pocket is rich in aromatic amino-acid residues, mostly Phe residues. Such aromatic side chains may interact with the drug molecules through hydrophobic or aromatic interactions. Several polar residues located in close vicinity may form hydrogen bonds with the drug molecules. The structure implies a voluminous binding pocket in which different sets of residues are used for binding of the different kinds of substrates. Such a strategy is similar to that previously observed in the regulatory protein QacR.

The critical role of aromatic side chains in the binding of lipophilic substrates is also demonstrated by site-directed mutagenesis studies of other MDTs. Three aromatic residues (Tyr 40, Tyr60, and Trp63) in EmrE, which is an extensively characterized SMR transporter from Escherichia coli, play a role in substrate recognition (41, 42). The importance of aromatic side chains in the equivalent positions has been demonstrated for other SMR proteins as well (43).

A distinguishing feature of MDTs that recognize cationic drugs seems to be the presence of a membrane-embedded negatively charged residue, namely glutamate or aspartate, that neutralizes the positively charged moiety of the drug. EmrE, which is a proton-coupled SMR transporter, has only a single membrane-embedded charged residue, Glu14. Through site-directed mutagenesis studies, Glu14 has been shown to be an essential part of the binding site of both substrate and protons (44, 45). In LmrP, which is an MFS multidrug transporter from Lactococcus lactis, two acidic residues have been shown to be essential for the recognition of the monovalent cation Hoechst 33342, whereas a single negatively charged residue in the binding pocket is sufficient for recognition of the monovalent cation ethidium (38). Also in the soluble regulatory proteins the equivalent of such a buried residue has been visualized. Glu253 of the transcription factor BmrR contributes to the protein–drug interaction through its carboxylate group (39, 47). In QacR, which is another regulatory protein, four partially buried glutamates in the drug-binding site exist, at positions 57, 58, 90, and 120, that surround the drug-binding pocket and interact with the positively charged moieties of the substrates (40).

Another structure of the soluble regulatory protein QacR, this time in complex with bivalent diamidine substrates, reveals a novel manner of drug binding through electrostatic interactions. In these structures the electrostatic neutralization of the positively charged substrate was achieved through drug interactions with the negative dipoles of several oxygen atoms from nearby side chains and the peptide backbone, and not through interaction with the carboxylate of an acidic residue (48). QacA and QacB are both MFS transporters from Staphylococcus aureus that are closely related. Although QacA confers resistance to both monovalent and bivalent cations, QacB confers resistance to monovalent cations only. Site-directed mutagenesis studies provided evidence that the substrate specificity differences are caused by the presence of an acidic residue at position 323 in QacA that plays a critical role in conveying resistance to bivalent cations (8).

In MdfA, however, an MFS multidrug transporter, it has been demonstrated that no single acidic residue plays an irreplaceable role, and although efficient transport of positively charged substrates does require a negative charge at position 26 (either Glu or Asp),-neutralization of this position does not necessarily abolish the interaction of MdfA with cationic drugs (49, 50). In the recently solved structure of AcrB in complex with minocycline and doxorubicin, no acidic residues seem to participate in cationic-substrate binding, and this finding may be a reflection of the fact that substrates of AcrB can be also uncharged and negatively charged (35).

The organic anion (OA Ts) and cation (OCTs) and multidrug transporters share a fundamental trait of polyspecificity. It is interesting to compare, from the knowledge that has accumulated so far, the strategy adopted by each such group of multispecific transporters that allows for broad-range substrate recognition.

The OCTs transport various small amphiphilic organic anions, uncharged molecules, and even some organic cations that are usually around the molecular weight of 400–500 Da and include clinically relevant drugs such as anti-HIV therapeutic agents, anti-tumor drugs, antibiotics, anti-hypertensives, and anti-inflammatories (11). OCTs substrates are organic cations and weak bases that are positively charged at physiologic pH, as well as noncharged compounds, with molecular mass of < 500 Da. Among transported substrates of OCTs are endogenous compounds, drugs, and xenobiotics (12). Many SLC22 family members are expressed in the boundary epithelia of the kidney, liver, and intestines and play a major role in drug absorption and excretion. Substrate translocation by most OCTs is energized by coupling the uptake of an organic anion into the cell to the extrusion of another organic anion from the cell; thus, OCTs use the existing intracellular versus extracellular gradients of anions, such as α-ketoglutarate, as a driving force (11). Similarly, OCT1-3 from the SLC22 family are exchangers of organic cations (12). Several OCT members have been identified and cloned from various eukaryotic origins. Human OCT1-4 and Uralot were more thoroughly studied. 

**Drug Transport in Living Systems**
than others, and so knowledge regarding substrate specificity, drug transport, regulation, and overall characteristics has accumulated. In contrast, a description regarding the functional characteristics of OA T5-9 is still lacking.

In the organic anion transporter rOA T3, site-directed mutagenesis studies lead to the proposal that rOA T3 contains a large binding pocket with several interaction domains responsible for the high-affinity binding of structurally diverse substrates. Two essential membrane-embedded basic residues (Arg 454 and Lys 370) attract negatively charged substrates through electrostatic interactions, but their role in determining specificity is not identical. Interestingly, replacement of the basic amino acids at positions 370 and 454 with the corresponding residues of OCTs, thus generating the mutant R454DK370A, changed the charge selectivity of the protein that could now transport the cation MPP⁺. Several conserved aromatic residues that have been shown to be important for the transport of different substrates by rOA T3 are assumed to mediate substrate recognition through aromatic interactions that could include π–π interactions with aromatic residues. A conservative replacement of Asp475 in rOCT1 to glutamate yielded a protein that exhibited higher affinities for some of its cationic substrates but also impaired transport rates. The authors suggested that the mutation at position 475 from Asp to Glu alters the structure of the cation binding site and harms the translocation step, thus altering the selectivity of rOCT1. Additional studies of the binding site of rOCT1 reveal six more amino acids critical for substrate affinity, among which are aromatic and polar amino acids.

The general feature emerging is that of a multifaceted binding pocket, in which different drugs bind to separate, yet overlapping, sets of residues, many of them aromatic.

**Transport Mechanisms of MDTs**

**AcrB**

Structures of the RND transporter AcrB suggest a possible mechanism for this class of transporters. AcrB functions together with the auxiliary MFP AcrA and the OMP TolC as a tripartite transport system. The AcrB structures reveal an asymmetric trimer, in which each protomer adopts a different conformation according to its proposed role in the catalysis of the transport reaction. The "binding" protomer is occupied by bound substrate, and the "extrusion" protomer is outwardly open in a way that suggests it is the form present just after the extrusion of the substrate through the TolC funnel; the "access" protomer seems to be in the state just before substrate binding. The three different protomer conformations are suggested to represent consecutive states of the transport cycle that result in the guiding of the substrate through AcrB toward TolC. How is the proton gradient coupled to the drug efflux and to the conformational changes that shift the transporters between states? In AcrB it seems that the proton translocation pathway is separate from that of the substrate translocation pathway. Site-directed mutagenesis studies in AcrB and MexB, which is an AcrB homolog from Pseudomonas aeruginosa, show that any replacements of either of the three transmembrane charged residues Asp407, Asp408, and Asp409, and completely abolishes the ability of the protein to confer drug resistance [55, 56]. No other charged residues reside within the transmembrane domains making this triplet the most likely proton translocation pathway. The structure reveals that in the "access" and "binding" protomers Lys940 is adjusted by salt bridges with Asp407 and Asp408, whereas in the "extrusion" protomer, Lys940 is turned toward Thr978 and the salt bridges are eliminated. Additional studies identify also Thr978, which is close to the essential Asp-Lys-Asp network, as important for AcrB's function and as a putative part of the proton relay network [57]. Some unanswered questions remain on the way to grasp fully the AcrB transport mechanism, among which is how this movement caused by protonation and deprotonation is transmitted to the large conformational changes that must occur throughout the transport cycle. Also, the essential role of AcrA and the interaction with TolC have yet to be decoded.

**MFS transporters**

In the case of MFS transporters (Fig. 1), it is tempting to speculate that they may be acting by an alternative access mechanism similar to that suggested originally by Lardéčsky [58] and recently supported for the extensively characterized MFS transporter Lacy, which is a lactose permease that cotransports protons and sugar [28]. The substrate and proton binding sites and translocation pathways are proposed to be separate, and the coupling is transmitted by a series of conformational changes induced on substrate binding. The sugar-binding site located in the approximate middle of the molecule at the apex of a deep hydrophilic cavity has alternating access to either side of the membrane as a result of reciprocal opening and closing of hydrophilic cavities on either side of the membrane.

**EmrE**

The simplest coupling mechanism between drug and proton transport has been described for the MDTs from the SRM family (Fig. 1). A single membrane-embedded charged residue, Glu14 in the case of EmrE, an E. coli SMR, is evolutionarily conserved throughout the SMR family and is essential for function. This residue provides the core of the coupling mechanism because its deprotonation is essential for substrate binding [59, 60]. Conversely, substrate induces proton release, and both reactions (substrate binding and proton release) have been observed directly in the detergent solubilized preparation of EmrE [61]. The estimated pKₐ for Glu14 is unusually increased (about 8.3–8.5) compared with 4.25 for the same carboxyl in aqueous environment. The fact that the binding sites for both substrates and protons overlap is that the occupancy of these sites is mutually exclusive provides the basis of the coupling mechanism [45, 59]. The fine-tuning of the pKₐ is essential because replacement of Glu14 with Asp results in a decrease in the pKₐ of the carboxyl and generates a protein that at physiological pH has already released the previously bound protons, can still bind substrate but cannot longer couple the substrate flux to the proton gradient [44, 61].
SMR proteins function through a binding site shared by protons and substrates, which can be occupied in a mutually exclusive manner and provide the basis for the simplest coupling of two fluxes (15, 59).

Physiologic Roles and Natural Substrates

MDTs provide a survival strategy for living organisms that are constantly assailed by a host of harmful chemicals from the environment. Because of the diversity of these “xenobiotics”, cellular survival mechanisms must deal with an immense variety of molecules. MDTs supply one such strategy. The patterns of abundance of drug efflux systems in different organisms do not correlate directly with the above-suggested role. Although MDTs are highly abundant in the soil and environmental bacteria, they exist in relatively large numbers in other organisms as well, such as intracellular bacteria (5). It is not clear, however, whether this is the primary function of all of those identified as MDTs. It is also not clear why so many different ones are required and why there is such a high redundancy in substrate specificities among different MDTs within the same genome. This may be from the necessity to cover a range as wide as possible of substrates, and overlapping provides a backup strategy in case of failure of one of the systems.

In some cases, a direct correlation is found between MDTs and adaptation to specific environments. The natural environment of an enteric bacterium such as E. coli is enriched in bile salts and fatty acids. Bile acids are amphipathic molecules that...
Drug Transport in Living Systems

Figure 2 The structure of the AcrB-drug complex and proposed mechanism of drug transport (from Reference 54). (a) The complex is observed from the side, with the drug shown as a red hexagon. The dotted line indicates a possible pathway for substrates moving from the cytoplasm. The complex contains three molecules of AcrB, Acr accessory proteins, and the TolC channel to the exterior. The drug is proposed to enter AcrB when it is in the access (A) conformation before binding more closely to the porter domain of AcrB in the binding (B) conformation. It is then transported to the opposite face and is released from the extrusion (E) conformation of AcrB. Transport of the xenobiotic is powered by the proton (H\(^+\)) gradient across the membrane. (b) The proposed ordered multidrug binding change mechanism of the three-unit AcrB complex.

act as emulsifying agents, thus possessing antimicrobial activity. Deletion of the acrAB operon leads to increased susceptibility to bile acids and fatty acids, such as decanoate, which suggests that efflux of these compounds may be one of the physiologic roles of the AcrAB efflux system in E. coli. Consistent with that, the expression of AcrAB was increased by growth in the presence of 5 mM decanoate (62). In another study, both bile and mammalian steroid hormones (estradiol and progesterone) have been shown to be substrates of the two major efflux systems of E. coli, AcrAB-TolC and the MFS EmrAB-TolC (63). CmaABC, which is a tripartite efflux pump from Campylobacter jejuni, was also shown to contribute to the bile resistance of the bacterium (64) and so have similar efflux systems from other bacteria, including Salmonella typhimurium (65) and Vibrio cholerae (66). In some cases, deletion of MDTs decreases the pathogenicity of given microorganisms, most likely because of their impaired adaptability (67). It has been suggested that some bacterial MDTs might be involved in export of signals for cell-cell communication (68). For some MDTs, a specific function has been well documented. BLT, which is an MFS multidrug transporter of Bacillus subtilis, has been shown to function as a substrate-specific transporter of the polyamine spermidine, which is a natural cellular constituent. BLT is encoded in an operon with BltD, a spermidine/spermine acetyltransferase, which emphasizes its possible physiologic role in downregulating cellular spermidine levels through efflux, in concert with an enzyme that chemically modifies spermidine (69).

The vesicular monoamine transporters (VMATs) from the SLC18 family provide another example of polyspecific proteins with a very defined physiologic role. These essential proteins are involved in the storage of monoamines in the central nervous system and in endocrine cells in a process that involves exchange of 2H\(^+\) with one substrate molecule. The VMATs interact with various native substrates and clinically relevant drugs, and they display the pharmacologic profile of multidrug transporters (70).

More recently possible roles in supporting pH homeostasis and alkali tolerance have been suggested for TetL, which is a tetracycline efflux protein from Bacillus subtilis, and the multidrug transporter MdfA (71).

In higher multicellular organisms, the strategy for drug elimination is a multistep process that may involve metabolism, binding to proteins in the circulatory system and binding to specific receptors and excretion processes. A variety of polyspecific transporters (from the ABC, MATE, RND, and MFS superfamilies) are expressed throughout the organism. Some of these transporters provide mechanisms that protect the brain and other sanctuaries from exposure to toxins, both endogenous and exogenous. Others are found in the boundary epithelia of kidney, liver, and intestine and play a major role in drug absorption and excretion.

The discussion of the role of MDTs is tightly connected with the question of the evolution of polyspecificity. One can suggest two possible paradigms of multidrug transporter evolution: The first one states that because MDTs confer broad-range drug
resistance, they evolved to protect cells from environmental "xenobiotics." A second possible scenario is that each MDT has evolved as a substrate-specific transporter, and its ability to transport toxins is only an opportunistic side effect (72).

These issues and the others discussed above provide us with a glimpse of this fascinating and complex world of polyspecific proteins. These ubiquitous proteins in many cases play central roles in survival of organisms by providing the means to remove "xenobiotics" away from their targets. On the other hand, in some cases, they present a serious threat for treatment of drug-resistant cancers and infectious diseases. Understanding of the molecular mechanisms that underly the mechanism of action of these proteins is therefore highly relevant.

Acknowledgment
YE is an Adams Fellow from the Israel Academy of Sciences and Humanities. SS is Mathilda Marks-Kennedy Professor of Biochemistry at the Hebrew University of Jerusalem. Work in the author’s laboratory is supported by Grant NS16708 from the National Institutes of Health, Grant 2003-399 from the United States Israel Binational Science Foundation, Grant 119/04 from the Israel Science Foundation, and Grant PS0 GM73210 from the Center for Innovation in Membrane Protein Production.

References


Further Reading

Membrane_Proteins.html.

See Also

ATP-Binding Cassette (ABC) Transporters
Bacterial Resistance to Antibiotics
Ion Transport
Transport: Antiporters and Symporters
Ion Channels in Medicine

Frank Lehmann-Horn and Karin Jurkat-Rott, Institute of Applied Physiology, Ulm University, Ulm, Germany

doi: 10.1002/9780470048672.wecb260

Our current understanding of ion channels has been made possible by techniques such as electrophysiology and molecular biology, but genetics-based approaches inspired by associated diseases have pointed to regions of functional significance on the ion channels. The ion channelopathies are caused by mutations of corresponding genes or autoantibodies. Although the channels form a highly diverse group (cation, anion, voltage-gated, ligand-gated) and are expressed in excitable and nonexcitable tissues, underlying mutations show recurrent patterns within functionally essential protein parts and lead to common clinical features and almost predictable mechanisms of pathogenesis. This knowledge will prove useful for development of treatment strategies for individuals with various genetic backgrounds and will contribute to both the effectiveness and safety of drugs in the future.

Life’s chemistry of aqueous solutions employs ions as carriers of cell signals. To overcome the impermeable lipid bilayer membrane, the cell invented transporters, channels, and pumps. Ion gradients are established by pumps and serve as batteries of potential energy. Ion channels are membrane-spanning proteins that possess a selectivity filter and a pore that usually is closed by gates in the resting state. The discharge occurs when ion channels open their gates and allow ions to flow down their electrochemical gradients. The cells depolarize when cations diffuse inward or when anions diffuse outward; the cells repolarize when the opposite occurs. Precise control of channel opening and closing is necessary for proper cell excitability and for the production of electrical signals and particularly the action potentials of nerve and muscle. Structures of importance like pore, selectivity filter, voltage sensors, ligand-binding sites, and opening and closing gates were highly conserved for more than 600 million years. Therefore, it is no surprise that an alteration of such functionally important structures can cause a disease.

The episodic clinical features are provoked by environmental factors and caused by abnormal cell excitability, whereas the progressive manifestations that occur in some diseases (periodic paralysis, epilepsy, ataxia, and nephrocalcinosis) are caused by secondary cell degeneration.

### Ion Channels in Disease

The link to human disease came from applying electrophysiological in vitro measurements on diseased muscle cells. These studies showed that an ion channel malfunction could cause a disease (1, 2). Later, the term ion channelopathies was introduced to define this class of diseases characterized by increased or decreased electrical cell excitability (3).

Most channelopathies known to date do not lead to death, but rather they require an abnormal situation, a so-called trigger, to present with recognizable symptoms. Frequently, the attacks can be provoked by rest after physical activity or by exercise itself, hormones, stress, and certain types of food. Most channelopathies have a certain clinical pattern in common: Typically, the symptoms occur as episodic attacks lasting from minutes to days that show spontaneous and complete remission, onset in the first or second decade of life, and—for unknown reason—show amelioration at the age of 40 or 50 years. Surprisingly, many patients with channelopathies respond to acetazolamide, a carbonic anhydrase inhibitor. Most channelopathies show no chronic progression; however, a few exceptions exist. Channelopathies are caused by mutations or by autoantibodies. Although rare, they are important models for frequent disorders.

As the hereditary channelopathies form a group of diseases too diverse to discuss in this short review, we refer to Table 1. In addition, we will describe in more detail some diseases of the largest group of hereditary voltage-gated cation channelopathies, the voltage-gated Na⁺ channelopathies of skeletal muscle, brain, and heart. Furthermore, we will give an example for an autoimmune ligand-gated cation channelopathy and an anion channelopathy.

<table>
<thead>
<tr>
<th>Table 1: Hereditary Voltage-Gated Cation Channelopathies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hypokalemic periodic paralysis (HPP)</td>
</tr>
<tr>
<td>2. Myokymia</td>
</tr>
<tr>
<td>3. Benjamin-Bruton congenital amaurosis</td>
</tr>
<tr>
<td>4. Dystonia / myotonia congenita</td>
</tr>
<tr>
<td>5. Neuronal sodium channelopathy</td>
</tr>
<tr>
<td>6. Hypomagnesemia</td>
</tr>
<tr>
<td>7. Epilepsy</td>
</tr>
<tr>
<td>8. Channelopathy of the heart</td>
</tr>
</tbody>
</table>

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
### Table 1: Overview of hereditary human channelopathies

<table>
<thead>
<tr>
<th>Gene</th>
<th>Locus</th>
<th>Protein</th>
<th>Disease</th>
<th>Trait</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sodium channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCN1A</td>
<td>2q24</td>
<td>Na(_{\text{v}1.1})</td>
<td>SMEI, Familial hemiplegic migraine 3</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>SCN1A</td>
<td>2q24</td>
<td>Na(_{\text{v}1.1})</td>
<td>GEFS(^+)</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>SCN1B</td>
<td>19q13.1</td>
<td>p1 subunit</td>
<td>BENF, Benign familial neonatal/infantile seizures</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>SCN2A</td>
<td>2p23-24</td>
<td>Na(_{\text{v}1.2})</td>
<td>D Gain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCN4A</td>
<td>17q23.1-25.3</td>
<td>Na(_{\text{v}1.4})</td>
<td>Hyperkalemic periodic paralysis PC, PM, Brugada</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>SCN5A</td>
<td>3p21</td>
<td>Na(_{\text{v}1.5})</td>
<td>Long QT syndrome 3</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>SCN9A</td>
<td>2q24</td>
<td>Na(_{\text{v}1.7})</td>
<td>Erythromyalgia, extreme paroxysmal pain, Insensitivity to pain</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>SCNN1B/G</td>
<td>16p12.2-p12.1</td>
<td>ENaC</td>
<td>Liddle’s syndrome, Pseudohypoaldosteronism I</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>Calcium channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CACNA1S</td>
<td>1q31-32</td>
<td>Ca(_{\text{v}1.1})</td>
<td>Hypokalemic periodic paralysis 1</td>
<td>Unclear</td>
<td></td>
</tr>
<tr>
<td>CACNA1C</td>
<td>1q31.3</td>
<td>Ca(_{\text{v}1.2})</td>
<td>Timothy syndrome</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>CACNA1A</td>
<td>19p13.1</td>
<td>Ca(_{\text{v}1.3})</td>
<td>Episodic ataxia 2, spinocerebellar ataxia 6</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>CACNA1F</td>
<td>Xq11.23</td>
<td>Ca(_{\text{v}1.4})</td>
<td>Familial hemiplegic migraine 1</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>CACNB4</td>
<td>2q32-23</td>
<td>p4 subunit</td>
<td>Congenital stationary night blindness</td>
<td>R</td>
<td>Loss</td>
</tr>
<tr>
<td>RYR1</td>
<td>19q13.1</td>
<td>RYR1</td>
<td>Malignant hyperthermia susceptibility</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>RYR2</td>
<td>1q42.1-43</td>
<td>RYR2</td>
<td>Central core disease, multiminicoire disease</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>PKD1</td>
<td>16p13.3</td>
<td>Polycystin-1</td>
<td>Polycystic kidney disease, protein like Ca(_{\text{v}})</td>
<td>Unclear</td>
<td></td>
</tr>
<tr>
<td>Potassium channel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KCNA1</td>
<td>12p13</td>
<td>K(_{\text{v}1.1})</td>
<td>Episodic ataxia 1</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCNC3</td>
<td>1q31.3-1.3</td>
<td>K(_{\text{v}3.3})</td>
<td>Spinocerebellar ataxia</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>KCNE1</td>
<td>21q22.1-22.2</td>
<td>M(_{\text{ink}})</td>
<td>LQT-5</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCNE2</td>
<td>21q22.1</td>
<td>M(_{\text{IRP1}})</td>
<td>Jervell and Lange-Nielsen</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>KCNQ1</td>
<td>11p15.5</td>
<td>K(_{\text{v}7.1})</td>
<td>LQT syndrome inducible, atrial fibrillation</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCNQ2</td>
<td>10q13.3</td>
<td>K(_{\text{v}7.2})</td>
<td>Jervell and Lange-Nielsen</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>KCNQ3</td>
<td>8q24.22-24.3</td>
<td>K(_{\text{v}7.3})</td>
<td>Benign familial neonatal convulsions</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCNQ4</td>
<td>1p34</td>
<td>K(_{\text{v}7.4})</td>
<td>Dominant deafness</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCNQ2</td>
<td>17q25-36</td>
<td>K(_{\text{v}1.1}), HERG</td>
<td>LQTs-2</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCN1</td>
<td>11q24</td>
<td>K(_{\text{v}1.1})</td>
<td>Andersen variant of Bartter</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCN2</td>
<td>17q24.2</td>
<td>Kir2.1</td>
<td>Andersen syndrome</td>
<td>Loss</td>
<td></td>
</tr>
<tr>
<td>KCN11</td>
<td>11p15.1</td>
<td>Kir6.2</td>
<td>Short QT syndrome, atrial fibrillation</td>
<td>Gain</td>
<td></td>
</tr>
<tr>
<td>ABCB8</td>
<td>11p15.1</td>
<td>SUR1</td>
<td>Hyperinsulinemic hypoglycemia</td>
<td>R</td>
<td>Loss</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Locus</th>
<th>Protein</th>
<th>Disease</th>
<th>Trait</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCC9</td>
<td>12p12.1</td>
<td>SUR2</td>
<td>Dilated cardiomyopathy</td>
<td>D</td>
<td>Loss</td>
</tr>
<tr>
<td>KCNMA1</td>
<td>10q22.3</td>
<td>Kᵥ1.1/BK</td>
<td>Epilepsy, paroxysmal dyskinesia</td>
<td>D</td>
<td>Gain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Less-selective cation channels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNCG1</td>
<td>4p12-ce</td>
<td>cGMP-gated</td>
<td>Retinitis pigmentosa</td>
<td>R</td>
<td>Loss</td>
</tr>
<tr>
<td>CNCK1</td>
<td>4q12-ce</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3q27-28</td>
<td>CLC2</td>
<td>Idiopathic epileptic syndrome</td>
<td>D</td>
<td>G/L</td>
</tr>
<tr>
<td>CLCN5</td>
<td>Xp11.22</td>
<td>CLC5</td>
<td>Dent disease</td>
<td>R</td>
<td>Loss</td>
</tr>
<tr>
<td>CLCN7</td>
<td>16p13</td>
<td>CLC7</td>
<td>Osteopetrosis</td>
<td>D/R</td>
<td>Loss</td>
</tr>
<tr>
<td>CLCN8</td>
<td>7q31.2</td>
<td>CFTR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLRA1</td>
<td>5q31.2</td>
<td>GLRA1</td>
<td>Hyperekplexia</td>
<td>D/R</td>
<td>Loss</td>
</tr>
<tr>
<td>GABRA1</td>
<td>5q34-35</td>
<td>GABRA1</td>
<td>GEFS+, Absence epilepsy with febrile seizures, juvenile myoclonic seizures</td>
<td>D</td>
<td>Loss</td>
</tr>
<tr>
<td>GABRG2</td>
<td>5q31.1-33.1</td>
<td>GABRG2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: *G*, gain; L, loss; D, dominant; R, recessive.

The names of the genes and their chromosomal location are given in column 1 and 2; the names of the corresponding proteins are listed in column 3 for voltage-gated (indicated by a subscribed "v") and ligand-gated channels. Columns 4 and 5 list the diseases and their inheritances. The effects of the mutations on the function of the channel complex are categorized in column 6 as gain or loss of function. Gain stands for a gain that leads to a depolarization block. In general, gain-of-function mutations exert dominant effects if they dominate cell function, whereas loss-of-function mutations only cause dominant inheritance if the channel complex is multimeric (dominant negative effect) or if the second gene cannot compensate for it (haploinsufficiency).

Biochemistry of Voltage-Gated Cation Channels

Basic motif of the main cation channel subunit, the so-called α subunit, is a tetrameric association of a series of six transmembrane α-helical segments, numbered S1-S6, connected by both intracellular and extracellular loops, the interlinkers (Fig. 1). The α subunit contains the ion-conducting pore and therefore determines main characteristics of the cation channel complex conveying ion selectivity, voltage sensitivity, pharmacology, and binding characteristics for endogenous and exogenous ligands. Although for Ca²⁺ and Na⁺ channels the α subunit consists of a monomer, K⁺ channels form homotetrmers or heterotetramers because each α subunit consists only of one domain with six transmembrane helices. Accessory subunits termed β, γ, or δ do not share a common structure; some have one to several transmembrane segments and others are entirely intracellular or extracellular. Functionally, they may influence channel expression, trafficking, and gating.

Voltage-sensitive cation channels have at least one open state and at least two closed states, one from which the channel can be activated directly (the resting state) and one from which it cannot (the inactivated state). This implies that least two gates regulate the opening of the pore, an activation and an inactivation gate, both of which are usually mediated by the α subunit. Although activation is a voltage-dependent process, inactivation and the recovery from the inactivated state are time dependent. The voltage sensitivity of cation channels is conveyed by the S4 segments that are thought to move outward on depolarization and channel opening (4). During channel closing, not all voltage sensors move back at once, which generates a variety of closed states that explain the distribution of voltage sensor mutations to...
Voltage-gated Na\(^+\) channel and associated disorders. The \(\alpha\) subunit consists of four highly homologous domains (repeats I–IV) that contain six transmembrane segments each (S1–S6). The S5 loops, S6 loops, and the transmembrane segments S6 form the ion selective pore, and the S4 segments contain positively charged residues that confer voltage dependence to the protein. Note the small, modulatory \(\beta\) subunit. Mutations associated with channelopathies are indicated by conventional 1-letter abbreviations for the replaced amino acids.

Phenotypes in Na\(^+\) channels. The ion-conducting pore is thought to be lined by the S5-S6 interlinkers that contain the selectivity filter. Although the localization of the activation gate may be within the pore, the inactivation gate has been shown to be located in different regions in the Na\(^+\) and K\(^+\) channels.

The crystal structure of a channel clarified the basis for selection of ions that can pass through the open channel pore and the mechanism by which the channel proteins sense changes in transmembrane voltage that control the open or closed conformational states of the channel (5). Thus, investigations of ion channel proteins employ fundamental physics to study the function of biologically critical proteins.

Physiology of Voltage-Gated Cation Channels

At the resting potential, the open probability of voltage-gated cation channels is extremely low, which indicates that very few channels open randomly. Depolarization causes channel activation by markedly increasing open probability. During maintained depolarization, open probability is reduced time-dependently and not voltage-dependently by channel inactivation, which leads to a closed state from which the channels cannot be reactivated immediately. Instead, inactivated channels require repolarization and a certain time for recovery from inactivation. On the other hand, repolarization of the membrane prior to the process of inactivation will deactivate the channel (i.e., reverse activation that leads to the closed resting state from which the channels can be activated). In this simple, approximate model, transitions from one state into another are possible in both directions, which permits also the transition from the resting to the inactivated state at depolarization as well as the recovery from inactivation via the open state (for review see Reference (6)). Forward and backward rate constants for the transitions determine the probabilities of the various channel states.

Hereditary Voltage-Gated Cation Channelopathies

Hereditary diseases of voltage-gated ion channels cover the diverse fields of medicine, myology, neurology, cardiology, and nephrology. As ion channels do not come alone, but rather in whole families of related proteins that conduct each ion type with slightly modified function and varying tissue expression patterns, the underlying mutations are restricted to single genes expressed in a specific tissue such as brain, skeletal muscle, cardiac muscle, sensory tissues, and secretory tissues (Table 1). Examples are myotonia, periodic paralyses, cardiac arrhythmia, long QT syndrome, migraine, episodic ataxia, epilepsy, and nephrocalcinosis. This trick is evolutionary: On one hand, it mediates many functions with the aid of one basic mechanism. On the other hand, it compensates for an eventually disturbed function by closely related channel siblings. The localization of the disease-causing mutations in the various channel proteins and their functional consequences can be similar in these disorders.
Skeletal muscle voltage-gated Na\textsuperscript{+} channelopathies

Clinically, skeletal muscle Na\textsuperscript{+} channelopathies appear as recurring episodes of muscle stiffness or weakness triggered by typical circumstances such as cold, exercise, oral K\textsuperscript{+} load, or drugs. Muscle stiffness, termed myotonia, ameliorates by exercise and can be associated with transient weakness during quick movements that lasts only for seconds. It is the clinical phenotype brought about by uncontrolled repetitive firing of action potentials that lead to involuntary muscle contraction. On the other hand, the weakness is characterized by lack of action potentials or inexcitability.

Several types of inherited myotonias have been observed because of mutations in the Na\textsuperscript{+} channel gene, SCN4A (Table 1, Fig. 1). Potassium-aggravated myotonia (PAM) may be distinguished clinically from other more frequent forms of Na\textsuperscript{+} channel myotonia by its sensitivity to K\textsuperscript{+}. On the other hand, paradoxical myotonia or paramyotonia (PC) worsens with exercise and cold, and it is followed by long spells of limb weakness that last from hours to days. Two other Na\textsuperscript{+} channel disorders are characterized by episodic types of weakness, with or without myotonia, and are distinguished by the serum K\textsuperscript{+} level during attacks of tetraplegia: hyperkalemic or hypokalemic periodic paralysis (HyperPP and HypoPP). All are autosomal and are transmitted dominantly.

For PAM, PC, and HyperPP, the underlying Na\textsuperscript{+} channel pathogenesis mechanism is the same; mutations are present that cause a gating defect of the Na\textsuperscript{+} channel, which leads to slow or incomplete inactivation—a so-called "gain of function" mutation (normally, Na\textsuperscript{+} channels are inactivating rapidly). This mutation results in an increased tendency of the muscle fibers to depolarize. Although Na\textsuperscript{+} influx at slight depolarization generates repetitive muscle action potentials and myotonia, stronger depolarizations lead to general inactivation of Na\textsuperscript{+} channels and to abolition of action potentials, which causes muscle weakness. Heterozygotes have both mutant and wild type channels, but the mutant channel determines the change in cell excitability, and hence, these disorders are dominant. The mutations (Fig. 1) are located mainly 1) in the voltage sensing S4 segment of domain IV, which is suggested to couple the inactivation to the activation process (PC; [7]); 2) in the III–IV interlinker, which is known to contain the inactivation gate (PC and PAM; [8]); and 3) at several intracellular-facing positions involved potentially as acceptor for the inactivation particle (Fig. 2) or involved in steric hindrance of the binding of acceptor and inactivation particle (PAM and HyperPP; [9, 10]). Overlapping clinical phenotypes of the three disorders probably are caused by a similar increase in channel open probability (Fig. 3). The more severe membrane depolarization found in PC and HyperPP correlates with the higher transient intracellular Na\textsuperscript{+} accumulation, whereas depolarization and accumulation are small in PAM ([11]).

Local anaesthetics and antiarrhythmic drugs of Class I, such as mexiletine and lidocaine derivatives, are useful antimyotonic agents for therapeutic treatment of PAM and PC. Their antimyotonic action occurs because they stabilize the inactivated state, which leads to a use-dependent block. The spontaneous weakness typical of HyperPP is not affected by mexiletine because no repetitive action potentials occur. However, for HyperPP, diuretics such as hydrochlorothiazide and acetazolamide can be useful; these drugs decrease the frequency and severity of paralytic episodes by lowering serum K\textsuperscript{+} and by other unknown mechanisms, perhaps, for example, by affecting pH or K\textsuperscript{+} channels.

Voltage-gated Cation Channel

Figure 2 : Hinged-lid model of fast inactivation of Na\textsuperscript{+} channels. Bird’s eye view of the channel that consists of four similar repeats (I–IV). The channel is shown cut and spread open between repeats I and IV to allow a view of the intracellular loop between repeats III and IV. The loop acts as the inactivation gate whose hinge GG (a pair of glycines) allows it to swing between two positions: the open channel state and the inactivated closed state where the inactivation particle IFM (the amino acids isoleucine, phenylalanine, and methionine) binds to its acceptor.
The disease in which mutations have only been found in voltage sensors, are hypokalemic periodic paralyses (HypoPP) types 1 and 2. In both types, episodes of generalized muscle weakness occur occasionally, often during the second half of the night after a day of intensive exercise. Another trigger is a carbohydrate-rich meal. Glucose and released insulin induce a rapid uptake of K⁺ into the muscle fibers. The resulting hypokalemia correlates with the clinical expression of the paralytic attack and gave the disease its name. If no K⁺ is substituted, the weakness can last several hours or days until the serum level is normalized by a hypokalemia-induced rhabdomyolysis or K⁺ retention. As muscle strength is normal between attacks, at least in young patients, the underlying ion channel defect must be well compensated. The intermittent attacks of weakness in HypoPP lead to the requirement of trigger mechanisms. Insulin secretion, as after carbohydrate-rich meals, is one such trigger. Insulin activates the electrogenic Na⁺/K⁺-ATPase; insulin per se and the resulting decrease in [K⁺]o normally lead to a membrane hyperpolarization. In contrast to normal muscle, HypoPP-1 and HypoPP-2 muscle fibers depolarize to ~50 mV at a reduced [K⁺]o of 1 mmol/L and loose force (12). This explains the hypokalemic weakness of the patients.

In HypoPP-1, mutations have been identified in the Caα1.1 voltage sensors (54 segments of repeats II or IV (13, 14). HypoPP-2 mutations are located in 54 segments of Naα1.4 repeats II and III (Fig. 1; (12, 15)). The resulting changes in the pore currents were minor for HypoPP-1 and showed reduced function for HypoPP-1/2 rather than gain of function. Recent results on K⁺ and Na⁺ channels indicate that voltage sensor mutations may create an accessory ion pathway that generates a hyperpolarization-activated cation leak independent of the main channel pore, and that this leak current is responsible for the pathogenesis of the disease (16–18).

Neuronal voltage-gated Na⁺ channelopathies

Of the many neuronal voltage-gated Na⁺ channelopathies (Table 2), only generalized epilepsy with febrile seizures plus (GEFS⁺) and severe myoclonic epilepsy of infancy (SMEI) will be described in more detail. GEFS⁺ is an autosomal dominant childhood-onset syndrome that features febrile convulsions and a variety of afebrile epileptic seizure types within the same pedigree (19). SMEI is characterized by clonic and tonic-clonic seizures in the first year of life that are often prolonged and associated with fever. Developmental stagnation with dementia occurs in early childhood. In contrast to GEFS⁺, the syndrome usually is resistant to pharmacotherapy. Sometimes, patients with SMEI have a family history of febrile or afebrile seizures, and families exist in which GEFS⁺ and SMEI overlap, so that SMEI can be regarded as the most severe phenotype of the GEFS⁺ spectrum (20).

The first genetic defect in this group of diseases was discovered in a large GEFS⁺ pedigree (19). The authors identified a C121W mutation that disrupts the disulfide bridge of the β₁-subunit extracellular loop and leads to a loss of β₁-subunit function.
function. Subsequently, several groups found linkage to a cluster of genes that encode neuronal Na+ channel α-subunits on chromosome 2q21-33 (Table 1). The first two point mutations were detected in SCN1A to predict amino acid changes within the S4 segments of domains II and IV (T875M, R1648H; (21)). Many more SCN1A mutations have been described since then in addition to the mutations in the GABA_A receptor (Table 1). A few of the Na+ channel mutations were expressed in human embryonic kidney cells or Xenopus oocytes that revealed both gain- and loss-of-function mechanisms. Gain-of-function alterations were an acceleration of recovery from inactivation that shortened the refractory period after an action potential for R1648H (22), increased persistent sodium currents predict membrane depolarization for T875M, W1204R, and R1648H (23); a hyperpolarizing shift in window current for W1204R (24); and a reduced channel inactivation upon high frequency depolarizations for D188V (25). Whether the S4 mutations generate a reduced channel inactivation upon high frequency depolarization shift in window current for W1204R (24); and a reduced channel inactivation upon high frequency depolarizations for D188V (25). Whether the S4 mutations generate a hyperpolarization-activated cation leak through an accessory ion pathway has not yet been studied. In contrast, loss-of-function mechanisms were described in part for the same mutations such as enhanced fast and slow inactivation for T875M and R1648H (22, 24) or a depolarizing shift of the steady-state activation curve for I1656M and R1657C (26), which all reduce the amount of available sodium channels. A complete loss-of-function was described for two other GEFS+ point mutations, V1255L and A1685V (26). Hence, loss-of-function mechanisms seem to predominate for GEFS+, which is in agreement with the genetic findings in SMEI as will be outlined below. In contrast to the point mutations found in GEFS+ families, most SMEI patients carry de novo nonsense mutations that predict truncated proteins without function (27). One SMEI point mutation was also shown to yield nonfunctional channels when expressed in human embryonic kidney cells (26). The sodium channel blocker lamotrigine, which is the only drug of this class that is in use in patients with idiopathic generalized epilepsies, is considered the only effective therapy to terminate channel reopenings (33) whereas beta blockers are not effective (34).

Mutations in SCN5A have also been associated with idiopathic ventricular fibrillation and with Brugada syndrome (35). A family history of sudden unexplained death is typical. Individuals with Brugada syndrome often exhibit a characteristic ECG pattern that consists of ST elevation in the right precordial leads, an apparent right bundle branch block, and normal QT intervals. They have an increased risk for potentially lethal polymorphic ventricular tachycardia. Administration of Class I antiarrhythmics and local anesthetics can expose this ECG pattern in latent cases. Accordingly, the proposed cellular basis of Brugada syndrome involves a primary reduction in myocardial sodium current. In fact, most Brugada mutations cause premature stop codons, frameshift errors, and splice site defects, and they are expected to cause nonfunctional channels. Heterologous expression of missense mutations revealed a reduced channel open probability also in agreement with a reduced channel function. Currently, an implantable cardioverter-defibrillator is considered the only effective therapy to terminate ventricular arrhythmias in symptomatic patients with Brugada syndrome (36).

### Acquired or autoimmunecation channelopathies

Many years before the term channelopathies had been established, muscle diseases were identified to be caused by autobody antibodies against ligand- or voltage-gated ion channels. Often, these disorders occur in association with malignancy. For example, many patients with thymoma develop autobody antibodies to the nicotinic acetylcholine receptor (nAChR) of the skeletal muscle.
neuromuscular junction. Normally, ACh molecules can bind to nAChR, which is situated in the neuromuscular junction of the muscle cell membrane and is the first cation channel cloned (37). ACh binding triggers opening of the ion channel and elicits an inward Na\(^+\) and Ca\(^2+\) current with fast rising and decaying phases. The inward current elicits with a high safety factor an action potential that propagates along the muscle fiber membrane and activates excitation-contraction coupling. The Lambert-Eaton syndrome is associated with small cell lung cancer and is caused by antibody binding to specific extracellular epitopes on the voltage-gated P/Q-type calcium channel glycocalyx of the peripheral motor neuron terminus. It is characterized by muscle fatigue (38). Peripheral nerve hyperexcitability (PNH) is a malfunction of the peripheral nerve and leads, in addition to other symptoms, to spontaneous skeletal muscle overactivity. In some patients, PNH is hereditary, whereas in most patients it is acquired and often an autoimmune disorder caused by serum antibodies to voltage-gated, desorbin-in sensitive K\(^+\) channels of peripheral nerves (40). As these autoimmune disorders fulfill the criteria of channel disorders, this group of diseases has been included recently in the broader classification of channelopathies.

Hereditary Anion Channelopathies

Cystic fibrosis (CF) is one of the most frequent genetic diseases with one case in 2000 to 4000 live births. It is a very serious disease with a mean life expectancy of 20 to 40 years. The patients suffer mostly from chronic pulmonary infections that lead to lung destruction, right heart insufficiency, and heart failure. The defective gene product, the cystic fibrosis transmembrane conductance regulator (CFTR), was identified in 1991 (41). The complex molecule functions as Cl\(^-\) channel and as a regulator of other ion channels and transporters. A defect in CFTR leads to reduced NaCl and water secretion in the airways and in other epithelia. In addition, NaCl absorption is enhanced. As a result, the clearance of the airways is impeded, and chronic colonization by pathogenic bacteria such as Pseudomonas aeruginosa leads to airway destruction. CF is one of the first genetic diseases in which genetic therapy is being attempted. However, the initial enthusiasm has been spoiled by the very limited success in animal models and the lack of convincing benefit for the patient. Currently, new approaches for gene and even more so for “classic” therapy are under study. In addition, the CFTR molecule with all its complex functions is target of basic research. It is entirely feasible that the closer understanding of this molecule, its synthesis, maturation, and interaction with other transporters will lead to new and maybe unexpected therapeutic strategies.

Tools and Techniques Used for the Study of Ion Channels

The patch clamp technique

Hodgkin and Huxley unraveled the ionic basis of nerve excitation in the squid giant axon by the first detailed description of the processes of activation and inactivation of voltage-gated sodium and K\(^+\) currents that use the voltage-clamp technique (42). After this technique had become the principal tool for the study of channels for decades, a more recent development has revolutionized this field. The patch clamp technique, developed by Hamill et al. (43), is a specific application of voltage clamp developed to record the current through a membrane patch conducted by a single channel molecule. The patch clamp technique allows for the direct electrical measurement of ion channel currents while simultaneously controlling the cell’s membrane potential. It uses a fine-tipped glass capillary to make contact with a patch of a cell membrane to form a GΩ seal. Originally, this high resistance seal was yielded on skeletal muscle fibers by enzymatic treatment that removed the basal membrane, the glycocalyx, and the connective tissue. Thus, the treated preparation allowed Hamill et al. (43) to attach the glass pipette to the plasma membrane with a leak resistance of 10 to 50 MΩ. Suction resulted in a GΩ seal and enabled the measurements of currents in the sub-pA range with a small noise. Current variants of this technique make possible the application of solution on the exterior and interior of whole cells and on the membrane patches torn from the cell (outside-out or inside-out)—every thinkable configuration of solution and ion channel orientation craved by the ion channel researcher. Usually, primary cultured cells or cell lines are preferred as they reveal a relatively clean surface membrane (44) and require no enzymatic treatment that damages the plasmamembrane. The patch clamp technique is now the gold standard measurement for characterizing and studying ion channels and is one of the most important methods applied to physiology. Single-channel recordings have shown that many channels (e.g., the voltage-gated Na\(^+\) channel) possess only two conductance levels: zero when the channel is closed and a constant conductance when the channel is open. After depolarization, a brief delay occurs before channels open. The intervals are not identical during each depolarization; in fact, the opening and closing of a given single channel is a random process even though the open probability depends on the voltage and is more sensitive to the voltage than an electronic device such as a transistor. After a subsequent short interval, the open time, the current jumps back to zero as the channels close. The stochastic nature can be understood by certain energy barriers that must be overcome before a channel can flip from one conformation (e.g., open) to another (e.g., closed). The energy needed for this purpose comes from the random thermal energy of the system. One can imagine that each time the channel molecule vibrates, bends or stretches, it has a chance to surmount the energy barrier. Each channel is like a binomial trial with a certain probability of success. Clearly, because the protein movements are on a picosecond time scale, the chance of success at each trial must be small, and many trials will be needed before the channel shuts. Usually, a normal Na\(^+\) channel does not reopen even though the depolarization may be retarded by the
Drug Treatment

The functional expression of the mutations in “expression systems” allows one to study the functional alterations of mutant channels and to develop new strategies for the therapy of ion channelopathies (e.g., by designing drugs that specifically suppress the effects of malfunctioning channels). The limitations in throughput are overcome by the automation of the patch clamp technique by which it will become both cost-effective and fast and, at the same time, will enable the highest sensitivity and most accurate description of drug effect compared with any other ion channel drug screening method. Several strategies for the automation of the patch clamp technique have been pursued with set ups already available on the market.

Currently, ion channel modulator drugs account for several billion U.S. dollars in worldwide sales, such as Ca2+ channel blockers (verapamile-type anti-arrhythmics), Na+,K+ channel blockers (lidocaine-type anti-arrhythmics), and Na+ channel blockers (sulphonylurea antidiabetics, amiodarone-type anti-arythmics). Ca2+ channel blockers (verapamile-type anti-arythmics) and Na+ channel blockers (sulphonylurea antidiabetics, amiodarone-type anti-arythmics) are the most accurate description of drug effect compared with any other ion channel drug screening method. Several strategies for the automation of the patch clamp technique have been pursued with set ups already available on the market.

Acknowledgments

We thank for the continuous support of the German Research Foundation (DFG, Ju470/1).

References


Further Reading


Mechanosensitive (MS) channels function as molecular transducers of mechanical forces into electrical and/or chemical intracellular signals in living cells. These channels play a major role in the physiology of mechanosensory transduction encompassing cellular processes that range from regulation of cellular turgor and growth in microorganisms to touch, hearing, and blood pressure regulation in vertebrates. Together with the recent work on MS channels in eukaryotic cells studies of prokaryotic MS channels using a multidisciplinary approach based on the patch-clamp recording, X-ray crystallography, computer simulations, and electronparamagnetic resonance and fluorescence resonance energy transfer spectroscopy, have significantly contributed to our understanding of the relationship between the structure and function in these membrane proteins. In particular, experiments employing chemical manipulation of the physical properties of the membrane lipid bilayer have shed light on 1) how MS channels detect physical forces in cellular membranes, and 2) what causes structural conformational changes in MS channels in response to membrane tension.

Mechanosensitive (MS) channels are the main signaling molecules of mechanosensory transduction. They convert mechanical forces acting on cellular membranes into electrical and/or biochemical signals in various types of nonspecialized cells as well as specialized mechanoreceptor neurons (1). The patch-clamp technique, which allows electrophysiologists to study the activity of single-channel molecules, allowed the first recordings of MS channel currents in real time some 20 years ago (2). Although at some point MS channels were considered to be a possible patch-clamp recording artifact, our knowledge of their structure and function has grown to the point that nowadays these channels are central players in our understanding of protein–lipid bilayer interactions. The cloning of the bacterial MscL and MscS channels, the elucidation of their 3-D crystal structures (Fig. 1a and b), and the demonstration of their physiologic role in bacterial osmoregulation (3) have provided a basis for intensive research and rapid progress in studies of the structure and function in the MS class of ion channels. Furthermore, the cloning and genetic analysis of the mec genes in C. elegans (4), genetic and functional studies of the TRP-type MS channels (5), as well as molecular biologic and functional studies of the TREK-1 family of 2P-type potassium channels (6) have also contributed to our understanding of the role of MS channels in the physiology of mechanosensory transduction and in the pathology of several major human diseases (2). The major aim of this article is to summarize recent developments regarding MS channels, with a major focus on the physico-chemical principles that can account for the stretch sensitivity of these membrane proteins.

### Biologic Background

Living cells are exposed to a variety of mechanical stimuli acting throughout the biosphere. The range of mechanical stimuli extends from thermal molecular agitation to potentially destructive cell swelling caused by osmotic pressure gradients. Since cellular membranes present a major target for mechanical stimuli, MS ion channels are the main signaling molecules designed to detect and translate these stimuli into biologically meaningful signals as illustrated by several examples given below:

1. In their natural environment, bacterial cells need to adapt to a wide range of osmotic conditions. Escherichia coli cells exposed to hypo-osmotic shock respond by a rapid release of cellular osmolytes such as proline, potassium glutamate, trehalose, and ATP. This ability prevents the cells from lysis by decreasing the turgor pressure on the challenge of a sudden shift in osmolarity. Bacterial MS channels, MscL and MscS (Fig. 1a and b), are major components of adaptation mechanisms to hypo-osmotic shock. Being located in the cytoplasmic membrane, MscL and MscS are activated by an increase of membrane tension.
Mechanosensitive Channels

2. Genetic screens of the nematode C. elegans have identified several membrane proteins being required for touch sensitivity of this worm. Four of these proteins, MEC-2, MEC-4, MEC-6, and MEC-10, form a mechanically gated ion channel complex (4). These proteins belong to the DEG/ENaC superfamily of amiloride-sensitive Na⁺ channels of the transporting epithelia with many of them suspected to be gated directly by mechanical stimuli. MEC-2, MEC-4, MEC-6, and MEC-10 proteins have been shown to underlay mechanoreceptor currents in this nematode. Consistent with the role of these channels in mechanotransduction in C. elegans is the finding that mutations in MEC-4 result in touch insensitivity, and dominant mutations in the same gene result in swelling-induced degeneration and lysis of the mechanosensory neurons.

3. TREK-1 channels belong to a superfamily of 2P-domain K⁺ channels. They are polymodal (i.e., gated by a variety of chemical and physical stimuli) K⁺ channels. They are opened by both physical (stretch, heat, voltage, cell swelling, and intracellular acidosis) and chemical stimuli (lysophospholipids, polyunsaturated fatty acids, membrane crenators, and volatile general anesthetics) (Fig. 2). Their main function is to maintain the resting level of membrane potential. Recent studies using TREK-1 knockout mice suggest a central role for TREK-1 in anesthesia, neuroprotection, pain reception, and depression (6).

4. TRPC1 is a member of the canonical TRP (transient receptor potential) subfamily of another large and diverse family of ion channels. TRP channels are expressed in many tissues in numerous organisms where they function as cellular sensors mediating responses to a variety of physical (e.g., light, osmolarity, temperature, and pH) and chemical stimuli (e.g., pheromones, odors, and nerve growth factor) (3). TRP channels function as specialized biologic sensors that are essential in processes such as hearing, vision, taste, and tactile sensation. Several TRP channels may be inherently mechanosensitive, including the TRPC1 channel, which has been identified as MscCa, the Ca⁺ permeable M5 channel in Xenopus oocytes (2).

Figure 1 Physical and chemical stimuli affecting the gating of bacterial MS channels. (A) The structure of the pentameric MscL channel (left) and a channel monomer (right) from E. coli based on the 3-D structural model of MscL obtained by X-ray crystallography has revealed that the channel folds as a homopentamer (7). MscL is activated by membrane stretch, amphipaths (e.g., lysophospholipids, chlorpromazine, and triiodothyronine) and parabens. The channel activity is inhibited by Gd³⁺ and high hydrostatic pressure (HHP) and is modulated by temperature and intracellular pH (3). (B) The structure of the MscS heptamer (left) and the channel monomer (right) from E. coli based on the 3-D structural model of MscS (8) most likely depicting an inactive or desensitized functional state of the channel (3). MscS is activated by membrane stretch, amphipaths, and parabens and is modulated by voltage. The activity of the channel is inhibited by Gd³⁺ and high hydrostatic pressure (HHP) (3). The amino acid point in membrane structures (i.e., channel protein and/or lipid bilayer) affected by the specific stimuli.

Chemistry, structure, and gating mechanism

MS channels are composed of amino acids, which are the building blocks of all proteins. The number of amino acids varies largely between different types of MS channels. For example, a single monomer of the bacterial channel MscL of E. coli is made of 136 amino acids folded in several α-helices connected by loops. A short N-terminal α-helix of the MscL monomer is followed by two transmembrane helices TM1 and TM2 and a C-terminal cytoplasmic α-helical domain (Fig. 2a). The TM1 helix is connected to TM2 by a loop that extends into the pore region and lines the periplasmic side of the channel. A 3-D structure of MscL obtained by X-ray crystallography has revealed that the channel folds as a homopentamer (Fig. 1a) (7).

Diversity and heterogeneity of the MS class of channels is well illustrated by the fact that MscS, the second type of MS channels found in bacteria, differs significantly from MscL in its primary as well as quaternary structure. Each channel belongs to a separate subfamily of the large family of prokaryotic MS channels (3). A monomer of MscS is a small membrane protein of 286 amino acids. A 3-D crystal structure of MscS shows that the functional channel is a homohexameter having a large, cytoplasmic region (Fig. 2b) (8). Each of the
Mechanosensitive Channels

Figure 2. Polymodal activation of TREK-1 by physical and chemical stimuli. TREK-1 is opened by stretch, heat, intracellular acidosis, depolarization, lipids, and volatile general anesthetics, and it is closed by protein kinase A (PKA) and protein kinase C (PKC) phosphorylation pathways. TREK-1 is inhibited tonically by the actin cytoskeleton. The cytosolic carboxy C-terminal domain has a key role in the regulation of TREK1 activity. Phosphorylation of Ser333 by PKA and phosphorylation of both Ser333 and Ser300 by PKC in this region arrest TREK-1 opening. cAMP, cyclic AMP; DG, diacylglycerol; pHi, internal pH (reproduced from Reference 6, with permission).

seven MscS subunits contains three transmembrane domains, TM1–TM3, with N-termini facing the periplasm and C-termini at the cytoplasmic end of the channel. According to the crystal structure, the TM3 helices line the channel pore, whereas the TM1 and TM2 helices constitute the sensors for membrane tension and voltage (3).

Eukaryotic MS channels are far more diverse than the prokaryotic MS channels suggesting that various types of MS channels known today may have become adapted independently and at several occasions during the evolution to specific tasks in mechanosensory transduction of living organisms. In other words, the evolution of MS channels likely converged from independent genetic origins toward a common function of transducing mechanical stimuli into meaningful biologic signals. For example, a monomer of the TREK-1 channel consists of 411 residues. Since its quaternary structure has not been determined experimentally, the functional channel is thought to be composed of two, most likely identical 2P-domains (6). Each 2P domain has four transmembrane segments, an extended extracellular loop, and intracellular N- and C-termini. Although structurally very different from MscL and MscS, the TREK-1 channel functionally closely resembles the bacterial channels. The modality of stimuli that affect its activity (Fig. 2) is very similar to the modality of stimuli affecting the activity of other bacterial channels (Fig. 3a and b). The mechanosensitivity of MscL, MscS, and TREK-1 have been well characterized (3, 6). Consequently, one would expect that a fundamental question on the extent to which the mechanism of gating by physical and chemical stimuli characteristic of prokaryotic and archaean MS channels has been conserved and adapted to gating of MS channels in eukaryotes can be answered fairly by comparing mechanosensitive properties of the bacterial MS channels with TREK-1.

Currently two basic models describe gating of MS channels by mechanical force: the bilayer and the more speculative tethered model (3). A third model of indirect gating of MS channels by mechanical stimuli, which combines elements of the bilayer and tethered models and requires an intermediary mechanosensitive membrane protein to interact with an ion channel, has also been considered (2, 9). According to the bilayer model, the tension in the lipid bilayer alone is sufficient to gate directly the MS channels. This model was proposed initially for the gating of bacterial MS channels. To date, it has been documented in a large number of MS channels from both prokaryotic and eukaryotic organisms (Table 1). Purified MscL, MscS, and other prokaryotic MS channels remain mechanosensitive when reconstituted into artificial liposomes (3). Several eukaryotic MS channels including 2P-type potassium channels TREK-1 and TRAAK, TRP-type channels TRPC1, as well as calcium-dependent stretch-activated potassium channels (SAKCa) have also been shown to be gated by the bilayer mechanism (Table 1). In contrast, the tethered model invokes direct connections between MS channels and cytoskeletal proteins and/or extracellular matrix (ECM) and requires relative displacement of the channel gate with respect to the cytoskeleton or ECM for channel gating. Originally proposed for gating of MscS channels in hair cells and chick skeletal muscle, this model should apply to eukaryotic MS channels in specialized mechanoreceptor cells (2, 10).

The evidence showing that lipids play an essential role in opening and closing not only of prokaryotic MS channels but also of the MS channels of fungi, plants, and animals has led recently to a proposal of a possible unifying principle for mechanosensation based on the bilayer mechanism (11). The main idea of the unifying principle is that forces from lipids gate MS channels independently of their evolutionary origin and
4 results from the bilayer–protein interaction. The hydrophobic mismatch between the protein and the bilayer, which bilayer of artificial liposomes requires energy because of hy-

Notes: The bacterial MS channels, MscL and MscS, are the only MS channel proteins whose 3-D structure has been determined. The gating mechanism and/or physiologic function of some MS channels have not been characterized fully. Note that in contrast to the bilayer mechanism, no single experimental result provides direct support for the tethered model of MS channel gating (1). NT indicates that the effect of amphipaths (CPZ and TNP) has not been tested in the particular type of MS channels (adopted from Reference 10, with permission).

A likely gating mechanism or physiologic function that has not yet been established firmly.

Mechanosensitive Channels

Table 1 Summary of prokaryotic and eukaryotic MS channels identified at the molecular level

<table>
<thead>
<tr>
<th>MS Channel</th>
<th>Source</th>
<th>Gating mechanism</th>
<th>Amphipaths</th>
<th>Physiological function</th>
</tr>
</thead>
<tbody>
<tr>
<td>MscL</td>
<td>Bacteria</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Cellular turgor Growth</td>
</tr>
<tr>
<td>MscS</td>
<td>Bacteria</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Cellular turgor Growth</td>
</tr>
<tr>
<td>MscA1</td>
<td>Archaea</td>
<td>Bilayer</td>
<td>NT</td>
<td>Cellular turgor*</td>
</tr>
<tr>
<td>MscA2</td>
<td>Archaea</td>
<td>Bilayer</td>
<td>NT</td>
<td>Cellular turgor*</td>
</tr>
<tr>
<td>MscM</td>
<td>Archaea</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Cellular turgor*</td>
</tr>
<tr>
<td>MscMJLR</td>
<td>Archaea</td>
<td>Bilayer</td>
<td>No</td>
<td>Cellular turgor*</td>
</tr>
<tr>
<td>MscT</td>
<td>Archaea</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Cellular turgor*</td>
</tr>
<tr>
<td>MEC4</td>
<td>C. elegans</td>
<td>Tether*</td>
<td>NT</td>
<td>Touch</td>
</tr>
<tr>
<td>TREK-1</td>
<td>Brain, heart</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Resting membrane potential</td>
</tr>
<tr>
<td>TRAAK</td>
<td>Brain, spinal chord</td>
<td>Bilayer</td>
<td>Yes</td>
<td>Resting membrane potential</td>
</tr>
<tr>
<td>ENAc</td>
<td>Rat, human, C. elegans</td>
<td>Bilayer/tether*</td>
<td>Yes</td>
<td>Touch</td>
</tr>
<tr>
<td>TRPC1</td>
<td>Xenopus oocytes</td>
<td>Bilayer</td>
<td>NT</td>
<td>Unknown</td>
</tr>
<tr>
<td>TRPY</td>
<td>Fungi</td>
<td>Bilayer*</td>
<td>NT</td>
<td>Cellular turgor</td>
</tr>
<tr>
<td>TRPA1</td>
<td>Hair cells</td>
<td>Tether*</td>
<td>NT</td>
<td>Hearing</td>
</tr>
<tr>
<td>TRPN</td>
<td>Drosophila, zebramice</td>
<td>Tether*</td>
<td>NT</td>
<td>Touch, Hearing</td>
</tr>
<tr>
<td>SAKCa</td>
<td>Chick heart</td>
<td>Bilayer/tether</td>
<td>Yes</td>
<td>Myogenic tone</td>
</tr>
</tbody>
</table>

Notes: The bacterial MS channels, MscL and MscS, are the only MS channel proteins whose 3-D structure has been determined. The gating mechanism and/or physiologic function of some MS channels have not been characterized fully. Note that in contrast to the bilayer mechanism, no single experimental result provides direct support for the tethered model of MS channel gating (1). NT indicates that the effect of amphipaths (CPZ and TNP) has not been tested in the particular type of MS channels (adopted from Reference 10, with permission).

A likely gating mechanism or physiologic function that has not yet been established firmly.

Mechanosensitive Channels

Chemical Tools and Techniques

Several experimental techniques have proven extremely useful in studies of MS channels. Besides the already classic patch-clamp technique used for functional studies of ion chan-

thickness of the bilayer adjacent to the membrane protein will tend to match the length of the protein’s hydrophobic exter-

Fig. 3 because any uncompensated mismatch will add a high energetic cost by exposing hydrophobic groups of amino acid residues to water. The energetic cost of transferring a hydrophobic protein surface from an organic solvent to an aque-
ous environment is approximately 17 mJm$^{-2}$ (1). Hydrophobic matching can be achieved by stretching and/or tilting of the lipid chains, which are more flexible than relatively rigid pro-

tens. Localized changes in bilayer thickness and curvature may also compensate for the mismatch (Fig. 3) (1). Consequently, bilayer dilation/thinning or changes in local membrane curvature could affect an MS-channel protein by shifting the equilibrium between the closed and the open channel conformations. This process has experimentally been demonstrated by reconstituting the MscL channel into phospholipid bilayers of different thick-

ness and examining the channel function by the patch clamp as well as examining its structure by EPR spectroscopy. The experiments showed that although hydrophobic mismatch was not the driving force that could trigger the channel opening, the MscL reconstitution into thin bilayers (e.g., PC16, 16:1 di-
palmitoylphosphatidylcholine) decreased the energy required to open the channel, whereas reconstitution into thick bilayers (e.g., PC20, 20:1 eicosenoylphosphatidylcholine) increased the energy (13). Consequently, although thinning the bilayer is not sufficient to fully open the channel, it contributes to decreasing the energy barrier that has to be overcome to open the channel.

The second possibility of gating MS channels by bilayer tension takes into consideration changes in intrinsic bilayer cur-

vature as a possible trigger for the channel opening (Fig. 3). Experiments on bacterial MS channels suggested that a diverse group of substances with amphiphilic or amphiphilic properties
Mechanosensitive Channels

Hydrophobic Mismatch

Intrinsic Curvature

Figure 3 Possible mechanisms of MS channel activation by bilayer deformation forces. Hydrophobic mismatch and bilayer curvature are considered as deformation forces of pressure-induced changes in the lipid bilayer causing conformational changes in MS channels as indicated by the example of MscL (13). These changes were studied experimentally by reconstituting purified MscL proteins in liposome bilayers prepared from synthetic phospholipid/cholesterol lipid of well-defined composition. The changes in functional properties were examined by the patch-clamp technique, whereas the structural changes were determined by EPR and FRET spectroscopy. (Reproduced from Reference 12, with permission).

such as trinitrophenol (TNP), chlorpromazine (CPZ), local anesthetics, or lysophospholipids (LPC) could trigger gating in these MS channels in the absence of externally applied membrane tension (14). This finding was subsequently confirmed on several prokaryotic (3) and eukaryotic MS currents (10, 15–17). Experiments examining the effect of amphipaths on an MS channel structure and function by EPR spectroscopy and patch-clamp recording side by side were undertaken using again MscL as a model MS channel. Addition of LPC to one monolayer of liposomes reconstituted with MscL generated local stresses leading to redistribution of the transbilayer pressure profile in the lipid bilayer and causing the channels to open without applying membrane tension. These studies thus showed that by chemically modifying membrane lipid content, one could stabilize multiple conformational states in MS channels for a time sufficiently long for biophysical studies of membrane proteins. Furthermore, in contrast to hydrophobic mismatch, insertion of an amphipath into only one monolayer of the membrane bilayer is sufficient to activate fully an MS channel. Together, these findings support theoretical considerations, which suggested that variations in the lateral pressure within lipid bilayers of cell membranes could serve as a mechanism for modulation of protein function (18).

Practical Applications

The following examples illustrate the practical applications in nanotechnology and medicine that have resulted from the basic research on structure and function of MS channels:

1. A light-driven nanovalve was constructed using the MscL channel of E. coli as a template (19). The nanovalve is opened by long-wavelength ultraviolet radiation and can be closed by visible light. For this purpose, MscL was modified by attaching light-sensitive synthetic compounds that undergo light-induced charge separation causing reversible opening and closing of the channel. Such a light-driven nanovalve can, for example, be used in liposome-based drug delivery systems.

2. Bacterial MscL and MscS channels have been shown recently to have potential as selective targets for
novel types of antibiotics (20). Namely, both channels are not only opened by amphipaths (3) but also by parabens, the alkyl esters of p-hydroxybenzoic acid, which seem to interact directly with the gate of the channels (20). The novel antibiotics based on parabens as lead compounds would, therefore, work as openers of MscL and MscS and thus compromise cellular turgor of bacterial cells, which inhibits cell growth and proliferation. In particular, MscL, identified to date in many bacterial pathogens, seems to be an ideal target for the novel antibacterial agents because MscL is highly conserved in prokaryotic cells, but its homologues have not been found in animal and human cells (1). Consequently, an antibiotic targeting MscL could be broad-spectrum and selective, potentially targeting a range of pathogenic bacteria with expected minimal side effects to infected patients.

3. The growing number of human diseases, including cardiac arrhythmias, polycystic kidney disease, Duchenne muscular dystrophy, and tumorigenesis, to name a few, have been associated directly with changes in expression and/or gating of Msc channels (2). The spider venom peptide GsMtx-4 seems to apply to several channels from evolutionary diverse origins. Additional questions to address will be on the structural origins of both diseases.

References


Further Reading


Mechanosensitive Channels


This article deals with the passive transport of electrolytes, water, and small organic molecules across biologic membranes (cell membranes and membranes that confine cellular compartments). Passive transport may be classified into diffusion, facilitated diffusion, and bulk flow. The focus in the current article is on simple diffusion, a process described by the random movement of solutes that results in the net transport (flux) along a concentration gradient. This process, at first, may be described by the first law of Fick. More detailed models take into account the partitioning of solutes between the aqueous media and the barrier itself, as well as the flux resistance that arises from water layers adherent to the barrier, which results in the second law of Fick. The actual extent of the diffusion of molecules across biologic membranes is caused by the interaction between the composition and the structural arrangement of the membrane and by the physico-chemical characteristics of the permeant. For the sake of simplification, the primary structural element of a biomembrane, which is a phospholipid bilayer, can be considered a continuous, lipophilic phase between two aqueous compartments. Commonly used permeability screening approaches include in silico modeling, liposome-based experimental models, cell culture models, and in situ perfusion.

The permeability of compounds through cell membranes is of great interest and importance for the elucidation of many biologic cell functions. Most metabolically important substances are transported across membranes by active transport. Many other intrinsic compounds, as well as most drugs, are known to pass the membrane by passive diffusion.

Biologic Background

Biologic membranes as transport barriers

The most prominent function of biologic membranes is to control selectively the molecular transport into and out of cells. The ability of certain molecules to traverse such biologic barriers depends on their composition as well as their structural and functional features. These features are discussed in more detail on a molecular level elsewhere in this encyclopedia. In brief, the plasma membrane of a cell, as well as any other biologic membrane, is organized basically as a bilayer structure of two sheets composed of phospholipids and neutral lipids, like cholesterol, in which membrane proteins are embedded. The polar head-group of the phospholipids is facing toward the aqueous phases, which are the cytosol on the one side and the medium around the cell on the other, whereas the lipophilic chains of the sheets face to one another to form the center of the membrane.

At first glance, such a membrane might be assumed to be a tight barrier that separates the cell interior from its surrounding or cellular compartments from each other, not allowing for any transport of compounds across this barrier. However, particular molecules need to pass through cell barriers to maintain continuously the basic functionality of the cells; they need to introduce and load substrates as well as remove and discharge toxic compounds. Furthermore, overall cell volume, osmotic pressure, intracellular pH, and ionic composition are controlled by the passage of molecules through the cell membrane.
Types of transport processes across cell membranes

Transport of molecules across biomembranes can be classified mechanistically into the following types: 1) active transport via carriers, 2) passive transport, which comprises a) simple diffusion and b) facilitated diffusion, and 3) endocytosis/transcytosis. Active transport consumes energy through coupling with cellular metabolism. It is accomplished typically via substrate-specific carrier proteins. Therefore, active transport can be maintained against a concentration gradient, which means, for example, certain substances selectively can be picked up independently of their concentration in the cell. Also, a toxic substance may be removed completely from the cell while an appreciable concentration of the same substance still exists in the surrounding medium.

In contrast to active transport, passive transport as a whole does not involve energy consumption and, therefore, only can work down a concentration gradient (or other types of gradients, such as electrochemical potential, thermal, or pressure gradients). In other words, passive transport of molecules equilibrates their chemical potential on both sides of the membrane. The process of passive transport can be subdivided into two different mechanisms: passive diffusion and facilitated transport. Passive diffusion is a physico-chemical process, whereas in facilitated transport, molecules pass through the membrane via special channels or are translated via carrier proteins. Both passive diffusion and facilitated transport, in contrast to active transport, follow a gradient, where facilitation merely lowers the activation energy for the transport process.

In contrast to the above-discussed molecular transport mechanisms, endocytosis, exocytosis, and transcytosis represent the transport of compounds across plasma membranes along with the bulk of their surrounding aqueous medium through vesicle formation (invagination), translocation of the vesicle, and subsequent membrane fusion, a process that also requires energy. These topics are not covered in this article.

Transport across epithelial barriers

Transport across epithelial barriers, such as the gastrointestinal (GI) wall or the blood/brain-barrier, is the result of a series of different, sequential, and parallel processes. Nevertheless, they may be modeled in many cases in the same way as transport across a single membrane.

As a class of tissue, epithelium demarcate body entry points, predisposing a general barrier function with respect to solute entry and translocation. The intestine is lined with enterocytes, which are polarized cells with their apical membrane facing the intestinal lumen that is separated by tight junctions from the basolateral membrane that faces the subepithelial tissues. In addition to their barrier function, the epithelium that line the basolateral membrane that faces the subepithelial tissues.

Passively absorbed compounds diffuse either through the cell itself (transcellular pathway) or in between cells (paracellular pathway). The lipid bilayers of which the mucosal and basalateral epithelial cell membranes are composed of, define the primary transcellular diffusion resistance to solute transport across the intestinal barrier. Transcellular permeability, particularly of lipophilic solutes, depends on their partitioning between intracellular membrane and aqueous compartments (Fig. 3).

Physico-Chemical Approach to Passive Transport

Passive transport is the process of mass transfer across any separating barrier, diaphragm, membrane, or partition “wall” toward equilibrating any differences in chemical potential because of gradients of concentration, electrostatic potential, thermal, or (partial) pressure gradient.

Individual molecules and/or ions (if we, for the moment, disregard the solvation shell) traverse the barrier to equilibrate the gradient. The velocity of the passage of individual molecules affects the overall kinetics of transport with respect to both the lag time (time until steady-state of transport is reached), the time period after which the equilibrium is reached, and—in the case of non-symmetrical systems like cell membranes—the equilibrium state. Properties of both the barrier (membrane) and the diffusant (traversing molecule) are of importance for the mechanism of passage, particularly their interaction on the molecular level. Here, solvation shells and energies of solvation/resolvation play a major role. Solvation states and intermediate states affect activation energies for the passage of the molecules.

Quantitative description of the diffusion process

Let us consider diffusion of molecules between two compartments in the first place, for the sake of simplicity, without a (rate-influencing) barrier between them (Fig. 2, Scheme 2a).

The donor compartment contains a higher concentration of diffusant \( C_D \) compared with the concentration in the acceptor compartment \( C_A \). In other words, a concentration gradient exists. Furthermore, also for the sake of simplicity, we consider transport along a one-dimensional direction.

Definitions

**Flux \( J \)** of a species is the mass \( M \) (or number of molecules of this species) transported per unit time across the barrier, normalized by the cross-sectional surface area \( A \) of the barrier:

\[
J = \frac{\text{d}M}{\text{d}t} \cdot \frac{1}{A} \tag{1}
\]

(dimensions of \( J \): g sec\(^{-1}\) cm\(^{-2}\))

A plot of the mass transported versus time is linear if the donor concentration is virtually constant and the acceptor concentration is virtually zero; the flux is constant when the...
system has reached the equilibrium called steady state (i.e., the linear part in Fig. 3):

$$M = J \cdot A \cdot t$$  \hspace{1cm} (1a)

The slope of the curve corresponds to $J \cdot A$. In most experimental setups, the surface area $A$ is known and flux $J$ can be derived. Constant flux is maintained as long as $C_D$ is much larger than $C_A$ and as long as both are practically unchanged. In biologic systems, a good example for such conditions may be the uptake of xenobiotics from the GI-trait, where the GI tract acts as a reservoir for the molecules, which are transported through the intestinal epithelium into the blood, which acts as a sink.

The quantitative value of flux is widely dependent on the concentration in the donor compartment $C_D$ (figuratively spoken, the pressure). Ideally, $C_D$ should not change over the period of observation; if $C_D$ is not constant, its change with time needs to be considered, which in many cases applies to real situations. However, the concentration in the acceptor compartment should be kept negligibly small (sink conditions) to make calculations easier. Fortunately, in biologic systems, this scenario is very often the case and experimental conditions in most cases can be chosen accordingly. It is handy to define permeability $P$ as the inverse of the resistance of the membrane, which is the flux normalized by the concentration in the donor compartment.

$$P = J / C_D$$  \hspace{1cm} (2)

Permeability, therefore, comes in the dimension of cm/sec, a self-explanatory unit.

Combining Equation 1 and 2 reveals:

$$\frac{dM}{dt} = P \cdot A \cdot C_D$$  \hspace{1cm} (3)

In cases where $C_D$ is not constant over time, the flux $J$ also changes with time. Exchanging mass by concentration, $M = C_D \cdot V$, where $V$ is the concentration of the donor phase, reveals:

$$\frac{dC_D}{dt} = \frac{P \cdot A}{V} \cdot C_D$$  \hspace{1cm} (4)

Setting $k = P \cdot A / V$, integration of Equation 4

$$\int_{C_{D0}}^{C_D} \frac{1}{C_D} \, dC_D = k \int_{0}^{t} \, dt$$  \hspace{1cm} (5)
Passive Diffusion Across Membranes

**Figure 2** Schematic illustration of concentration gradients. \( C_D \) concentration in the donor compartment; \( C_A \) concentration in the acceptor compartment. Scheme 2a: without a physical barrier; Scheme 2b: lipid barrier-controlled; Scheme 2c: water layer-controlled; and Scheme 2d: combined lipid- and water layer-controlled.

**Figure 3** Diagram illustrating mass transport versus time, lag time, and steady-state flux.

The sigma minus plot (i.e., plotting \( \ln C_D \) vs. time) yields permeability as the slope \( k \), where still \( k = P \frac{A}{V} \).

A similar general expression also can be used under non-sink conditions. Here it is useful to describe the velocity of transport (i.e., flux \( J \) ) expressing the effect of the concentration gradient (\( \Delta C \)) and the length of the dimension along the line (\( \Delta x \)), as well as to introduce the diffusion coefficient as the proportionality constant (the Fick law).

\[
J = -D \frac{\Delta C}{\Delta x} \quad (7)
\]

This description is particularly useful because the diffusion coefficient is reasonably well defined in aqueous solutions, it is related to molecular properties in aqueous solutions, and it can be predicted. However, in biologic systems, the observed length \( x \)—for a single transport step—will be widely unchanged. Preferably, steady state and sink conditions are studied, which simplifies the Fick law and focuses on permeability as stated above.

Let us now consider a barrier between the donor and the acceptor compartment, where the barrier has different properties in terms of the solvation of the traversing molecules, as a simplification of a biomembrane. Such a model is shown in Fig. 2, beginning with Scheme 2b, where a homogeneous lipid barrier between the donor and the acceptor compartment is...
Passive Diffusion Across Membranes

introduced. The permeability also is defined in this case as described above and as indicated in the Scheme. Looking closer into the system, the diffusant will partition into the lipid barrier according to the ratio of solubility both in the aqueous phases and in the lipid phase (the partition coefficient, i.e., the ratio of chemical potentials in the two phases). The concentration ratios in the two interfaces will be equal, but not the absolute concentrations, as the concentration in the donor compartment is higher than in the acceptor compartment. The gradient between the two interfaces will make the diffusion occur (see Fig. 2).

The diffusion coefficient in the aqueous phase (donor and acceptor phase) would not play a role if practically no concentration gradient exists within these compartments, in other words, if the diffusion in the aqueous compartments is not rate determining. The partition coefficient between the lipid phases and aqueous phases, $K$, determines whether $C_{D}$ and $C_{A}$, are higher or lower than $C_{L}$ and $C_{A}$, respectively.

If the diffusion of the permeant in the lipid barrier phase is rate determining, then the thickness of the barrier has to be taken into account as well. Here, the permeability coefficient is proportional to its diffusion coefficient $D$ in the membrane (slow diffusion in the membrane).

\[ \text{lag time} = \frac{D}{K C_{D} h} \]  
\[ \text{M} = k \cdot t \]  

Also in this example, flux $J$ is constant during the steady state and permeability $P$ can be revealed. However, here we have to consider a lag time $t_{L}$. For a constant flux, we have to wait until the lipid barrier is saturated and shows a constant gradient of the diffusant according to the partition coefficient. The lag time that represents the intercept with the time axis is found by extrapolating the linear part of the curve (Fig. 3).

\[ \text{lag time} = \frac{D}{K C_{D} h} \]  
\[ \text{M} = k \cdot (t - t_{L}) \]  

Diffusion with respect to physico-chemical characteristics of the permeate

In the beginning of the last century, it was observed that the permeability of a substance across a cell membrane is proportional to the relative partition coefficient of the permeating substance between phases of oil and water. In later years, it was found that the product of the permeability coefficient $P$ and the square root of the molecular weight of a permeant shows a better correlation with the partition coefficient than does permeability alone. This correlation has lead to the idea that permeation is limited not only by the lipid solubility of the permeant but also by a "screen-like" property of the membrane because small molecules penetrate faster than would be predicted from their oil-water partition coefficients. For larger molecules, however, one would expect a product of permeability with the cube root, rather than the square root, of the molecular weight to be more closely correlated with the oil-water partition coefficient. Likewise, the apparent partition coefficient to the membrane plays a major role for the transport processes across the membrane. In strict physico-chemical terms, a membrane is not a phase, but it is most common to treat it as such anyway.

Partitioning of a substance between two phases (e.g., the aqueous phase and the membrane) may be expressed by the difference in Gibbs free energy between the two phases. Major energy differences of the permeant between the two phases are because of the hydrogen-bonding abilities of the substance and the energy of hydration and/or solvation. In addition, nonpolar molecular interactions and entropy terms also contribute to the energy difference. A plot of the logarithm of the partition coefficient versus the number of hydrogen bond donors and acceptor of the permeant gives a straight line (3, 4). The partitioning of a molecule at the water-membrane interface is greatly dependent on its polar nature. The more polar a substance, the higher are the energy expenses for the substance to enter into the membrane phase. For example, the permeability and partition into the membrane decrease approximately exponentially with the number of OH groups of the permeating molecule (3, 4).
Membrane uptake of the noncharged form of a solute is favored over its charged form. As a result, solute membrane permeability versus pH curves is shifted toward the right for weak acids and toward the left for weak bases. This condition also explains the irregular permeability behavior of ion pairs.

**Passive transport of water and protons**

The movement of water and protons/hydroxyl ions across the cell membrane is of particular importance for the existence of the cell (6). Two passive processes move water through membranes: diffusion and hydrodynamic flow. A semipermeable membrane is a membrane that will allow certain molecules (i.e., water) or ions to pass through it by diffusion. If the concentrations of a substance for which the membrane is impermeable are different on the two sides of the membrane, then water will be transported to equilibrate concentrations, for example, a stream of water toward the higher concentration is driven by osmotic pressure.

For the permeation of protons and hydroxyl ions, it has been shown that a prerequisite is the formation of membrane spanning water aggregates in which just the electrons are switched between water molecules. This prerequisite is in contrast to what is found for other ions.

**Simple diffusion of nonelectrolytes**

In the case of nonequilibrium spatial distribution of a certain species, characterized by a nonuniform distribution of chemical potential, diffusion of this species will occur in accordance with the gradient of its chemical potential. If a gradient exists in the chemical potential for one of the species, then a statistical force will be exerted on the particle distribution. The average velocity of a particle will be given by the product of its mobility and the sum of the forces acting on it (4). During the transport of a substance across a membrane, it must move through phase boundaries, such as aqueous phase/membrane/aqueous phase. As mentioned before, the substance has different affinities to each phase encountered. In most cases, the diffusion within the membrane is much slower than in the liquid phase.

So, permeability is proportional to the diffusion coefficient in the membrane and the partition coefficient between the membrane and aqueous phases, while being inversely proportional to the thickness of the membrane. The molecular weight and shape of the permeant are related closely to all of those characteristics.

It is reasonable to expect that nonelectrolytes normally have limited partitioning readily into the lipid barrier. Biologic fluids contain a variety of such charged species that originate from the dissociation of salts, acids, and bases. Another kind of transport, facilitated diffusion, is prevalent for them because their permeation via simple diffusion is rather limited because of limited partitioning into the biomembrane. Facilitated diffusion is a process where permeants diffuse across membranes with the aid of transport proteins. Water-filled pores in the membrane formed by proteins are the main pathway for ions. These pores may be gated so that they can open and close, thus regulating the ion flux. Another kind of membrane proteins, carrier proteins that change shape as the molecules are being carried through, are responsible for the facilitated transport of larger molecules like glucose and amino acids. In contrast to active transport, facilitated diffusion does not require energy and only can carry molecules or ions down a concentration gradient. The transport proteins involved are intrinsic, that is, they span the membrane. They are solute-specific, for example, they have binding sites for the specific molecule or ion that they transport. They also have a defined capacity of how many solutes they can transport. An example of facilitated diffusion is glucose that permeates through cell membranes rather slowly via conventional diffusion because it is not soluble in the membrane. However, glucose passes quickly across a cell membrane via facilitated diffusion.

**Tools and Techniques to Study Diffusion Across Membranes**

Both the prediction and the studying of transport processes through biologic membranes are of great interest in many fields. For example, in drug development, permeability across the intestinal barrier is of essential importance for the destiny of a drug candidate and in molecular biology, knowledge about the pathways in cells of endogenous compounds is of interest. Many different techniques to study such phenomena, thus, have been developed during the years. The complexity of the used techniques ranges from chromatographic systems with covalently bound membrane phospholipid to in situ studies with segments of animals.

**Prediction of diffusion based on experimentally determined physico-chemical characteristics**

Chromatographic methods

Chromatography also can be used to predict permeability. The most widely used methods are immobilized artificial membrane chromatography (IAM) and immobilized liposome chromatography (ILC).

Immobilized artificial membrane chromatography (IAM)

IAM chromatography is a reverse-phase liquid chromatography method that emulates the lipid environment of cell membranes.
by using a chromatographic surface prepared by covalently immobilized cell membrane phospholipids. The technique is experimentally simple and can screen many compounds within a short time. The predominant factor that regulates the passive diffusion across a membrane is its ability to pass through the lipid cell layer, and the capacity factor log K, determined by this technique shows reasonable correlation with permeability across cell monolayers and partitioning into liposomes (8, 9).

Immobilized liposome chromatography (ILC)

Compared with IAM, which uses a monolayer of phospholipid, the liposomal phospholipid bilayers in ILC have the advantage of closely resembling biologic membrane bilayers and constitute a 2-D fluid in which lipid molecules and other components are free to diffuse (10). With this technique, the phospholipid composition can be changed to mimic the membrane of interest. Membrane lipids extracted from human cells also can be used; the technique then is called immobilized biomembrane chromatography (IBC) (11).

Partitioning and distribution (log P/log D)

The lipophilicity of a drug is the most-used single physicochemical property to predict permeability across biologic membranes (12). The most common distribution model is octanol/water (log P) or octanol/buffer (log D, at fixed pH).

Several experiments have demonstrated that a correlation exists between log P and the degree of transcellular absorption of a homologous series of compounds. However, for structurally different compounds, such correlation could not be shown. The lipidosome partitioning system is employed increasingly as an alternative to the IAM approach (13, 14).

In vitro artificial membrane methods

Phospholipid vesicle-based methods

Liposomes

Liposomes can be used to investigate the passive diffusion of compounds across a membrane. Unilamellar liposomes (Fig. 4) can be loaded passively with the compound of interest, the external water phase replaced by an acceptor medium, and the concentration of this compound measured in the medium as a function of time. Liposome studies allow a complete manipulation of solute environment both inside and outside of the vesicles, thus making it a suitable system for passive mechanistic absorption studies. The so-called “stopped-flow technique” can be used to study permeability through liposome bilayers. The major advantage with this technique is that even very rapid permeation processes can be measured.

The distribution coefficient between a liposome phase (see above about the membrane phase) and a water phase also has been used instead of an octanol/water partition coefficient (log P). The liposome phase provides a biomimetic environment to a much larger degree compared with the octanol phase and has shown to the ability to predict in vivo permeability more precisely (13). Lipid bilayer-containing partition systems thus have been considered to model the hydrogen-bonding ability of biologic membranes better than bulk solvents. Biologic lipid membranes are composed of charged lipids that provide a polar environment at the membrane surface, which in biologic systems, often holds a negative charge. A strong attraction of positively charged molecules, thus, is observed. Liposome partitioning, in contrast to octanol/water partitioning, can take into account hydrogen bonding and electrostatic effects. However, the partition coefficient measured in liposome systems may not reflect permeation always. In some cases, it may reflect more drug interactions with the membrane surface (10, 12).

The phospholipid vesicle-based barrier

The phospholipid vesicle-based barrier is a novel predictive medium-throughput screening method for drug permeability, with the use of a tight barrier of liposomes on a filter support (Fig. 5). The liposomes are deposited into the pores and onto the surface of a filter support by centrifugation. Solvent evaporation and freeze-thaw cycling then are used to promote the fusion of liposomes, and a tight barrier is obtained as shown with calcin permeability and electrical resistance (16). The technique is much less labor intensive than comparable cell-based methods and seems, therefore, appropriate for screening of a larger number of drug candidates in earlier stages of drug development. The advantage of this model compared with the parallel artificial membrane permeability assay (PAMPA) models is that the use of liposomes instead of a mixture of phospholipids in an organic solvent leads to a model that resembles the biologic membrane to a larger extent.

Parallel artificial membrane permeability assay (PAMPA)

The PAMPA model is a simple model for fast determination of transcellular in vivo drug absorption. Coating a hydropho-
models are that they underestimate the permeability of actively transported drugs and hydrophilic molecules with low molecular weight and that the use of UV detection excludes drug compound that do not display UV absorbance (8, 9, 20),

Cell-based tools

Isolated membrane vesicles

Cell homogenates precipitated by centrifugation and resuspended in a buffer result in the formation of vesicles, which are mixed with the permeant in the buffer and filtrated after a fixed time. The amount of permeant taken up by the vesicles then is determined. Vesicles offer a unique opportunity to study the properties of drug and nutrient transport at the cellular level. In addition to offering a possibility for separating the brush border and basolateral membranes, membrane vesicles studies allow a complete manipulation of the solute environment both inside and outside the vesicle, thus making it an ideal system for mechanistic absorption studies. Uptake studies can be performed with a small amount of the substance to be tested. A unique advantage for this technique is that the vesicles can be cryopreserved and used for a long duration. The drawbacks are that membrane vesicles are used primarily for studying concentrative processes, for example, active carrier-mediated transport, the need for radiolabeled compounds, and day-to-day variations in the preparation (8, 22, 23).

In vitro cell models

In vitro techniques for the measuring of permeability are less labor- and cost-intensive compared with the more complex in situ and in vivo models. In addition, they are more suitable as screening models because they require smaller quantities of test compound. The capacity of an in vitro system to predict permeability across a biologic barrier depends on how closely the model resembles the in vivo system (8, 23).

Single-cell studies

Cells from animal or human origin can be isolated from different tissue and used as uptake systems. Normally, the isolated cells are suspended in a buffer under O2/CO2 in the presence of the permeant and shaken well. After a certain time period, the cells are separated from the buffer and extracted to determine the amount of permeant inside the cells. Because of the low volume of the cells, the assay is based mostly on use of radiolabeled compounds (23). Cells can be isolated from different origin, and every animal can give rise to experiments with many different compounds or conditions. However, many disadvantages exist as well, like the lack of reproducibility with large variations in viability, because of the chemical or mechanical stress during isolation. This method also only allows for uptake, not transport or permeability, to be studied, but the polarity of the uptake; whether a compound is taken up on the apical or basolateral side cannot be decided (22).

Cell monolayer studies

Several different cell monolayer models that mimic in vivo barriers have been developed and are very popular models in industry as well as academia. The ideal model would be a monolayer of polarized normal human enterocytes, but attempts to isolate and grow them have failed because of the low viability and the complicated requirement for attachment, monolayer formation, and differentiation. However, tumor cells grow

![Figure 5](image1.png) Schematic drawing of the experimental setup used in the permeation studies with the phospholipid vesicle-based barrier (reprinted from Reference 15).

![Figure 6](image2.png) Schematic drawing of the PAMPA permeation cell (reprinted from Reference 15).

Passive Diffusion Across Membranes

![Diagram](image3.png)
than tissue models, but their culture can be labor-intensive and relatively accessible and more amenable to higher throughput. Caco–2 assays are used. The system also is amenable to automation. The major drawback of cell monolayer models is the intrinsic variability that can be seen in permeability data. The results obtained in several laboratories can differ by more than an order of magnitude, probably because of the underlying variability in, for example, culturing techniques, genetic drift, and variable architecture is preserved and closely resembles the in vivo situation and the viability of the tissue can be monitored closely by electrical parameters during the experiment. However, because these techniques also involve the determination of the active transport and efflux mechanisms, we leave these topics to other parts of the encyclopedia.

In vivo and in situ models In vivo studies are performed in unanesthetized animals or humans. Epithelial permeability has been studied in vivo mainly by perfusion techniques such as double balloon catheter. In situ perfusion methods involve a surgical procedure on anesthetized animals whereby a segment of the intestine is isolated and perfused with drug solution. In situ perfusion of intestinal segments frequently is used to study the permeability and absorption kinetics of drugs. The biggest advantage of the in situ system compared with the in vitro tissue systems is the presence of an intact blood and nerve supply in the experimental animals (8). But because these models are dealing with active transport mechanisms, a more detailed presentation would be beyond the scope of this article.

References
2. Vávra J. Dipeptidomimetics as Pro-Moieties for hPEPT1 Targeted Prodrugs. In: Department of Pharmacy, University of Tromsø, 2004.

Figure 7 Schematic drawing of a cell monolayer assay with the cell monolayer on a filter; support separates the apical and basolateral chamber from each other (reprinted from Reference 15).
Passive Diffusion Across Membranes


Further Reading


See Also

Drug Transport
Ion Channels and Pores, Chemical Biology of Small Molecule Transport
Transport in Cells, Mechanisms of Water Channels
Chemistry of Lipid Homeostasis

Paul N. Black, Whitney Sealls and Concetta C. DiRusso,
Center for Metabolic Disease, Ordway Research Institute and Center for Cardiovascular Sciences, Albany Medical College, Albany, New York
doi: 10.1002/9780470048672.wecb284

Lipid metabolism encompasses a large part of intermediary metabolism and includes highly dynamic processes, which are subject to regulation at transcriptional and enzymatic levels. The maintenance of lipid metabolism or lipid homeostasis is multilayered and includes biosynthesis and degradation, the impact of external stimuli proceeding through signal transduction, the function of lipid-responsive transcription factors, and lipid trafficking. It also includes the production of bioactive lipids, which have regulatory roles in different lipid metabolic pathways, effect specific transcription factors, and influence a myriad of cellular processes.

Lipid metabolism is responsive to external supplies of fatty acids and other precursors required for neutral lipid, glycerol phospholipid, and sphingolipids synthesis. In some cell types, fatty acid composition within membrane lipids must necessarily adapt to changes in temperature to maintain appropriate fluidity essential for biologic activity. Finally, lipids are diverse in both structure and abundance and differ among animals, plants, fungi, and bacteria. This article will focus on the fatty acid metabolism and, to some extent, its relationship to sterol metabolism, which underpins lipid homeostasis.

Lipid homeostasis is maintained through a complex network of hormonal, neuronal, and environmental regulators. Inability to modulate lipid metabolism to maintain homeostasis is a hallmark of many disease states, including metabolic syndrome, obesity, diabetes, cardiovascular disease, and some cancers. Together, these diseases are the leading causes of morbidity and mortality in the United States and most other developed countries. It is hypothesized that high circulating levels of free fatty acids, occurring as a result of various metabolic disturbances, lead to excessive fatty acid internalization and the resultant lipotoxicity of normal cells and tissues. Furthermore, free fatty acids and activated fatty acids (fatty acyl-CoAs) are toxic to cells because of their hydrophobic properties, which dissolve membranes. These compounds interfere with enzyme function and, when in excess, serve to increase both oxidation rates and the synthesis of ceramide, which in turn leads to increases in cell death by apoptosis. The correlation between chronically increased plasma free fatty acids and triglycerides with the development of obesity, insulin resistance, and cardiovascular disease has led to the hypothesis that decreases in pancreatic insulin production, cardiac failure, and cardiac hypertrophy are from aberrant accumulation of lipids in these tissues (1).

Biologic Background

Lipids represent a diverse group of compounds, which are readily soluble in organic solvents such as chloroform or toluene and are essential for the structure and function of all living cells. Lipids include oils, fatty acids, waxes, steroids (e.g., cholesterol and steroid hormones), and other related compounds. Specific classes of lipids are the primary structural components of membranes, provide sources of metabolic fuel, and function as bioactive signaling molecules.

Classes of lipids

Fatty acids, the simplest class of lipids, are carboxylic acids containing a long aliphatic tail generally consisting of an even number of carbon atoms, which are either saturated or unsaturated (Table 1). Natural fats and oils generally have
fatty acids with at least eight carbon atoms [e.g., caprylic acid (C8:0, n-octanoic acid)]. Saturated fatty acids (SFAs) lack double bonds and, as a result, can form straight chains, which can be tightly packed. Common SFAs include butyric (C4:0, n-Butyanoic acid), lauric (C12:0, n-Dodecanonic acid), myristic (C14:0, n-Tetradecanonic acid), palmitic (C16:0, n-Hexadanonic acid), stearic (C18:0, n-Octadecanonic acid), and arachidonic (C20:0, n-Eicosanonic acid). Monounsaturated fatty acids (MUFA’s) contain one double bond along the chain, which generally occurs in a cis-configuration. Although some monounsaturated fatty acids have a trans-configuration, most found in nature are cis. This cis-configuration results in a fatty acid, which when part of a phospholipid molecule functions to increase membrane fluidity. Common MUFA’s include palmitoleic acid (C16:1n7), cis-9-Hexadanonic acid and oleic acid (C18:1n9), cis-9-Octadecanonic acid. Polyunsaturated fatty acids (PUFA’s) contain two or more double bonds, each in a cis-configuration. These fatty acids include the essential fatty acids in humans, linoleic acid (LA, C18:2n6), cis-9,12-Octadecatrienic acid, and α-linolenic acid (LNA, C18:3n3). Monounsaturated fatty acids (MUFA’s) and polyunsaturated fatty acids (PUFA’s) are the primary dietary sources of essential fatty acids. Essential fatty acids support the cardiovascular, reproductive, immune, and nervous systems, including the production of prostaglandins, which regulate body functions such as heart rate, blood pressure, blood clotting, fertility, and conception, and are essential in the inflammatory response.

Fatty acids can be esterified to glycerol to form diglycerides, which can enter the phospholipid biosynthetic pathway or, with the addition of a third fatty acid, can be converted into triglycerides (TAG). TAGs are the predominant storage lipid but are also present in blood plasma and, in association with cholesterol, are components of lipoprotein particles. Fatty acids can be esterified to glycerol to form diglycerides, which can enter the phospholipid biosynthetic pathway or, with the addition of a third fatty acid, can be converted into triglycerides (TAG). TAGs are the predominant storage lipid but are also present in blood plasma and, in association with cholesterol, are components of lipoprotein particles. Fatty acids can be esterified to glycerol to form diglycerides, which can enter the phospholipid biosynthetic pathway or, with the addition of a third fatty acid, can be converted into triglycerides (TAG). TAGs are the predominant storage lipid but are also present in blood plasma and, in association with cholesterol, are components of lipoprotein particles. Fatty acids can be esterified to glycerol to form diglycerides, which can enter the phospholipid biosynthetic pathway or, with the addition of a third fatty acid, can be converted into triglycerides (TAG). TAGs are the predominant storage lipid but are also present in blood plasma and, in association with cholesterol, are components of lipoprotein particles. Fatty acids can be esterified to glycerol to form diglycerides, which can enter the phospholipid biosynthetic pathway or, with the addition of a third fatty acid, can be converted into triglycerides (TAG). TAGs are the predominant storage lipid but are also present in blood plasma and, in association with cholesterol, are components of lipoprotein particles.

### Table 1: Chemical names and descriptions of common fatty acids

<table>
<thead>
<tr>
<th>Common name</th>
<th>Nomenclature</th>
<th>Carbon skeleton</th>
<th>Systematic name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saturated Fatty Acids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Butyric acid</td>
<td>C4:0</td>
<td>n-Butyanoic acid</td>
<td></td>
</tr>
<tr>
<td>Caprylic acid</td>
<td>C6:0</td>
<td>n-Octanoic acid</td>
<td></td>
</tr>
<tr>
<td>Caprylic acid</td>
<td>C8:0</td>
<td>n-Octanoic acid</td>
<td></td>
</tr>
<tr>
<td>C10:0</td>
<td>n-Decanonic acid</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lauric acid</td>
<td>C12:0</td>
<td>n-Dodecanonic acid</td>
<td></td>
</tr>
<tr>
<td>Myristic acid</td>
<td>C14:0</td>
<td>n-Tetradecanonic acid</td>
<td></td>
</tr>
<tr>
<td>Palmitic acid</td>
<td>C16:0</td>
<td>n-Hexadanonic acid</td>
<td></td>
</tr>
<tr>
<td>Stearic acid</td>
<td>C18:0</td>
<td>n-Octadecanonic acid</td>
<td></td>
</tr>
<tr>
<td>Arachidonic acid</td>
<td>C20:4</td>
<td>n-Eicosanonic acid</td>
<td></td>
</tr>
<tr>
<td>Behenic acid</td>
<td>C22:0</td>
<td>n-Eicosanonic acid</td>
<td></td>
</tr>
<tr>
<td>Lipoic acid</td>
<td>C24:0</td>
<td>n-Tetracanonic acid</td>
<td></td>
</tr>
<tr>
<td>Monounsaturated Fatty Acids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palmitoleic acid</td>
<td>C16:1n7</td>
<td>cis-9-Hexadanonic acid</td>
<td></td>
</tr>
<tr>
<td>Oleic acid</td>
<td>C18:1n9</td>
<td>cis-9-Octadecanonic acid</td>
<td></td>
</tr>
<tr>
<td>Vercenic acid</td>
<td>C18:1n11</td>
<td>cis-11-Octadecanonic acid</td>
<td></td>
</tr>
<tr>
<td>Gadonic acid</td>
<td>C20:1n9</td>
<td>cis-9-Eicosanonic acid</td>
<td></td>
</tr>
<tr>
<td>Polyunsaturated Fatty Acids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linoleic acid (LA)</td>
<td>C18:2n6</td>
<td>cis-cis-9,12-Octadecadienic acid</td>
<td></td>
</tr>
<tr>
<td>n-3 Linoleic acid (LNA)</td>
<td>C18:3n3</td>
<td>cis-cis-cis-9,12,15-Octadecatrienic acid</td>
<td></td>
</tr>
<tr>
<td>Highly Unsaturated Fatty Acids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eicosapentaenoic acid (EPA)</td>
<td>C20:5n3</td>
<td>cis-cis-cis-5,8,11,14,17-Eicosapentaenic acid</td>
<td></td>
</tr>
<tr>
<td>Docosahexaenoic acid (DHA)</td>
<td>C22:6n3</td>
<td>cis-cis-cis-5,8,11,14,17,20-Eicosapentaenic acid</td>
<td></td>
</tr>
</tbody>
</table>
sphingosine backbone can be O-linked to a charged head group such as choline. Fatty acids are attached through an amide linkage to sphingosine. Ceramides, sphingomyelins, and glycosphingolipids, which differ in their head groups, are the predominant forms of sphingolipids and function as structural components of membranes and in signal transmission and cell recognition. Finally, fatty acids are stored esterified to cholesterol in lipoprotein particles. Cholesterol is the major sterol found in plasma membranes and, in addition, provides essential functions in lipoprotein metabolism and lipid trafficking and bile acid metabolism. Readers are directed to Vance and Vance (2) and to the section entitled “Lipid Synthesis” in this series regarding the details of phospholipid, sphingolipids, and lipoprotein synthesis, including structure.

Fatty acid and fatty acid derivatives function as signaling molecules

Some of the earliest indications that fatty acids and fatty acid derivatives function as signaling molecules come from studies using the gram-negative bacterium Escherichia coli. The long chain acyl-CoA (LC-CoA) formed as a consequence of coupled transport-activation serves primarily as a substrate for β-oxidation, although it can be incorporated into membrane phospholipids. One particular feature of LC-CoA is that it functions as a bioactive lipid that governs the DNA-binding properties of a specific transcription factor (FadR), which negatively regulates genes required for fatty acid degradation and positively regulates key genes required for fatty acid biosynthesis. Studies summarized by DiRusso and Black (3), which culminated in three crystal structures of FadR, alone in the presence of DNA with the consensus-binding site and in the presence of myristoyl-CoA, clearly defined the bioactive role of LC-CoA as an important effector governing lipid metabolism in E. coli.

In mammalian systems, the HUFA s derived from LA and LN yield an array of compounds with unique roles in a variety of cellular metabolic and signaling events (see “Prostaglandins” in this series). AA and EPA-derived eicosanoids mediate pleiotropic responses including, for example, pain, vasodilatation, vasodilation, inflammation, and bronchoconstriction. Some HUFA s and eicosanoids are ligands for transcription factors, particularly members of the peroxisomal proliferator activated receptor family (PPAR) involved in fatty acid and triacylglycerol metabolism and adipocyte differentiation. Certain classes of highly unsaturated fatty acids also decrease the activity of the Steroid Receptor Element-Binding Protein (SREBP) family members, which activate genes required for LDL uptake, cholesterol, and fatty acid synthesis.

Dynamics of Fatty Acid Homeostasis

Lipid homeostasis is maintained through the multilayered regulatory networks of lipid metabolism, transport, and signal transduction. These processes are dynamic and respond to nutritional and environmental cues, which includes interplay between fatty acid synthesis, fatty acid degradation, and complex lipid and cholesterol synthesis, which are influenced by numerous cellular processes. Metabolic output to maintain homeostasis must be quickly adjusted to meet and adapt to the nutritional needs of the cell. In the context of lipid metabolism, it includes well-defined regulatory processes at both the enzyme and transcriptional levels. Inability for a cell or organism to maintain lipid homeostasis has been suggested to cause a number of diseases, including obesity, diabetes, atherosclerosis, and some cancers.

Fatty acid synthesis

Mammalian fatty acid synthesis is a complex process being partly governed by de novo pathways and also requiring essential dietary contributions (Fig. 1a). Quantitatively, de novo mammalian fatty acid synthesis produces primarily palmitate (C16:0) and oleate (C18:1), which constitute the bulk of the side chains of membrane lipids, storage TAGs, and sterol esters. Short and medium chain fatty acids are synthesized in lesser amounts but have some essential functions in cellular metabolism and growth. For example, myristate (C14:0) is added to proteins in posttranslational modification reactions and is often required for membrane association of proteins, many of which are involved in complex signal transduction events. Very long chain fatty acids (VLFA; ≥ C20) are essential components of sphingolipids, glycosylphosphoinositides, and the lubricating waxes of skin and hair. The essential dietary fatty acids include linolenic acid (LA, C18:3 ω3), linoleic acid (LN, C18:2 ω6), and alpha-linolenic acid (LN, C18:3 ω3). Mammals are missing a fatty acid desaturase capable of introducing a double bond between C9 and the methylene end of the fatty acid. Hence, to synthesize essential HUFA s, LA and LN are both required in the diet (Fig. 1b). LA is the precursor to arachidonic acid (AA, C20:4 ω6) and derivative eicosanoids, whereas LN is the precursor to eicosapentaenoic acid (EPA, C20:5 ω3), docosahexaenoic acid (DHA, C22:6 ω3), and ω3-derived derivatives (Fig. 1b).
ecosanoids, sometimes called resolvins (4). EPA and DHA are required in large quantities during embryonic and neonatal development, particularly for the formation of neural tissues (5). Some current evidence indicates the de novo synthetic pathway for DHA and EPA may be too limited to meet the growth and development requirements of infants and dietary supplementation is recommended for nursing mothers and synthetic formulas.

Four enzyme families are required for fatty acid synthesis. The first, acetyl-CoA carboxylase (ACac), synthesizes malonyl-CoA from acetyl-CoA and CO₂. Malonyl-CoA is the substrate for two other protein families, the fatty acid synthases (Fasn) and the fatty acid elongases (Elovl). The length of the acyl chain is determined by the specificity of the terminal reaction catalyzed by the thioesterase active site of Fasn. The mammalian liver Fasn synthesizes primarily palmitate. The palmitate prolongs the steps to condense and reduce each 2-carbon unit added (7).

The HUFAs (AA, EPA, and DHA), required by mammals, are synthesized by the microsomal elongases and Δ5 and Δ6 desaturases (8). The substrates for synthesis, LA and LN, are obtained primarily from vegetable oils in the diet. Elongation and desaturation of LA produces AA, whereas elongation and desaturation of LN produces EPA and DHA. Synthesis of DHA, which is concentrated in the brain, proceeds through the elongation and desaturation pathways to the product C24:6ω3, which is then the substrate for one cycle of peroxisomal β-oxidation to yield C22:6ω3 (8).

The liver is quantitatively the most important site of de novo fatty acid synthesis and processing and modification of dietary fatty acids. In liver, fatty acids are incorporated into phospholipids, triglycerides, and cholesterol esters, packaged in very low density lipoproteins, and exported to blood. Uptake of fatty acids in extrahepatic cells requires degradation of the complex lipids into component parts by lipases. Fatty acids are then imported by protein-mediated transport mechanisms involving fatty acid transport protein (FATP), fatty acid translocase (FAT/CD36), and long chain acyl-CoA synthetase (A-ACS) (9). This process is poorly understood, although efforts are underway to define the components and mechanisms mediating the selectivity and efficiency of uptake into cells and tissues (10). Fatty acids are incorporated into complex lipids through de novo synthetic and remodeling pathways. As detailed below and shown in Figure 2a, intracellular pools of acyl-CoA are involved in processes outside of lipid metabolism and, in many instances, function as important regulatory molecules. Figure 2b illustrates an overview of glycerol phospholipid synthesis and how fatty acids in the form of acyl-CoA enter these metabolic pathways. Readers are referred to the article entitled “Lipid Synthesis” in this series for specific details regarding these pathways.

β-oxidation of fatty acids

Fatty acids represent an important source of metabolic fuel, especially for cardiac and skeletal muscle, and are degraded by β-oxidation. Before β-oxidation, fatty acids are first activated by acyl-CoA synthetase (A-ACS). This activation step proceeds through an acyl adenylate intermediate and results in the formation of acyl-CoA. These molecules must first traverse the mitochondrial membrane before further metabolism. Acyl-CoA is converted to acyl carnitine by carnitine palmitoyltransferase 1 (CPT 1) on the inner surface of the outer mitochondrial membrane, which is then transported across the inner mitochondrial membrane by carnitine-acylcarnitine translocase into the exchange for carnitine. Carnitine palmitoyltransferase 2 (CPT 2), located on the matrix side of the mitochondrial inner membrane, converts acylcarnitine back to acyl-CoA, which then enters the β-oxidation cycle. During each cycle of β-oxidation, the products are acetyl-CoA, NADH, and FADH2. The first step in β-oxidation is an oxidation step, which is catalyzed by acyl-CoA dehydrogenase (ACD) and yields FADH2. This step is followed by a hydration step catalyzed by enoyl-CoA hydratase and a subsequent oxidation step catalyzed by hydroxyl acyl-CoA dehydrogenase yielding NADH. A acyl-CoA is liberated by ketolactone after cleavage of the bond between the ω-2 and γ-carbons with the addition of CoA to the shortened fatty acid, which undergoes another round of β-oxidation. Regulation of β-oxidation is largely controlled by the concentration of free fatty acids available. Malonyl-CoA, a key intermediate in fatty acid biosynthesis, inhibits CPT 1, thereby inhibiting the transport of acyl-CoA into mitochondria, which represents an important metabolic regulatory switch and functions to modulate the fatty acid biosynthetic and fatty acid oxidative pathways in such a manner that these processes will not occur together. Very long chain fatty acids are initially oxidized in the peroxisome where the initial oxidation step is catalyzed by acyl-CoA oxidase and the subsequent steps in β-oxidation are catalyzed by a multi-enzyme complex with hydratase, dehydrogenase, and thiolase activities. Unsaturated fatty acids require additional enzymatic activities, including enoyl-CoA isomerase and dienoyl-CoA reductase. Readers are directed to Vance and Vance (2) for additional details regarding β-oxidation, including the details of the metabolic reactions.

Fatty acid trafficking

Fatty acids are very apolar compounds and readily partition into a lipid bilayer and biologic membrane. Given the expense to synthesize fatty acids, most cell types will take up exogenous fatty acids, which as acyl-CoA, enter pathways of β-oxidation and complex lipid and TAG synthesis. The concentration of free fatty acids in the circulation, extracellular milieu, and within cells is extremely low as a consequence of their relative insolubility under aqueous conditions. Cells with high levels of fatty acid metabolism (either degradation or storage) transport exogenous fatty acids at higher rates when compared with those with low levels of lipid metabolism. In a number of cell types, fatty acid transport is inducible and commensurate with the expression of specific sets of proteins thought to participate in this process (11). Insulin regulates fatty acid uptake.
Figure 2  (A) Acyl-CoA metabolism within the cell. Fatty acids must be activated to acyl-CoA before metabolism. As shown, acyl-CoA pools derive from vectorial acylation [import/activation], lipolysis of triglyceride stores, chain turnover from complex lipids, and de novo biosynthesis. Acyl-CoAs are important signaling molecules, incorporated into complex lipids and triglycerides, are substrates for energy production through β-oxidation, and are substrates for protein N-myristoylation and palmitoylation. Acyl-CoAs are also subject to chain modification, including desaturation and elongation. Acyl chains in complex lipids serve both structural and regulatory roles. (B) Overview of glycerol lipid metabolism in eukaryotes, with an emphasis on how fatty acids as acyl-CoAs enter these pathways. Glycerol-3-phosphate is converted into phosphatidic acid (PA) through a two-step process, with the intermediate monoacylglycerol phosphate (MAG-P). PA serves in the synthesis of diacylglycerol (DAG) or CDP-diacylglycerol (CDP-DAG). The former is required in the CDP-choline (CDP-C) and CDP-ethanolamine (CDP-E) pathways in the synthesis of phosphatidylcholine (PC) and phosphatidylethanolamine (PE). DAG also serves in the synthesis of triacylglycerol (TAG). PE is converted into phosphatidylycerine via phosphatidylethanolamine-serine transferase; the reverse reaction requires phosphatidylycerine decarboxylase. CDP-DAG is involved in the biosynthesis of phosphatidylinositol (PI), phosphatidylglycerol (PG), and cardiolipin (CL). The asterisks show where acyl-CoA is a substrate in the reaction, including the conversion of lysoPC (LPC), lysoPE (LPE), and lysoCL to PC, PE, and CL, respectively. Also shown is the requirement of acyl-CoA for the generation of cholesterol esters (CHO-FA) from cholesterol (CHO).
is some indication that FAT/CD36 deficiency contributes to the etiology of hereditary hypertrophic cardiomyopathy (14). Genetic linkage studies suggest a deficiency of FAT/CD36 is associated with hypertriglyceridemia and hyperinsulinemia in the spontaneously hypertensive rat. The most informative data on FAT/CD36 derives from studies of transgenic overexpressing and knockout mice. The over expression of FAT/CD36 in transgenic mice results in slightly lower body weight than control litter mates, reduced levels of triglycerides (LDL fraction), and elevated levels of circulating fatty acids. Mice with engineered deletions in the gene encoding FAT/CD36 are viable, yet have a significant decrease in binding and uptake of oxidized low density lipoprotein in peritoneal macrophages. These animals also have significant increases in fasting levels of cholesterol (HDL fraction), nonesterified free fatty acids, and triacylglycerol (LDL fraction). Each of these phenotypes is consistent with alteration in lipid trafficking pathways but do not necessarily indicate a deficiency in cellular uptake of free fatty acids.

Fatty acid transport protein 1 (FATP1) and long chain acyl-CoA synthase 1 (ACSL1) were identified as components of a fatty acid transport system in expression cloning experiments, which specifically targeted cellular fatty acid uptake (15). A sub set of FATP and ACSL 1 were selected in these experiments, it was suggested that fatty acid transport occurred by a coupled transport-activation (15). This mechanism was first shown to occur in gram-negative bacteria and is referred to as vectorial acylation (16). Six different isoforms of FATP have subsequently been identified experimentally in mice, rats, and humans (e.g., in mice, mmFATP1-6).

The mouse FATP1 structural gene contains a peroxisome proliferator-activated receptor response element (PPRE) from -456 to -474; FATP1 expression is induced fourfold and 5.5-fold, respectively, by PPARα and PPARγ in the presence of their respective activators in a PPRE-dependent manner (16). FATP1 expression is down regulated by insulin in cultured 3T3-L1 adipocytes and up regulated by nutrient depletion in murine adipose tissue (17). An insulin response sequence has been identified in the mouse FATP1 structural gene, which is also found in the regulatory region of other genes negatively regulated by insulin, including those encoding phosphoenolpyruvate carboxykinase, tyrosine aminotransferase, and insulin-like growth factor-binding protein 1 (16).

Knockout mice have been generated for FATP1 (18), FATP2 (19), FATP4 (20), and FATP5 (21), which have provided some insights into the roles of these proteins in fatty acid trafficking. Deletion of FATP1 protects animals from developing insulin resistance on a high fat diet and during chronic lipid infusion (18). Animals with a deletion in the gene encoding FATP2 were only evaluated for changes in very long chain acyl-CoA synthetase activity, which was partially diminished, and for accumulation of very long chain fatty acids, which were unchanged. The animals have not been evaluated for fatty acid uptake and no other phenotypes were noted (19). Deletion of FATP4 results in neonatal lethality (20), which is believed to be due to a restrictive dermopathy that prevents expansion of the diaphragm upon birth. Deletion of FATP5 in the mouse results in aberrant bile acid conjugation, which in some manner is linked to the regulation of body weight (21).

Bioactive fatty acids and cellular signaling

Work describing the transcriptional control of fatty acid metabolism in E. coli set the stage showing LC-CoA is an important regulatory molecule that influences a number of different metabolic processes (Fig. 2a).

A noted above, fatty acids also serve as important bioactive fatty acids, which impact on a number of cellular processes related to and distinct from lipid homeostasis (22). In mammalian systems, the highly unsaturated fatty acids derived from LA and LN yield a range of compounds with unique roles in a variety of cellular metabolic and signaling events. AA and EPA-derived eicosanoids mediate pleiotropic responses including, for example, pain, vasoconstriction, vasodilation, inflammation, and bronchoconstriction. Some HUFA and eicosanoids are ligands for transcription factors, particularly members of the peroxisomal proliferator activated receptor family (PPAR) involved in fatty acid and triacylglyceride metabolism and adipocyte differentiation. Certain classes of HUFA also decrease the activity of the SREBP family members, which activate genes required for LDL uptake, cholesterol, and fatty acid synthesis (23). Numerous studies have implicated HUFA in the prevention and management of diabetes, cardiovascular disease, and metabolic syndrome, although the mechanisms are largely undefined (24). Current evidence indicates eicosanoids derived from EPA may counter some effects of eicosanoids derived from AA (22). Hence, there is much interest in establishing a balanced ratio of ω6:ω3 fatty acids, generally recommended to be 2:1 (26, 27).

Lipid-responsive transcription factors and fatty acid homeostasis

The peroxisomal proliferator activated receptors (PPAR) are considered to be internal fatty acid sensors. To date, three distinct PPARs have been discovered, α, γ, and δ, each encoded by a separate gene with distinct tissue specific distributions. PPARα is primarily found in the liver and is involved in regulating the genes involved in fatty acid catabolism whereas PPARγ is primarily involved in the adipocyte differentiation and fatty acid storage. PPARα has a wide range of tissue distribution, but its specific role in the regulation of fatty acid metabolism in the liver has not been well elucidated. The natural ligands for the PPARs are ω3 and ω6 PUFA ω6 fatty acids, HUFA s, and fatty acid-derived metabolites (e.g., leukotriene B4, and 5(S)-hydroxy-eicosatetraenoic acid (29, 31)). Recent work has shown that very long chain fatty acids, including C20:4-ω6 and C22:6-ω3 are potent activators of the PPARα/δ heterodimer and have been shown to induce target genes associated with fatty acid and cholesterol uptake and metabolism (32). PPARγ is primarily expressed in adipose and muscle tissue and is involved in the transcriptional regulation of genes involved in fatty acid and cholesterol synthesis and storage (33). PPARγ is activated by drugs such as thiazolidinediones (34), which are used to treat type II diabetes. The PPARα/γ heterodimer is also important in the transcriptional regulation of genes involved in fatty acid and cholesterol synthesis and storage.
C20:5 ω-3, C22:6 ω-3, and branched-chain fatty acids are potent activators of PPARα (31). Although there is some disagreement whether PUFA or HUFAs are more potent PPAR α agonists, HUFAs are more potent toward PPARγ (32). PPARα regulates genes involved in mitochondrial β-oxidation, including those encoding straight-chain acyl-CoA dehydrogenase (ACDH) and carnitine palmitoyl transferase (CPT1 and 2). PPARα also regulates key peroxisomal genes, including those encoding acyl-CoA oxidase (AOX), L-bifunctional protein, and 3-ketoacyl-CoA thiolester (33), involved in the β-oxidation of very-long chain fatty acids (VLCFAs), which is required in the synthesis of DHA.

Sterol regulatory element-binding proteins (SREBP1α, SREBP1c, and SREBP2) are transcription factors that control both cholesterol and fatty acid biosynthesis. SREBP-1c regulates genes involved in fatty acid synthesis and storage, whereas SREBP-2 regulates genes involved in cholesterol and bile acid synthesis. SREBP-1a is thought to regulate both fatty acid and cholesterol biosynthesis. SREBPs are regulated by intracellular oxysterol levels such that when elevated, their proteolytic cleavage and subsequent activation is prevented. SREBPs serve to balance PPAR-regulated fatty acid oxidation. PUFA and HUFA are believed to suppress fatty acid synthesis by decreasing the activities of SREBP1c and LXR (34). The active form of SREBP is negatively regulated by PPAR, PUFA, and HUFA (22, 35, 36). In addition, SREBP transcription is controlled by LXR demonstrating crosstalk between PPARs and LXR (35). More specifically, PPAR competes with LXR for their shared heterodimer partner, RXR, thereby decreasing LXR activity. The balance between fatty acid biosynthesis and breakdown is important in many disease states and all of the above-mentioned transcription factors play a role in maintaining this balance.

Chemical Tools and Techniques For Studying Lipid Metabolism

Extraction of lipids and the foundations for current analytical studies

The fundamental step in all analytical studies on lipids and lipid metabolites begins with extraction methods. Two papers, Folch et al. (37) and Bligh and Dyer (38), are routinely cited methods describing lipid extraction and should be consulted as general guides. Details of these methods can be found in the Lipid Library (http://www.lipidlibrary.co.uk/topics/extract/index.htm) and in Cyberlipids (http://www.cyberlipid.org/extract/extract0002.htm).

Thin layer and high performance liquid chromatography

Early studies employed thin layer chromatography (TLC) in the separation, identification, and quantification of individual classes of lipids. These techniques have been optimized for different classes of glycerol phospholipids, neutral lipids, sphingolipids, and fatty acids. Methods and general references to TLC can be found in the Lipid Library (http://www.lipidlibrary.co.uk/topics/tlc/index.htm) and in Cyberlipids (http://www.cyberlipid.org/fraction/fraction0005.htm). High performance liquid chromatography (HPLC) has now become routine in the separation and quantification of lipids. Given the optical properties of lipids, the use of evaporative light-scattering detectors is most optimal for detection and quantification. The most comprehensive information on the use of HPLC in analytical lipid studies can be found at Cyberlipids (http://www.cyberlipid.org/index.htm).

Gas chromatography and mass spectrometry

The identification of individual classes of fatty acids has relied on the use of gas chromatography (GC), equipped with a flame ionization detector. Lipids are saponified after extraction and the fatty acids converted to methyl esters. The fatty acid methyl esters (FAME) are separated using GC. The use of standards allowed for the identification of individual species of fatty acids based on retention time. The method is quite sensitive and permits the quantification of fatty acid species. The mode of detection has been enhanced with the use of mass spectrometry (MS), which allows for the detection and quantification of unknowns, thus increasing the utility of these methods. The Lipid Library section on the gas chromatography of lipids (http://www.lipidlibrary.co.uk/GC/lipid0003_index.html) provides a comprehensive description of these methods.

Fluorescent fatty acids using fatty acid trafficking and signaling

A number of different fluorescent fatty acids are in use to investigate various aspects of fatty acid homeostasis. Among these are the C18, C16, and C12 fatty acid-5-aminofluorescein conjugates octadecylaminofluorescein (OAF), hexadecylaminofluorescein (HAF), and dodecylaminofluorescein (DAF), which have been used to study membrane dynamics in gram-negative bacteria. Studies on the intracellular fatty acid-binding proteins have employed anthroloxy-labeled fatty acids, whereas biophysical studies addressing fatty acid transport have used the naturally occurring cis-paranaric acid. Many of the current studies addressing fatty acid trafficking use BODIPY (4,4-difluoro-5-methyl-4-bora-3a,4a-diaza-s-indacene)-labeled fatty acids, which has been adapted for high throughput technologies to select for small molecules that inhibit one or more steps in fatty acid trafficking (39). The BODIPY-derived fatty acid molecules can be visualized using fluorescence microscopy, which provides a tool to monitor intracellular trafficking.

Research Directions—2007

High resolution studies using LC-MS/MS

The identification and quantification of lipids, particularly minor species that are likely to have bioactive roles, has lagged behind...
their hydrophilic counterparts. Understanding these minor lipid compounds is essential to have a complete picture of all of the factors involved in governing lipid homeostasis. With the advent of developing analytical tools that combine HPLC with tandem mass spectrometry, we are now in a position to acquire a significant and novel body of information on complex lipids and lipid metabolites, and to determine how these compounds fit into larger metabolic schemes essential for addressing issues of health and disease.

**Lipid maps**

The emerging field of metabolomics is defined as the study of metabolites and metabolic interrelationships and encompasses the global analysis of the plethora of small molecules generated in the process of metabolism. To effectively use this information for understanding both normal physiology and pathophysiology, we must first have a firm grasp on the interrelationships between metabolic enzymes and the pathways in which they participate. The study of the metabolic intermediates of lipid metabolism and the enzymes participating in these metabolic pathways comprise the field of lipidomics (see "Lipidomics" in this series). A new resource in lipidomics is the Lipid MAPS ([LipidMaps.org](http://www.lipidmaps.org/)), which has been established to aid investigators. Lipid MAPS serves 1) to separate and detect all of the lipids in a specific cell and discover and characterize any novel lipids that may be present, 2) to quantify each of the lipid metabolites present and the changes in their levels and location during cellular function, and 3) to define the biochemical pathways for the synthesis of each lipid and develop lipid maps that define the interaction networks ([40]). This consortium is directed to finding and cataloging the myriad of lipid compounds, including those that are least abundant, those that are transient intermediates in the metabolic pathways, and those with pronounced biologic activity. The consortium is establishing a comprehensive database available to researchers addressing questions within the lipidomics field (e.g., see Reference [41]). Independent researchers and smaller research consortia of investigators with common interests in lipidomics, such as those described within this proposal, can access these databases and, as importantly, can provide information to be cataloged within Lipid MAPS.

**References**


Further Reading

See Also
Lipidomics
Lipid Bilayers, Properties of
Lipid Synthesis
Lipid Signals
Chemistry and Chemical Reactivity of Lipids
Lipoproteins, Chemistry of
Mass Spectrometry: Small Molecules
Metabolism, Cellular Organization of...
Electron Transfer Chemistry in Photosynthesis

Art van der Est, Department of Chemistry, Brock University, Ontario, Canada
doi: 10.1002/9780470048672.wecb450

In oxygenic photosynthesis, light-induced electron transfer reactions occur in two photosystems embedded in the thylakoid membrane. These complexes bind many cofactors and are among the handful of membrane proteins for which the X-ray crystal structures have been determined. The availability of this structural information, in combination with advanced spectroscopic techniques, has allowed the electron transfer reactions to be studied in detail. This work has revealed the pathways, rates, and yields of the electron transfer reactions but has also raised several intriguing questions about how these properties are governed by the protein–cofactor interactions. In this article, a brief overview of the main features of these very complex reactions is given, along with a summary of some techniques used to study them. A short survey of some recent advances in our understanding of the chemistry of electron transfer in photosynthesis is given.

The energy needed by all living organisms is derived ultimately from the sun. However, most energy contained in sunlight is dissipated as heat, which is insufficient to break and form all but the weakest chemical bonds. Hence, to use the energy of the sun, living organisms require a mechanism for capturing individual photons and storing the energy they contain. Moreover, a system that can perform multiple turnovers is required, because even the energy contained in a single photon of visible light is not sufficient for many biochemical reactions. Photosynthesis is the process by which this energy storage is achieved, and because all living organisms are dependent on it, it is arguably the most important biologic process.

Biologic Background

At the heart of this solar energy conversion process are two light-induced electron transfer reactions that generate oxidizing and reducing agents and a transmembrane proton concentration gradient. The nature of this reaction and the type of oxidant it produces distinguishes the various types of photosynthetic organisms.

Occurrence of photosynthesis in the three domains of life

Photosynthesis is most commonly associated with higher plants, which are eukaryotes. However, in terms of photosynthetic electron transfer chemistry, a wider variety of species exists among the bacteria. Indeed, eukaryotic photosynthesis is believed to have evolved through a symbiotic relationship between a non-photosynthetic eukaryote and an oxygenic photosynthetic bacterium. Hence, a complete classification of the various types of photosynthetic electron transfer apparatus can be performed within the domain of the bacteria.

Types of photosynthetic organisms

Photosynthetic bacteria can be divided into two classes: anoxygenic and oxygenic. The anoxygenic bacteria are thought generally to have evolved before the advent of oxygenic photosynthesis, and in these species, the electron transfer oxidizes electron-rich compounds such as thiols. The limited availability of such compounds strongly restricts the environments in which anoxygenic bacteria can survive. In contrast, the oxygenic organisms can live in a much wider variety of habitats because they have evolved the ability to oxidize water, which is much more abundant. Although considerable diversity exists, both at a cellular and molecular level between the various photosynthetic organisms, all of them possess photosynthetic reaction centers, which are membrane-bound oxidoreductase enzymes in which the light-induced electron transfer takes place. The reaction centers are divided generally into two classes on the basis of the terminal electron acceptors, and all known reaction centers can be placed into one of these two types. A unique feature of oxygenic organisms is that they have two different reaction centers, one of each type. Hence, the oxygenic photosynthetic bacteria provide a convenient framework within which most features of photosynthetic electron transfer can be discussed.
Components of the photosynthetic apparatus in oxygenic photosynthesis

The two different reaction centers in oxygenic species seem to have evolved so that they can generate the very positive oxidation potential needed to split water without sacrificing the reducing potential generated on the other side of the membrane. In both reaction centers, an extensive network of chlorophyll molecules, used to capture light, is housed in the same complex as the electron transfer catalysts, and hence, these complexes are called photosystems. The two photosystems are called more correctly light-driven oxidoreductases, and they are referred to as Photosystem I (PS I) and Photosystem II (PS II) or plastocyanin:ferredoxin oxidoreductase and water:plastoquinone oxidoreductase, respectively. In addition to the photosystems, the cyanin:ferredoxin oxidoreductase and water:plastoquinone oxidoreductase participate in the electron transfer in oxygenic photosynthesis. The proton gradient generated by these enzymes is driven by ATP synthase to phosphorylate ADP. These components make oxygenic photosynthesis one of the best characterized biologic pathways.

Electron Transfer in Oxygenic Photosynthesis

The electron transfer pathway shown in Fig. 1 is driven by light-induced charge separation in the two photosystems. The first of these systems is PS II, shown on the left of Fig. 1. In PS II, the radical cation generated by the charge separation extracts electrons from a manganese cluster located on the lumen side of the membrane. After four turnovers of PS II, the manganese complex can remove four electrons from two water molecules to produce molecular oxygen and four protons. On the stromal side of the complex, four turnovers of the complex reduce two molecules of plastoquinone to plastoquinol, which diffuses into the membrane. The plastoquinol is then reoxidized to plastoquinone in the cytochrome b6f complex. The electrons and protons released in this process are both transferred to the stromal side of the membrane. The electrons reduce the soluble protein plastocyanin, whereas the two protons are released into the stroma. In the second photosystem, PS I, the radical cation generated by light-induced charge separation reoxidizes reduced plastocyanin. The electrons transferred through PS I reduce ferredoxin, which is a soluble protein located in the stroma. Ferredoxin is then reoxidized by ferredoxin:NADP+ reductase (FNR), which converts NADP+ to NADPH. The oxidation of water in PS II and the conversion of plastoquinone to plastoquinol and back again lead to the accumulation of protons in the lumen, which creates a proton gradient across the thylakoid membrane. Cyclic electron transport, in which reduced ferredoxin passes electrons to the cytochrome b6f, also can contribute to establishing a difference in H+ concentration across the membrane. This proton gradient is used to drive the ATP synthase enzyme shown on the right, which converts ADP to ATP on the stromal side of the membrane. The final step of photosynthesis, the production of carbohydrates from CO2 via carbon fixation, derives its energy from ATP and NADPH. This step is not shown in Fig. 1.

The Z-scheme

The elucidation of the overall mechanism of photosynthesis described above is the product of many years of intense research. Initially, the focus of most of this work was to determine the nature of the light-induced reactions and their energetics. In 1960,
Hill and Bendall (10) presented a hypothesis that suggested that two light-induced steps were involved and showed a diagram in which the energy was plotted along the horizontal axis. Because the two reactions in this diagram resembled the letter “Z,” it became known as the Z-scheme. Now it is drawn according to the more usual convention of energy as a vertical axis, and many details of the electron transfer reactions have been added (11). This scheme is shown in Figure 2. The vertical axis in the figure indicates the midpoint potential for oxidation or reduction for the various species. The two vertical steps indicate the absorbance of light by PS I and PS II, and it is apparent that the midpoint potentials of the two photosystems are shifted by ~0.8 V with respect to one another. Because of this shift, PS II can use 680 nm photons to generate an oxidizing potential of ~1.0 V, while PS I uses photons of similar energy (700 nm) to produce a reducing potential of ~0.7 V. The various steps after light absorption by PS I and PS II refer to electron transfer along the chain of acceptors in the two systems. Here, we will focus on the details of these electron transfer reactions. Electron transfer in the cytochrome b6f complex is a multistep process also, but these steps are not shown in Figure 2 and readers are referred to a recent review (12) for details.

Photosystem II and type II reaction centers

PS II is representative of the Type II class of reaction centers, and in these systems, the primary acceptor is pheophytin (shown in blue and labeled Pheo in Figure 3, left). A problem that has intrigued researchers in the field for many years is the fact that despite the apparent symmetry of the complex, the electron transfer occurs exclusively to Pheo, on the left in Figure 3 (13). The exact nature of this initial charge separation remains a subject of debate. In purple bacteria, the chlorophyll dimer acts as the donor and the electron transfer is known to proceed via the accessory chlorophyll Chl+$\text{ }^\ddag$ (shown in light blue in Figure 3 (14)). However, in PS II, recent evidence suggests that the initial charge may occur between ChlA and Pheo, and that the oxidized Chl$^+$ then is reduced rapidly by the special pair, P$_{680}^\ddag$ (15). After this initial charge separation, the electron is transferred to the neighboring quinone Q$_{A}$ and then across the complex to quinone Q$_{B}$ (both shown in red in Figure 3). The nature of the quinone varies slightly between various organisms. In PS II, it is plastoquinone, but other quinones, such as ubiquinone, are found in anoxygenic organisms with Type II reaction centers.

In the latter organisms, the oxidized donor is re-reduced by a cytochrome whereas in PSII and P$_{680}^\ddag$ is reduced by the neighboring tyrosine $Y_{Z}$, which in turn is reduced by the oxygen-evolving complex (OEC) (represented by the four Mn atoms shown in green in Figure 3). A second turnover of the enzyme removes a second electron from the OEC and leads to double reduction and protonation of Q$_B$ to its quinol form. The quinol does not have a high affinity for the Q$_B$ site and diffuses out of the reaction center to be replaced by a quinone. The complex can then undergo two more turnovers that produce another quinol and extract an additional two electrons from the OEC. The OEC then has sufficient oxidizing potential to extract four electrons from two water molecules and release four protons and an oxygen molecule in the process.

Figure 2 shows the lifetimes for the various electron transfer steps along with the midpoint potentials for the acceptors, as indicated by the vertical axis. The charge separation to Q$_{A}$ ...
A rapid transformation into S0 accompanied by the release of O2 than maximal energy efficiency. As can be observed from the moving reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less a quantum yield of close to unity. To achieve this extremely high quantum yield, the rates of forward electron transfer have been optimized, thus avoiding nonproductive recombination. The further the electron is from the donor, the slower the back reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less.

The further the electron is from the donor, the slower the back reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less.

The energetics of electron transfer in reaction centers and the factors governing the rates are discussed in Reference 17. One of the most interesting and intensively researched aspect of the electron transfer chemistry in PS II is the function of the OEC. This function is also the most poorly understood part of the PS II enzyme because of its complexity and the difficulty in obtaining spectroscopic signatures that can be interpreted unambiguously in terms of a mechanism. Here, only a basic outline of our current understanding of the water-splitting process is given, and readers are referred to several recent and excellent reviews for more detail (18–21). As successive electrons are removed from the OEC, it passes through a series of five states labeled S0 through S4. The S0 state is not stable and rapidly transforms into S1 accompanied by the release of an O2 molecule. This sequence of steps is known as the Kok cycle, named after Bessel Kok (22) who first observed that oxygen was evolved after every fourth flash of light given to dark-adapted PS II samples. As successive electrons are removed from the OEC, it passes through a series of five states labeled S0 through S4. The S0 state is not stable and rapidly transforms into S1 accompanied by the release of an O2 molecule. This sequence of steps is known as the Kok cycle, named after Bessel Kok (22) who first observed that oxygen was evolved after every fourth flash of light given to dark-adapted PS II samples.

The geometric arrangement of the redox active cofactors in Photosystem I and Photosystem II as given by the X-ray crystal structures: PS I pdb entry 1B80 (2) and PS II pdb entry 2AK7 (3). One of the most interesting and intensively researched aspect of the electron transfer chemistry in PS II is the function of the OEC. This function is also the most poorly understood part of the PS II enzyme because of its complexity and the difficulty in obtaining spectroscopic signatures that can be interpreted unambiguously in terms of a mechanism. Here, only a basic outline of our current understanding of the water-splitting process is given, and readers are referred to several recent and excellent reviews for more detail (18–21). As successive electrons are removed from the OEC, it passes through a series of five states labeled S0 through S4. The S0 state is not stable and rapidly transforms into S1 accompanied by the release of an O2 molecule. This sequence of steps is known as the Kok cycle, named after Bessel Kok (22) who first observed that oxygen was evolved after every fourth flash of light given to dark-adapted PS II samples. As successive electrons are removed from the OEC, it passes through a series of five states labeled S0 through S4. The S0 state is not stable and rapidly transforms into S1 accompanied by the release of an O2 molecule. This sequence of steps is known as the Kok cycle, named after Bessel Kok (22) who first observed that oxygen was evolved after every fourth flash of light given to dark-adapted PS II samples.

as is apparent in Fig. 2, the distance between the primary acceptor (A0A,B) and PheoA takes place with an overall lifetime of ~200 ps and a quantum yield of close to unity. To achieve this extremely high quantum yield, the rates of forward electron transfer have been optimized, thus avoiding nonproductive recombination. The further the electron is from the donor, the slower the back reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less.

As is apparent in Fig. 2, the distance between the primary acceptor (A0A,B) and PheoA takes place with an overall lifetime of ~200 ps and a quantum yield of close to unity. To achieve this extremely high quantum yield, the rates of forward electron transfer have been optimized, thus avoiding nonproductive recombination. The further the electron is from the donor, the slower the back reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less.

As is apparent in Fig. 2, the distance between the primary acceptor (A0A,B) and PheoA takes place with an overall lifetime of ~200 ps and a quantum yield of close to unity. To achieve this extremely high quantum yield, the rates of forward electron transfer have been optimized, thus avoiding nonproductive recombination. The further the electron is from the donor, the slower the back reaction becomes, and the forward reaction rate, therefore, can be slower. The price for the high quantum efficiency is less.
primary role of PS I is to produce a strong reducing potential. From this point of view, it is surprising that PS I contains two phytoquinone molecules on its acceptor side, because quinones generally behave as oxidants and semiquinones behave as weak reductants. Thus, in PS I, the environment clearly has a strong impact on the redox properties of the acceptors. A comparison of the electron transfer rates given in Fig. 2 also shows that the electron transfer times generally are faster than in PS II. Presumably, the reason for this is because no need exists for multiple reduction of the acceptors. Therefore, the quinone can be placed closer to the primary acceptor and can pass the electron to $F_X$ faster than $Q_A^-$. As in all reaction centers, the initial charge separation is extremely fast and it has been assumed generally that it occurs between the chlorophyll dimer ($P_{700}$) and $A_0$. However, as in PS II, this assumption is being challenged by recent data, which suggest that the initial charge separation may occur between the accessory chlorophyll(s) and $A_0$, followed by rapid donation from $P_{700}$ (16).

An unusual feature of the electron transfer in PS I is the fact that it is heterogeneous. Many steps, for example, from $A_1$ to $FX$ and the absorbance is measured using a probe pulse at a known delay after the pump pulse. The time resolution of this method can be as short as $10^{-13}$ seconds. Transient fluorescence methods involve detecting emitted photons and correlating the time at which the photons arrive at the detector with the time of the excitation pulse. The ultrafast techniques are usually limited to times less than a few nanoseconds. To study electron transfer steps with lifetimes longer than this, a continuous detection beam and a fast digitizer can be used to follow the absorbance changes or the emitted light. These methods have provided most lifetimes given in Fig. 2.

Electron paramagnetic resonance methods

The electron transfer in reaction centers generates a series of radical pairs that can be detected by electron paramagnetic resonance (EPR) spectroscopy (see Reference 27 for a review). The advantage of this method is that only paramagnetic species are detected. Hence there are fewer background signals and less chance of errors when subtracting them. However, the method is limited to a time resolution of $\sim 10^{-9}$ seconds. Therefore, it only can be used to study the secondary electron transfer steps. A crucial advantage of magnetic resonance techniques in general is that they depend on tensorial properties and therefore give information such as the relative orientation of the radicals and their spin density distributions along with the rates of electron transfer. For photosynthetic systems, EPR has been used most widely to study trapped paramagnetic intermediates generated by the electron transfer at low temperature. In such experiments, all kinetic information is lost; however, the properties of the individual radicals can be studied in detail, and the pathway can be deduced from the species observed.

Recent Advances

Directionality of electron transfer in photosystem I

One peculiarity of PS I is that the two branches of electron cofactors converge at $F_X$, so that a priori reason exists for why electron transfer should use only one of them. On the other hand, no obvious need exists for both branches to be active. This problem has been addressed by a number of researchers in recent years by using the known structure of PS I to identify specific amino acid residues close to the cofactors in one branch or the other and making point mutations at these locations. Although no consistent picture of the electron transfer pathway in PS I has been developed yet, much of the data provide evidence that both branches may be active. If this model is correct, some data also suggest that it may be possible to influence the extent to which electron transfer occurs in a given branch (24).

Water oxidation and the structure of the water-splitting complex

As discussed, the holy grail of much research on PS II has been to elucidate the mechanism for water splitting. A major hurdle...
References


Electron Transfer Chemistry in Photosynthesis


Further Reading


See Also

ATP Synthesis/ F1Fo Synthase
Fluorescence Spectroscopy: Overview of Applications in Chemical Biology
Iron-Sulfur World
Porphyrin and Corrin Biosynthesis
Metabolomics: Topics in Chemical Biology

Julian L. Griffin, Helen J. Atherton and Michael R. Pears, Department of Biochemistry, Tennis Court Road, Cambridge, United Kingdom
doi: 10.1002/9780470048672.wecb655

With the development of systems biology, several approaches have been developed to profile a tier of organization in a cell, tissue, or organism globally. Metabolomics, also referred to as metabonomics, is an approach that attempts to profile all the small-molecule metabolites in a biological matrix. One major challenge of this approach, as with other “-omic” technologies, is that the metabolome is context dependent and varies with developmental stage/age, genetic modification, disease, and environment. Thus, by definition, the metabolome of an organism must take into consideration all the effects of these manipulations. Despite these challenges, the approach has already been applied to understand metabolism in a range of animal models and has more recently started to be applied to clinical studies. In this article, we will discuss some common approaches currently used in metabolomics and the results that they have produced. In particular, we will focus on the two most common analytical approaches, nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry, and some biological problems they have been used to address.

The postgenomic era seeks to define the biological function of genes and how these translate to define an organism’s phenotype. Unlike genome sequencing projects, the effective study of functional genomics necessitates a plethora of multidisciplinary techniques (1). The development of analytical tools that enable the high-throughput measurement of gene products has provided a comprehensive signature of the physiological state of the cell at several levels (Fig. 1). Metabolomics describes the large-scale analysis of endogenous metabolites that comprise the entire collection of small molecules in a cell, tissue, organism, or biofluid, and includes sugars, organic acids, amino acids, and nucleotides (2, 3).

Benefits of a Metabolomic Approach

A metabolomic approach has several advantages for use in functional genomics. Firstly, as with transcriptional and proteomic analyses, the technique is context-dependent, such that the metabolite complement varies according to the physiological, developmental, or pathological state of the cell, tissue, organ, or organism, which makes it a highly potent tool for measuring changes in phenotype. Secondly, the importance of metabolites in biological control and communication, as building blocks for complex macromolecules and energy transporters, makes them effective markers of cellular function. Indeed, metabolites constitute molecular endpoints farther down the line from gene to function. Thirdly, metabolic control analysis suggests that, although changes in the quantities of individual enzymes are expected to exert little influence on metabolic fluxes, they can and do have a significant impact on metabolite concentrations, even when changes in flux are negligible. Fourthly, metabolites span the species barrier because they have the same chemical structure irrespective of the organism. Metabolomics is therefore a universal “omic” technology in contrast to transcriptomics and proteomics in which a priori knowledge of DNA and protein sequences from each organism is required. Little time is required for reoptimizing protocols for a new species. Finally, metabolomics presents several practical benefits, including cheap cost on a per sample basis, high throughput, and full automation. For example, after the initial purchase of a nuclear magnetic resonance (NMR) spectrometer or mass spectrometer (MS), samples can be analyzed at a cost in the region of £~1.00 per sample, in terms of consumables, with analytical acquisition times typically taking 10 min (NMR) to 70 min (GC–MS). This cost compares very favorably with transcriptional and proteomic analyses.
Terminology

The terms metabolomics (2) and the related term metabonomics (3) were coined in the late 1990s to describe the use of global profiling tools combined with pattern recognition to define a metabolic phenotype of a cell, tissue, or organism. The confusion over what to call the field is still a persistent problem, and new researchers to the field should be aware of both terms. Many people use the terms metabolomics and metabonomics interchangeably (4, 5). However, Nicholson and coworkers state that metabonomics is the “quantitative measurement of the time-related multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification,” whereas metabolomics is the “measurement of metabolite concentrations and fluxes and secretion in cells and tissues in which there is a direct connection between the genetic activity, protein activity and the metabolic activity itself” (6). To add more confusion, additional terms have been put forward dependent on the analytical techniques adopted. "Metabolic profiling" has been proposed to refer to the detailed analysis of metabolites by hyphenated techniques such as gas chromatography-(GC)-MS and liquid chromatography-(LC)-MS. In contrast, "metabolic fingerprinting" is thought to measure a subset of metabolites and uses techniques such as NMR and direct infusion electrospray-MS to create a barcode of metabolism (7). The term “metabolomics” will be employed in this article throughout.

Analytical techniques

Metabolomics is challenged by the remarkable heterogeneity and complexity exhibited by metabolites. Metabolites span concentration ranges of the order $10^9$, polarity ranges of $\sim 10^{20}$, and mass ranges of the order of 1500 amu. Moreover, most extraction procedures do not provide complete representations, which results in only a portion of the metabolome being analyzed. A truly comprehensive coverage of the metabolome, therefore, necessitates multiple techniques and different sample preparation strategies to encompass such diversity. The principal analytical platforms for metabolomics include 1H NMR spectroscopy and
Figure 2. The two most frequently used approaches for metabolomics are NMR spectroscopy and mass spectrometry. To maximize the coverage of the metabolome, these techniques can be combined. This figure shows the analysis of the aqueous fraction of a section of heart tissue that uses NMR spectroscopy, gas chromatography mass spectrometry and liquid chromatography mass spectrometry. Although mass spectrometry approaches are inherently more sensitive than NMR spectroscopy, metabolite identification becomes more difficult.

GC- and LC-MS (Fig. 2), although others have also used other techniques including Fourier transform infrared spectroscopy, thin layer chromatography, high-pressure liquid chromatography, and Raman spectroscopy. The major techniques as judged by current number of publications will be discussed below.

Nuclear magnetic resonance spectroscopy

High-resolution $^1$H nuclear magnetic resonance (NMR) spectroscopy is a relatively rapid technique and is highly robust in terms of reproducibility of results. The ubiquity of protons in cellular metabolites and the fact that other nuclei are observable by NMR (e.g., $^{2}$H and $^{13}$C) mean that a relatively large number of different metabolites can be detected. Furthermore, the technique requires minimal sample preparation, and its nondestructive nature allows for more analyses to be conducted. However, $^1$H NMR is an inherently insensitive analytical tool that only measures high-concentration metabolites. Typically, using a simple one-dimensional pulse sequence, $^1$H NMR can measure 30–100 metabolites in urine, 20–30 metabolites in blood plasma, and 10–30 metabolites in tissue extracts (8). Despite this, however, $^1$H NMR has proved highly discriminatory for liver toxins in rats (9, 10), mouse models of cardiac and neurological diseases (11, 12), and silent phenotypes in yeast (13). It has been suggested that this finding may be because of the high-concentration metabolites that represent central hubs of metabolism, whereby a perturbation at one point in a metabolic network would be transferred to other pathways through these highly connected hubs (14, 15). However, restriction the coverage of the metabolome to 30 or so metabolites may hinder the isolation of metabolites as unique biomarkers for disease processes and confound the deduction of which pathways are perturbed during a given modification. Strategies are being...
developed to make NMR a more sensitive approach, including the use of cryoprobes and hyphenated techniques. Cryoprobes use liquid helium to cool down the NMR coil assembly and preamplifier to ~4 K, which, by reducing thermal noise, results in a three- to four-fold increase in the signal-to-noise ratio. Hyphenated techniques involve first separating out high and low concentration metabolites using liquid chromatography. This separation improves sensitivity by reducing the likelihood of corsonant peaks and improves the dynamic range of the NMR experiment (16).

Other NMR-based techniques allow the quantification of metabolite concentrations in intact tissue, either in vivo or ex vivo. For example, magnetic resonance spectroscopy (MRS) allows the noninvasive interrogation of metabolite concentrations directly in vivo within a specific localized region. The technique has proved invaluable in the investigation of Alzheimer’s disease (17, 18), traumatic brain injury (19), multiple sclerosis (20), and many brain tumors (21). However, one disadvantage of MRS is that it is compromised by low spectral resolution, partly as a result of the lower fields used in vivo but also founded by dipolar couplings, chemical shift anisotropy, and bulk magnetic field inhomogeneity effects, which act to broaden spectral resonances. These effects can be reduced dramatically by spinning the sample at the “magic angle” (54.7°).

In vivo spectroscopy, it is possible to produce high-resolution spectra in rat heart mitochondria that were not NMR-visible (22). Demonstrated subcompartments of acetoacetate and glutamine of investigating metabolic compartmentation. A recent study compared metabolites with those in commercially available mass spectral databases (e.g., http://www.nist.gov). Although this approach generally works well, distinction between sugar diastereomers is conditioned by both species having identical fragmentation patterns. In such cases, retention times of standards are imperative for accurate assignments. Furthermore, spectral databases are not exhaustive, and so not all peaks can be assigned. Currently, efforts are being made to create metabolomic-specific mass spectral libraries (26). More detailed characterization of unknown peaks can be conducted by running the same samples using chemical ionization as opposed to electron impact ionization or using ion trap mass spectrometers. Chemical ionization is a less energetic method of ionization and so produces less fragmentation. This method increases the probability of observing the molecular ion and so can facilitate compound identification.

The sensitivity of the GC-MS approach is impressive. Using a simple single quadrupole GC-MS, Frieh and coworkers working on Arabidopsis thaliana reported the detection of 326 metabolites, a number that has since increased to over 1000 using a Time of Flight GC-MS (25, 27). Furthermore, the technique has benefited from recent technological advances in two-dimensional GC (28).

LC-MS provides metabolite separation by LC before detection by MS. Sample derivatization is generally not required because no prerequisite for volatility is necessary. Moreover, LC-MS can analyze a wider array of metabolites relative to GC-MS because it is not limited to volatile compounds or molecules that can be rendered volatile. In particular, thermolabile or large molecules such as di- and triphosphates, CoA adducts, peptides, and lipids are all amenable to analysis by LC-MS (29). Following separation by LC, the compounds are ionized, typically using electrospray ionization (26). The sample is passed through a thin metallic capillary tube that contains positively or negatively charged ions, dependent on the solvent used. Electrospraying of the sample creates a cloud of charged droplets, and desolvation of these droplets results in ions that are sampled by the MS. Because electrospray ionization causes minimal fragmentation of the molecular ion, direct metabolite identification by comparison of mass spectra is often not possible. Structural information can be gained, however, by using tandem MS technologies, whereby sequential and selective fragmentation can be carried out. Moreover, compounds that contain a particular functional moiety can be analyzed by conducting neutral loss experiments.

Mass spectrometry

Mass spectrometry (MS) allows the analysis of lower-concent ration metabolites compared with 1H NMR spectroscopy. However, a comprehensive metabolic analysis by MS often requires sample prefractionation, the most common being gas chromatography (GC) and liquid chromatography (LC). Although this prefractionation improves the resolution of the technique and reduces ion suppression, adding a chromatographic step can introduce variability into a data set. In GC-MS, volatile and thermally stable compounds are first separated by GC and the eluted compounds detected by MS. To maximize the range of compounds that are volatile and thermally stable, samples are usually derivatized before analysis for metabolomic experiments. Methoximation in combination with trimethylsilylation is the most predominant protocol for examining aqueously soluble metabolites (25). Methoximation reduces the number of isomers present for sugars, with the number of isomeric forms being reduced from 5 to 2, which thereby dramatically simplifies chromatograms. Trimethylsilylation replaces exchangeable protons with silyl groups to make compounds less polar and therefore more volatile and to improve chromatography. The derivatized samples are vaporized within an inlet port and injected onto the column where they are carried by an inert gas. Compounds are separated according to their partition between the mobile gas phase and the column-bound stationary phase, and the eluted compounds are subsequently ionized within the ion source. The most frequently used ionization method is electron impact in which a beam of electrons ionizes the sample molecules, which concomitantly fragment. Because the energy of these electrons is preselected, the fragmentation pattern is reproducible and effectively provides a near-unique metabolite fingerprint. Metabolites, therefore, can be identified easily by comparing spectra with those in commercially available mass spectral databases (e.g., http://www.nist.gov). Although this approach generally works well, distinction between sugar diastereomers is conditioned by both species having identical fragmentation patterns. In such cases, retention times of standards are imperative for accurate assignments. Furthermore, spectral databases are not exhaustive, and so not all peaks can be assigned. Currently, efforts are being made to create metabolomic-specific mass spectral libraries (26). More detailed characterization of unknown peaks can be conducted by running the same samples using chemical ionization as opposed to electron impact ionization or using ion trap mass spectrometers. Chemical ionization is a less energetic method of ionization and so produces less fragmentation. This method increases the probability of observing the molecular ion and so can facilitate compound identification.

The sensitivity of the GC-MS approach is impressive. Using a simple single quadrupole GC-MS, Frieh and coworkers working on Arabidopsis thaliana reported the detection of 326 metabolites, a number that has since increased to over 1000 using a Time of Flight GC-MS (25, 27). Furthermore, the technique has benefited from recent technological advances in two-dimensional GC (28).

LC-MS provides metabolite separation by LC before detection by MS. Sample derivatization is generally not required because no prerequisite for volatility is necessary. Moreover, LC-MS can analyze a wider array of metabolites relative to GC-MS because it is not limited to volatile compounds or molecules that can be rendered volatile. In particular, thermolabile or large molecules such as di- and triphosphates, CoA adducts, peptides, and lipids are all amenable to analysis by LC-MS (29). Following separation by LC, the compounds are ionized, typically using electrospray ionization (26). The sample is passed through a thin metallic capillary tube that contains positively or negatively charged ions, dependent on the solvent used. Electrospraying of the sample then creates a cloud of charged droplets, and desolvation of these droplets results in ions that are sampled by the MS. Because electrospray ionization causes minimal fragmentation of the molecular ion, direct metabolite identification by comparison of mass spectra is often not possible. Structural information can be gained, however, by using tandem MS technologies, whereby sequential and selective fragmentations can be carried out. Moreover, compounds that contain a particular functional moiety can be analyzed by conducting neutral loss experiments.
This article, but we refer the reader to several excellent reviews of triglyceride in the heart. Similar modulations of cardiolipin in the myocardium following streptozotocin-induced diabetes proceeding accumulation of proatherogenic lysophosphatidylcholines and decreases in the antioxidant ether phospholipids, which suggests that in these obese individuals the lipid profile was already such that it would promote atherosclerosis, inflammation, and insulin resistance. LC-MS has also been used to follow changes in the aqueous fraction of tissue extracts (31), blood plasma, and urine (30) from rodent models of type II diabetes, but here the largest challenge is identifying the metabolites detected, and often such data has been largely used to classify samples as part of a metabolic fingerprinting study rather than to carry out biomarker identifi-

Relative to GC-MS, the application of LC-MS within the metabolomics field is still at a preliminary stage. Reproducibility is a major concern, and true quantification can be hindered by ion suppression effects whereby one co-eluting metabolite affects the ionization of another (26). The lack of electrospray ionization mass spectral libraries also makes identification by LC-MS a particularly challenging problem. Nevertheless, the technique is developing fast and has benefited from several tech-

nological advancements such as the acquisition of accurate mass data by the use of ion cyclotron resonance Fourier transform MS (26).

It is possible to bypass the chromatography stage and use direct infusion of lipid extracts into the mass spectrometer as part of an approach referred to as shotgun lipidomics. Han and colleagues (32) using this approach demonstrated a profound decrease in cardiolipin in the myocardium follow-

ing streptozotocin-induced diabetes proceeding accumulation of triglyceride in the heart. Similar modulations of cardiolipin metabolism were observed in the ob/ob mouse. Note that an integral part of the metabolomic approach is the application of pattern recognition techniques to deduce what variation in a data set is associated with a given disease, genetic modification, or other manipulation of the system. Because of space limitations, it is not possible to discuss this area as part of this article, but we refer the reader to several excellent reviews (33, 34).

Applications of Metabolomics

Metabolomics has proved extremely versatile with a diversity of applications including the study of gene function, toxicology, plant metabolism, environmental analysis, clinical diagnostics, investigation of disease, and discrimination of organism geno-

types. It is not possible to create a definitive list given the huge number and range of approaches that have now been published, but we have set out some key publications below to demonstrate the potential uses of metabolomics.

Study of gene function

One of the first major successes of metabolomics has been the definition of silent phenotypes in yeast where conventional growth rate analyses failed to distinguish different yeast mu-

tants (12). By using 1H NMR spectroscopy, it was apparent that “silent” yeast mutants have compensatory changes in intracellular metabolites to allow the cells to grow at a normal rate. Mutants deleted for genes of related biological activ-

ity resulted in similar metabolite perturbations and, hence, metabolic profiles, as did strains deleted for genes related to mitochondrial metabolism (12). The authors coined the term FANCY for functional analysis of coreponses in yeast (12) to describe this comparative metabolomics approach. Currently, FANCY is being used to predict gene function. Metabolic pro-

files of strains deleted for known genes are being compared with those deleted for unknown genes in the hope of assigning biological roles.

Plant metabolism

Some of the most significant developments in metabolic profi-

ling, particularly those involving GC-MS, have been made in the plant sciences discipline (35) (Fig. 3). Plants are thought to be exceptionally metabolite-rich relative to mammals and yeast. This richness stems not only from the size of their genomes, which ranges from 20,000 to 50,000 genes, but also from multiple substrate specificities for many enzymes, subcellular compart-

mentation, and nonenzymatic reactions. Currently, 50,000 different compounds have been elucidated in plants (36), and the final number is set to increase to approximately 200,000 (28). Concomitantly, the tools and analytical techniques used in metabolomics have been advanced rapidly by researchers within this field. One major goal of plant metabolomics is to expand the information about how plant biochemistry is composed and controlled. Metabolic profiles have been conducted on a diverse array of plant species, including Arabidopsis thaliana (25), tomato (37), potato (38), and rice (39). Efforts are being focused on characterizing silent plant phenotypes and trying to elucidate biological rules for genes of unknown function. For example, a modified potato plant line suppressed in expres-

sion of sucrose synthase isoform II displayed no overt pheno-
type in terms of morphology, yield, or growth rate compared with parental lines. However, by using GC-time-of-flight MS to analyze the metabolic profiles, approximately 1000 com-

pounds were quantified, and plant varieties were successfully discriminated (38). Similarly, other researchers have integrated transcriptomic and metabolomic analyses to investigate global responses to nutritional stress in Arabidopsis thaliana and have demonstrated that the genes and metabolites involved in glu-

cosinolate metabolism are coordinately regulated in response to either sulphur or nitrogen deficiency (40). Metabolomics has also been employed in characterizing the regulatory synthesis of novel plant products: for example, modified health-related carotenoid and flavonoid-based antioxidants have been isolated by the Metabolic Profiling Laboratory and have been commercialized.

Metabolomics: Topics in Chemical Biology
in transgenic tomatoes (41). Plant–host interactions are also a popular target for metabolomic studies (42, 43).

Environmental science

Metabolomics has made remarkable inroads into the environmental research community. Here, a major emphasis is to understand the impact that environmental stress, such as pollution and climate change, has on wildlife. Indeed, many government organizations monitor the prevalence of pollutants in certain species of wildlife as indicators of the exposure risk within the environment. Studies of Japanese medalia have been conducted to investigate the effects of trichloroethylene, a common environmental pollutant, and the pesticide dinoseb, on the development of fish embryos (44, 45). Similarly, cadmium toxicity has been examined in the bank vole and rat and has revealed changes in lipid metabolism that preceded classical nephrotoxicity (46, 47). A further study investigated the effects of environmental toxins on earthworms (48). In particular, the analysis of earthworm tissue extracts by 1H NMR spectroscopy identified maltose as a potential biomarker for ecotoxicity within a metal-contaminated site.

Clinical diagnostics

Metabolomics is well suited to clinical trials, and as more efficient and comprehensive analytical tools become available, a wealth of additional information can be gathered from costly clinical trials without increasing the level of discomfort to the patient or decreasing the practicality to the physician. The fact that all the technologies involved in metabolomics are relatively high-throughput means that large-scale clinical studies can be performed. Makinen and coworkers (49) examined the use of 1H NMR spectroscopy of blood serum to diagnose diabetic nephropathy in 182 sufferers of type 1 diabetes. This approach relied, in part, on modeling the different distributions of lipoprotein fractions and produced a test approximately as good as that used clinically. The group has since developed a Bayesian Markov chain Monte Carlo approach to modeling the constituents of the saturated lipid profile in blood serum of α null mice. Metabolites are identified from exact retention times and comparison of corresponding mass spectra with those in the NIST database. 97 metabolites were quantified.

Figure 3 An example of the use of GC-MS for metabolic profiling. Left: A section of the total ion chromatogram from the analysis of TMS-derivatized aqueous tissue extracts from the liver of PPAR-α null mouse. Metabolites are identified from exact retention times and comparison of corresponding mass spectra with those in the NIST database. 97 metabolites were quantified. Right: Summary of metabolite differences in the tissues of the PPAR-α null mouse. Red: increased relative to control, blue: decreased relative to control. The increased/decreased width of certain arrows reflects relative increased/decreased concentrations across these pathways, respectively.
Metabolomics is becoming a prominent phenotyping aid to the investigation of animal models of disease and discrimination of organism genotypes. Metabolomics is ideally placed as a phenotyping tool in the exploration of naturally occurring and transgenic disease models. The refinement of knockout and knockin strategies combined with accumulating sequence data has accelerated the generation of accurate disease models. Moreover, large-scale mouse mutagenesis programs have been set up, which have produced thousands of mutants in need of analysis (57). Screening for congenital malformations and biochemical, hematological, and immunological defects is extremely labor-intensive and time-consuming. Furthermore, many models do not express the same phenotype as humans and often display a milder pathology such that it is difficult to assess the effects of genetic perturbation. This situation, in part, reflects the shorter life span of model organisms such as mice when compared with humans.

Metabolomics is becoming a prominent phenotyping aid to characterize tissues and organisms rapidly within a biological context. For example, using a metabolomics-guided screen of mutant mice, a mouse model with a novel enzyme deficiency that resembles human maple syrup urine disease has been isolated (58). Mice were identified as having elevated levels of branched chain amino acids in blood and increased concentrations of branched chain u-keto acids in urine. Mere observation correlated these results with a deficiency in branched chain amino-transferase. Other researchers have used metabolomics to characterize the metabolic deficits associated with a particular disease process, for example, human metabolic syndrome. For example, the systemic effects of the PPAR**α** mutation, a gene important in regulating the feeding/fasting response, have recently been defined; using a combination of **1**H NMR and GC-MS, metabolic changes have been followed in the heart, liver, skeletal muscle, and adipose tissue of the PPAR**α** null mouse (31). Similar metabolic profiling approaches have also been conducted to investigate cardiac disease in the mdx mouse (10) and atherosclerosis in the ApoE3-Leiden mouse (59).

Metabolomic techniques have also been used widely to phenotype neurological disorders, with a diverse array of applications including the characterization of regional variation, brain tumors, and neurological disorders (60-64). Because the brain is heavily compartmentalized, a recent study used metabolic profiling to characterize distinct neuroanatomical regions in rats ex vivo by high-resolution magic angle spinning **1**H NMR (61). Such lipids are easily detectable in vivo by magnetic resonance spectroscopy and could be used to monitor the efficacy of gene therapy in patients with glioma. As a complement to this study (65), the low molecular weight intermediate composition of the same rat gliomas was quantified subsequently, and it was demonstrated that myo-inositol, glycine, and taurine concentrations correlated with tumor cell density, whereas the
Conclusions and Future Trends

Although the coinage of the term metabolomics has been relatively new, an explosion of publications has occurred in this field plus a realization that many researchers already were doing similar studies, albeit without an “-omic” tag before the word. It has not been possible to review all the applications of metabolomics fully, and the applications will increase. In addition to the development of new applications, the development of the analytical approaches will also take center stage as researchers push back the limits of detection of NMR spectroscopy, mass spectrometry, and other analytical approaches. In addition to these “wet lab” developments, both the pattern recognition approaches used to process metabolomics and the metabolomic databases used to identify metabolites need to be developed or expanded. In this respect, an excellent place to start on the arduous journey to biomarker discovery through metabolomics is the current metabolomic databases found on the web that make standard spectra freely available (68–70).

References


Further Reading


See Also

Proteomics
Transcript Profiling, Tools for
Metabolic Profiling
Genotyping
Functional Genomics
The Pentose Phosphate Pathway: An Overview

John F. Williams, Research School of Chemistry, Institute of Advanced Studies, The Australian National University, Canberra, Australia
doi: 10.1002/9780470048672.wecb433

The elementary textbook treatment of the pentose phosphate pathway (PPP) describes its nature and occurrence in the cytoplasm of most cells. The reaction scheme is of two segments, oxidative and nonoxidative. The reactions of the oxidative segment are few and involve the decarboxylation of glucose 6-phosphate (Glc 6-P) to ribulose 5-phosphate (Ru 5-P) and CO₂ via 6-phosphogluconolactone and 6-phosphogluconate (6-PG). Concomitant production of two moles of NADPH and H⁺ for each mole of Glc 6-P converted to Ru 5-P. The nonoxidative reactions are depicted classically with an ordered reaction sequence of reversible steps for the interconversion of other pentose phosphate products that originate from Ru 5-P by epimerase and isomerase enzymes. The reactions of the PPP also link the formation of sugar phosphate intermediates (glycolyl units) containing 3–7 carbon atoms that are generated by freely reversible equilibrium reactions catalyzed in fixed order by Transketolase and Transaldolase. These reactions permit the following three biosynthetic PPP functions to emerge: cellular energetics, growth, and repair. Specifically via 1) the contribution by the oxidative segment of a high NADPH / NADP⁺ redox potential that provides electrons for most reductive anabolic processes; 2) the formation of ribose 5-phosphate (Rib 5-P) for all nucleotide and nucleic acid biosynthesis, which is a demand that does not usually exceed 2% of all Glc metabolized by PPP activity; and (c) a storage pool of diverse phosphorylated glycolyl units that may be used directly or as signals for biosynthetic and energy-yielding reactions by other pathways. Finally, selected reactions of the nonoxidative segment are also part of the most extensive synthetic and life-sustaining event on the planet, namely the photosynthetic reductive path of CO₂ assimilation in all C-3 plants.

The Reactions and Significance

The Reactions and Significance of the Oxidative Segment

The first step in the discovery of the PPP commenced in the autumn of 1929 when Otto Warburg, a German biochemist and Nobel Laureate, was the dominant figure in the unravelling of the chemistry of the PPP (Fig. 1) reactions. Bernard L. Horecker, an American enzymologist, who had unjustifiably claim much of that distinction for the scheme (2).

The Reactions and Significance

The Reactions and Significance of the Oxidative Segment

The first step in the discovery of the PPP commenced in the autumn of 1929 when Otto Warburg, a German biochemist and Nobel Laureate, was the dominant figure in the unravelling of the chemistry of the PPP (Fig. 1) reactions. Bernard L. Horecker, an American enzymologist, who had unjustifiably claim much of that distinction for the scheme (2).
The Pentose Phosphate Pathway: An Overview

The Oxidative Segment of the Pentose Pathway


to witness a demonstration of some remarkable and puzzling results that Barron and Harrop had published recently (3). The demonstration involved a simple system, namely, the reaction of glucose with non-nucleated erythrocytes in the presence and absence of methylene blue (MeB). Methylene blue is one of a large number of organic dyes that form electromotively active oxidation–reduction systems. MeB had been used successfully by Torsten Thunberg, during the previous decade, for the identification of many dehydrogenase reactions. However, considerable diplomacy was needed for a peaceful presentation of the Barron reaction system because Warburg had been publicly scathing in condemnation of biochemical reactions that included MeB. Notwithstanding this prejudice, when Warburg observed the outcome of the Barron experiments, he was truly astonished. In 1929, it was accepted that all glucose metabolism in erythrocytes was by glycolysis with the formation of two equivalents of lactate, which were diffused by the cells. Contrary to this expectation, the Barron data with MeB as a reactant showed a 50% suppression of the extent of glycolysis and lactate formation while the rate and extent of glucose utilisation was maintained and accompanied by a rapid oxygen uptake. Florink (4) wrote, "In Baltimore Warburg clearly saw, for the first time in his experience, a dehydrogenase at work even though methylene blue was involved."

Warburg returned to Berlin where, with his colleague W. Christian, he repeated and extended the Barron work and over the next 18 months showed that the phenomenon, which was first witnessed in Baltimore, comprised the following sequence of reactions (reactions (1)–(5)).

D-glucose 6-P + H2O + 2 NADP+ → CO2 + pentose-5-P + 2 NADPH + 2 H+ (1)

2 NADPH + 2 H+ + 2 FMN → 2 NADP+ + 2 FMNH2 (Rapid reaction) (2)

2 FMNH2 + O2 → 2 FMN + 2 H2O (Slow reaction) (3)

2 FMNH3 + 2 MeB → 2 FMN + 2 MeB H+ + 2 H+ (Fast reaction) (4)

2 MeB H+ + 2 H+ + O2 → 2 MeB + 2 H2O (Very fast reaction) (5)

Reactions (1)–(5) encompass the disclosure of 1) the reactions of the oxidative segment of the PPP, 2) the discovery of a new pyridine nucleotide-NADP+ and 3) the detection of a flavin mononucleotide (FMN) flavoprotein (initially called "Old yellow enzyme") and the recognition of its role as a catalytic carrier of reducing equivalents from reduced pyridine nucleotide to molecular oxygen via MeB. No contradiction can exist that the unravelling of the reactions of reactions (1)–(5) marked a historic moment of high achievement in biochemical research. The achievement was to lead Warburg and Christian to discover and characterize glucose 6-phosphate dehydrogenase (Zwischenferment) (Glc 6-PDH) (EC 1.1.1.49) and NADP+ (Wasserstoffübertragenden).

In 1931, Warburg and Christian (5) reported the first discoveries on Glc 6-PDH and the conversion of glucose 6-phosphate (Glc 6-P) to 6-phosphoglucono-δ-lactone, $K_{eq} = 6 \times 10^7$ (Fig. 1). Warburg thereby established the existence of new reactions for the oxidative metabolism of glucose that differed from those of the Emden-Meyerhof glycolytic pathway. During the next decade, Warburg, Dickens, Lipmann, Diche, and others showed, among other things, that the above lactone, a potential acid, was hydrolyzed by lactonase (EC 3.1.1.17) to form 6-phosphogluconic acid (6-PG). The thermodynamic constants for the combined actions of Glc 6-PDH

The Oxidative Segment of the Pentose Pathway


to witness a demonstration of some remarkable and puzzling results that Barron and Harrop had published recently (3). The demonstration involved a simple system, namely, the reaction of glucose with non-nucleated erythrocytes in the presence and absence of methylene blue (MeB). Methylene blue is one of a large number of organic dyes that form electromotively active oxidation–reduction systems. MeB had been used successfully by Torsten Thunberg, during the previous decade, for the identification of many dehydrogenase reactions. However, considerable diplomacy was needed for a peaceful presentation of the Barron reaction system because Warburg had been publicly scathing in condemnation of biochemical reactions that included MeB. Notwithstanding this prejudice, when Warburg observed the outcome of the Barron experiments, he was truly astonished. In 1929, it was accepted that all glucose metabolism in erythrocytes was by glycolysis with the formation of two equivalents of lactate, which were diffused by the cells. Contrary to this expectation, the Barron data with MeB as a reactant showed a 50% suppression of the extent of glycolysis and lactate formation while the rate and extent of glucose utilisation was maintained and accompanied by a rapid oxygen uptake. Florink (4) wrote, "In Baltimore Warburg clearly saw, for the first time in his experience, a dehydrogenase at work even though methylene blue was involved."

Warburg returned to Berlin where, with his colleague W. Christian, he repeated and extended the Barron work and over the next 18 months showed that the phenomenon, which was first witnessed in Baltimore, comprised the following sequence of reactions (reactions (1)–(5)).

D-glucose 6-P + H2O + 2 NADP+ → CO2 + pentose-5-P + 2 NADPH + 2 H+ (1)

2 NADPH + 2 H+ + 2 FMN → 2 NADP+ + 2 FMNH2 (Rapid reaction) (2)

2 FMNH2 + O2 → 2 FMN + 2 H2O (Slow reaction) (3)

2 FMNH3 + 2 MeB → 2 FMN + 2 MeB H+ + 2 H+ (Fast reaction) (4)

2 MeB H+ + 2 H+ + O2 → 2 MeB + 2 H2O (Very fast reaction) (5)

Reactions (1)–(5) encompass the disclosure of 1) the reactions of the oxidative segment of the PPP, 2) the discovery of a new pyridine nucleotide-NADP+, and 3) the detection of a flavin mononucleotide (FMN) flavoprotein (initially called "Old yellow enzyme") and the recognition of its role as a catalytic carrier of reducing equivalents from reduced pyridine nucleotide to molecular oxygen via MeB. No contradiction can exist that the unravelling of the reactions of reactions (1)–(5) marked a historic moment of high achievement in biochemical research. The achievement was to lead Warburg and Christian to discover and characterize glucose 6-phosphate dehydrogenase (Zwischenferment) (Glc 6-PDH) (EC 1.1.1.49) and NADP+ (Wasserstoffübertragenden).

In 1931, Warburg and Christian (5) reported the first discoveries on Glc 6-PDH and the conversion of glucose 6-phosphate (Glc 6-P) to 6-phosphoglucono-δ-lactone, $K_{eq} = 6 \times 10^7$ (Fig. 1). Warburg thereby established the existence of new reactions for the oxidative metabolism of glucose that differed from those of the Emden-Meyerhof glycolytic pathway. During the next decade, Warburg, Dickens, Lipmann, Diche, and others showed, among other things, that the above lactone, a potential acid, was hydrolyzed by lactonase (EC 3.1.1.17) to form 6-phosphogluconic acid (6-PG). The thermodynamic constants for the combined actions of Glc 6-PDH
The F-type Pentose Pathway

![Diagram of the F-type Pentose Pathway]

The sum reaction for the oxidative segment of the PPP is

$$\text{Glc 6-P} + \text{NADP}^+ + \text{O}_{2} \rightarrow \text{G6P} + \text{NADPH} + \text{H}^+$$

The cellular oxidant/antioxidant balance is also dependent on the oxidative damage to DNA, lipids, and proteins. The regulation of the following reactive oxygen system (ROS) enzymes: glutathione reductase, glutathione peroxidase, catalase, superoxide dismutase, NADPH oxidase, and nitric oxide synthase depend on the high value of the NADPH/NADP⁺ ratio (approximately 80/1, compared with 8.6 × 10⁻⁴ for the cytoplasmic NADH/NAD⁺ ratio) as the source of cytoplasmic reducing power (electrons) for biosynthetic and other reactions that require reducing equivalents in aerobic organisms. A reducing equivalent is equal to one electron or one hydrogen atom. Although the PPP is the most usual source of biosynthetic reducing power in cytoplasm, a singular outstanding exception to this generality was found in adipose tissue by Flatt and Ball (12), who showed that 50% of the very large demand for NADPH was supplied by the action of L-malate:NADP⁺ oxidoreductase (oxaloacetate decarboxylating [EC 1.1.1.40]), and the remaining half was supplied by the PPP.

Glc 6-P DH exists in multiple molecular forms and is subject to wide genetic, endocrine, and nutritional variability. Its deficiency is the most common enzymopathy in humans. It is also the most sensitively regulated enzyme in the whole array of reactions comprising the oxidative and nonoxidative segments of the PPP (Figs. 1 and 2). More than a dozen pathways of biosynthesis use NADPH as a substrate (e.g., synthesis of fatty acids, steroids, some photosynthetic products, isoprenoids, sphingosines, phenylalanine and tyrosine, seminal fructose, 3-hydroxykynurenine, and NAD⁺). A high level of unbound NADP⁺ is mandatory for the reductions of oxidized glutathione, dihydrofolate, D-glucuronate, ribonucleotides, cytochrome P450, nitrate and nitrite, glucose production from pyruvate via malate, and the hydroxylation of some fatty acids. The cellular oxidant/antioxidant balance is also dependent on NADPH production principally derived from PPP activity. The regulation of the following reactive oxygen system (ROS) enzymes: glutathione reductase, glutathione peroxidase, catalase, superoxide dismutase, NADPH oxidase, and nitric oxide synthase depend on the high value of the NADPH/NADP⁺ redox couple. Thus, Glc 6-P DH has a pivotal role in the management of oxidative stress and its consequences in normal physiology and pathophysiologic events such as carcinogenesis and aging. Consult Reference 13 for a review of Glucose 6-phosphate dehydrogenase and Reference 14 for a brief review of all other enzymes shown in Figs. 1 and 2.
The Reactions of the Nonoxidative Segment

A synoptic survey of the unravelling of the nonoxidative segment is presented in the latest comprehensive review of the PPP (1). The treatment covers the chemistry of reactions, enzymes, methods, tissue distributions, and the mathematical theory for the quantitation of metabolism by a theoretical connection of the pathways of Figs. 1 and 2 into a pentose phosphate cycle (PC).

Early Discoveries

The initial investigations that lead to chemical knowledge of the pathway of Fig. 2 were indirect and not planned to that end as the following narration will show. Warburg's proposition linking NADPH and respiration (reactions (2)-(4)) was shared by Erwin Haas who was a member of Warburg's Berlin-Dahlem laboratory. Haas departed Germany in 1938 and proceeded to the laboratory of Professor T. R. Hogness (University of Chicago), where he met Bernard Horecker, who had just completed Ph.D. training in enzymology. Haas was a meticulous scientist who possessed much of Warburg's data and methods (of which he was something of a master) and a plan of research to test the proposed role of NADPH in respiration. Haas and Horecker set out to isolate a putative NADPH-cytochrome reductase (reaction (6)) in order to demonstrate the existence and nature of an enzyme that was hypothesized to be the missing link in a respiratory pathway between reduced pyridine nucleotide and oxygen via the cytochrome system (reactions (7) and (8)).

NADPH + H+ + Cyt-C Fe3+ → NADP+ + Cyt-C Fe2+ + 2H

Cytochrome Oxidase (EC 1.9.3.1)

2H + Cyt-C Fe2+ + 0.5O2 → H2O + Cyt-C Fe3+ (8)

Sum Reaction

NADPH + H+ + 0.5O2 → NADP+ + H2O

The search for an enzyme activity was successful, and a reasonably pure flavoprotein, NADPH-cytochrome c reductase (EC 1.6.2.4), was isolated from yeast (15). It was then 1940, and basic research was interrupted and not taken up again until the end of the Second World War. During that interval, Horecker worked on war-related programs, was appointed to the staff of the National Institutes of Health, and in 1945 returned to basic research in enzymology. He proceeded to isolate NADPH-cytochrome c reductase from acetone-dried extracts of pig liver (16), at a time that was almost coincident with the report of Lehninger (10) that NADH was the substrate of respiration. As listed above, the true acceptor for NADPH-cytochrome c is cytochrome P450. This natural disappointment drew attention to a much larger problem, namely an inquiry into the chemistry, enzymology, and metabolic fate of the pentose phosphate product of Fig. 1. The British biochemist, Frank Dickens (Courtauld Institute, London U.K.), had been making pioneering investigations on this topic since 1936. However, a solution of the problem was going to be made in the United States, where metabolic biochemistry and enzymology flourished in the post-war vigour of the 1950s. It is helpful to note the intense interest at the time in the nature and significance of the new pathway of Glc 6-P oxidation. Important findings were going to be made in the laboratories of Bernard Horecker, Ephraim Racker, Seymour Cohen, Bernard Axelrod, and Gilbert Ashwell. Moreover, a most promotive stimulus was the outstanding investigation then being made by Melvin Calvin and colleagues at UC Berkeley. Their work was to lead to a Nobel Prize and to our current chemical appreciation of the path of reductive carbon fixation by photosynthesis (PS). Calvin’s great progress, during 1950-1955, was heavily dependent on success by the above biochemists to solve the chemical and enzymological problems posed by the reactions of Fig. 2. A notable number of the intermediates and enzymes of Fig. 2 were identified as reactants in the path of carbon in PS (17). Also, Horecker made early and successful attempts to compete with Calvin in the quest to map the enzymology of the carbon reduction cycle. By 1950 all investigators possessed strong clues that were to serve as signposts for an ultimate elucidation of a PPP reaction scheme. These signposts were 1) clear evidence that an alternative path of Glc 6-P oxidation existed in yeast, some bacteria, red cells, liver, and other animal tissues. 2) Dickens (18) had confirmed that 6-PG was decarboxylated oxidatively at carbon one to yield ribose 5-phosphate (Rib 5-P) and other sugar phosphate products (all unresolved), including a putative tetrose P. He also demonstrated that Rib 5-P was oxidized at five times the rates of arabinose 5-phosphate (Ara 5-P) and xylose 5-P, both of which are theoretical products of 6-PG decarboxylation. 3) Finally, Dische (19) reported that inosine monophosphate was oxidized to xanthosine by xanthine oxidase. This last important finding of possible end products of Rib 5-P dissimilation was confirmed by Waldvogel and Schlenk (20) who showed Glc 6-P formation from Rib 5-P using rat liver extracts.

Post-1950 Discovery of Reactions for the Nonoxidative Segment of the PPP

With the above background, between 1950 and 1955, a breathing-taking series followed of discoveries of enzyme and substrate reactivities that were destined to be incorporated into a reaction scheme (mechanism) for the classic nonoxidative PPP (Fig. 2). The Fig. 2 diagram has also been given the prefix F-type (for fat-cell) PPP. Williams et al. (1), because its ordered reaction sequence was later shown to measure uniquely a large contribution (50%) to metabolism when Glc was converted to fatty acids.
acids and triglyceride by insulin-stimulated adipocytes (12, 21, 22). An assembly of reaction steps for the scheme of Fig. 2 may be theorized using the conjunction of results from the following temporal list of events. In 1951, Cohen’s group (23) showed that Rib 5-P and Ara 5-P were formed from 6-PG oxidation. Rib 5-P formation was confirmed by Horecker et al. (7) who also proved unequivocally that Ru 5-P was the first pentose-P formed when 6-PG was decarboxylated. A new enzyme, Ribose 5-phosphate isomerase (R 5-P I) (EC 5.3.1.6), that catalyzed the interconversion of the two pentose phosphates was also proposed (7) (Fig. 2). A very active preparation of R 5-P I from Alfalfa (24) was used to demonstrate its reaction and equilibrium. Horecker and Smythidios (25) used a liver enzyme preparation with Rib 5-P to note the formation of an initially high concentration of se-dihydroxyacetone 7-phosphate (Seh 7-P). This seven-carbon ketulose ester had been found originally and was characterized by Andy Benson et al. (26) in Calvin’s laboratory, and it was shown to be an early product of PS carbon fixation. Seh 7-P was formed by the action of Transketolase (TK) (EC 2.2.1.1) (Fig. 2). TK was discovered by Racker and his collaborators (28–29) who demonstrated that it catalyzed the transfer of a two-carbon fragment (an active glycolaldehyde group) from appropriately structured ketulose-sugar donors to a wide selection of aldo-sugar acceptors (Reference 1, p. 754). Two of its donor transfer actions, using different aldo-acceptors, are shown as blue rectangular panels in Fig. 2. TK requires Mg2+ and thiamine pyrophosphate (TPP) as a coenzyme, and a two-carbon group is transferred to its acceptor via an hydroxyethyl-TPP intermediate. A list of 15 of its glycolaldehyde acceptor substrates is tabulated (Reference 1, p. 754). Horecker’s group (30) discovered a second broad specificity group transferring enzyme, namely Transaldolase (TA) (EC 2.2.1.2), which catalyzed the reversible transfer of a three-carbon dihydroxyacetone-enzyme-bound moiety (shown as red panels in Fig. 2) from Seh 7-P to glyceraldehyde-3-phosphate (Gra 3-P). This reaction formed Fru 6-P and a presumed tetrose phosphate that owed its assigned identity more to biochemistry because it was neither isolated nor identified. The availability of synthetic erythrose 3-phosphate (Ery 4-P) enabled Kornberg and Racker (31) to demonstrate the reversal of the TA reaction [reaction (9)]; thus, an authentic reason for its inclusion as an intermediate in the reaction scheme of Fig. 2 was provided. Ery 4-P probably only exists in exceedingly low concentration in any tissue, and to date, no evidence exists that it has ever been measured correctly in, nor isolated from, any preparation carrying out PPP or PS metabolism in vivo or in vitro (32, 33).

Fru 6-P + Ery 4-P ←→ Seh 7-P + Gra 3-P (9)

A third ketopentulose ester, xylulose 5-phosphate (Xlu 5-P) (Fig. 2), was isolated as a product of Rib 5-P metabolism by Ashwell and Hickman (34). It was also shown by Srere et al. (35) that Xlu 5-P, rather than the earlier assigned Ru 5-P, was a definitive substrate of TK. Ribulose 5-phosphate-3′-epimerase (Fig. 2) catalyzed the formation of Xlu 5-P and imparted the transconfiguration to the hydroxyls at carbons 3 and 4, which is a necessary stereochemical condition for substrate reactivity with TK. The 3′-epimerase was purified from a bacterial source by Stumpf and Racker (36) and from muscle by Dickens and Williamson (37). In summary, the above research seemed to uncover an array of substrates and enzymes that satisfied a minimum requirement for inclusion in a new pathway that connected the product of 6-PG decarboxylation with the formation of hexose and triose phosphates. A prescient Racker (38) was the first to present Fig. 2, a theoretical cyclic metabolic pathway with stoichiometry for an ensemble of the above reactants and enzymes.

Search for an Order of Reactions in the Nonoxidative PPP

It is possible to draw various theoretical schemes that oblige the arithmetic conjuction of five-carbon sugars with a summary outcome of sugar products that contain six-carbon and three-carbon atoms, respectively (39). That variety is greatly enhanced if reactions catalyzed by aldolase (Ald) (EC 4.1.2.13) are included. Aldolase occupies the same cellular compartment as all other enzymes of the PPP; it is a dihydroxyacetone 3-phosphate (DHAP) group transferring enzyme, with a catalytic capacity that is usually much greater than TK or TA (a notable exception is adipose tissue where Ald activity is low (21, 22) and only approximates the activity of TK and TA). Ald also has a broad substrate array of aldo-sugar phosphate acceptors (Reference 1, p.754), most of which are the same substrates as those involved in TK and TA reactions. It has never been clear why the pioneering investigators of the nonoxidative PPP assigned aldolase a role of catalytic silence.

The results of the only foundational experiments that explored the identity of a reaction sequence (mechanism) for the PPP were published by the Horecker group (40, 41). Detailed treatment of the conduct and conclusions of these important, but rarely discussed, experiments is given in References 1 and 42. A predication labeling technique that used (1-14C) and (2,3,4-14C)-Rib 5-P as substrates was adopted. These substrates were reacted with fractionated enzyme extracts of acetone powder preparations from rat liver, pea leaf, and pea root tissues and formed, among other intermediates, C14-labeled Glc-6-P (Fig. 3). It was anticipated that both the position and the amount of 14C-label imparted to the Glc 6-P formed by the above substrates would reveal the nature and order of the reactions involved in its formation. Because the enzyme extracts were made from acetone-dried powders, they were free of all nucleotides (40, 42). This strategy confirmed the dissimilation of labeled Rib 5-P to a hexose 6-P end point. Glc 6-P could not recycle, and the labeled prediction pattern was not scrambled. Mg2+ was also omitted from the reaction mixture to inhibit the activity of Fructose bisphosphatase (EC 3.1.3.11) and production of a contaminating 6-P from 1,6-P2 formed by aldolase and the triose-P products of the TK reactions. It is curious that Fru 1,6-P2, which is an Ald product, was formed (Fig. 3) by the liver and pea tissue reaction mixtures but was omitted from the final construction of the scheme for Fig. 2. The experiments with liver enzyme preparation were of a 17-hour duration (Fig. 3). Ribose 5-P was used rapidly during the experiments.
with a C-1/C-3 isotope ratio of 3 (74% of the C-14 isotope in C-1 and 24% in C-3) (40). With a sense of caution, Horecker (40) proposed an tentative chart of the PPP or pentose cycle (PC). It is astonishing that such substandard and unacceptable disagreement between practice and theory was so uncritically ignored by the general community of biochemists. Harland Wood (46) was an exception: he reported a tolerant suspension of the [1-14C]-Rib 5-P data and total rejection of any claim for predictive value by the [2,3-14C]-Rib 5-P results. It was suggested (46) that an as yet unidentified pathway was operated by the above enzyme fractions. Notwithstanding the rejection by Wood (46), without anymore revision, a placid acceptance followed as well as an equally prompt inclusion of F-type PPP into the canon of metabolic biochemistry. Two independent investigators did test the mechanism of the PPP; the first by Katz et al. (47) used [1-14C]-Rib metabolism in liver slices, and the other by Hiatt (48) used the same labeled substrate in mouse liver in vivo. These investigators did not find 14C distributions in the glucose product with twice as much isotope in C-1 as C-3, but instead they found these carbons were equally labeled. These independent failures to confirm the predictions of Fig. 2 were also ignored. Finally it is important to note that inclusion of several PPP enzymes and reactants in the Path of Carbon in P5 gratuitously added an aura of confidence and prestige to the status of the tentative PPP. Thus, 1955 marked an end of the second contentious period in the unraveling of the pathway.

Other than immensely valuable research by Patricia McLean and her collaborators at the Courtauld Institute, U.K., and investigations by Karl Brand at the Max Planck Inst, Dortmund, Germany, fundamental research on the mechanism of the PPP essentially ceased by 1957 and was not resumed for another decade. Instead the era of the quantitative measurement of pathways of carbohydrate metabolism had dawned and PPP measurements featured hugely. This emphasis on quantitation is best summarized in the following quotation from H. G. Wood (46): “The determination of the relative role of different pathways in normal living cells is without doubt of the greatest fundamental importance to our understanding of life processes and will in the future require more attention in all fields of metabolism.” Later Wood and Katz collaborated (49, 50) and during the next eight years developed theory and methods for the measurement of an entity denoted by Wood (46) and later stringently defined by Wood and Katz (49) as PC. The PC definition directed the following set of ordered events: (i) the entry of three moles of Glc 6-P into metabolism by the oxidative

### The Pentose Phosphate Pathway: An Overview

<table>
<thead>
<tr>
<th>Metabolite Concn. (mM)</th>
<th>Time (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>3.0</td>
<td>2.0</td>
</tr>
<tr>
<td>4.0</td>
<td>3.0</td>
</tr>
<tr>
<td>5.0</td>
<td>4.0</td>
</tr>
<tr>
<td>6.0</td>
<td>5.0</td>
</tr>
<tr>
<td>7.0</td>
<td>6.0</td>
</tr>
<tr>
<td>8.0</td>
<td>7.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.0</td>
</tr>
<tr>
<td>10.0</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Figure 3 - Conversion of ribose 5-phosphate to sugar phosphate products by rat liver enzyme preparation (RLEP) (40). Details of the incubation, protein content and methods used are described in (42). For measurement of the max. catalytic capacity of RLEP see (43). Reproduced from International Journal of Biochemistry, 19. Williams John E., Arora Krishan K. and Longnecker John P. The pentose pathway: A random harvest, 69 Pages, 1987, with permission from Elsevier, http://www.sciencedirect.com/science/journal/13572725
Pentose Cycle

The Pentose Phosphate Pathway: An Overview

Measurement of the Pentose Cycle: Theory and Practice

Search for an Alternative Reaction Scheme for the Pentose Pathway

A detailed treatment of this topic, including the derivation of some important mathematical expressions is given in Reference 1 (pp.753–762), and only a summary of the essentials will be presented here. Between 1958 and 1979, a dozen theoretical papers were published that provided the mathematical basis and formulas for measuring the F-type PC by the application of 14C-specifically labeled substrates. The method of measurement depended on an exact solution of the problem posed by the recycling and, thus, changes to 14C isotope distributions in Glc 6-P, which emerge from the metabolism of the substrates (2,14C)-glucose or (3,3-14C)-glucose in PC. Calculating the different distributions of labeled carbon to infinite cycles, for all percentage contributions of PC, is a difficult mathematical problem that was solved entirely by Joseph Katz. Katz is not only a gifted biochemist but also an equally talented mathematician and innovative metabolic theorist. The acceptance of a PC definition imposed an agreement that all 14C-labeled Fru 6-P formed by the reaction sequence is converted to Glc 6-P and recycled again through the oxidative segment reactions. The quintessence of all measurement methods has involved the development of mathematical expressions that describe the mathematical and ordered redistributions of either carbons 2 or 3 from labeled substrate glucose, into positions 1, 2, and 3 of the hexose 6-P products for any percentage contribution of PC. Such a theoretical distribution is a unique property of the PC. Experimental data for the C-1/C-2 and C-3/C-2 ratios, which have definite limits and values imposed by Katz and Wood theory (49, 50), are used in specially derived equations that calculate the PC contribution relative to the total metabolism of glucose. The above statements cannot be qualified. They derive from the fixed order of the reactions of Fig. 2, which is the mechanistic basis for all the formulas of all the measurement papers (Reference 1, p. 753). The amounts of 14C in C-1, C-2, and C-3 are derived from their earlier reservation and criticism of the substandard evidence for the reaction sequence of PPP (48) and the failure (by JK) to confirm its presence in liver (Reference 1, p. 766). Liver is a very rich source of the enzymes of the PPP (6, 14, 42, 52), and it provides, among other things, a ready display of the reactions of Figs. 1 and 2. Thus, the failure by all measurement investigations to find an F-type PC in liver was mystifying. The mystery deepened when it was noted that the formulation of their elegant measurement theory required Katz and Wood to abandon, indeed fail to mention, their earlier reservation and criticism of the substandard evidence for the reaction sequence of PPP (48) and the failure (by JK) to confirm its presence in liver (47).
The L-Type Pentose Cycle

The following three sets of findings summarize selected aspects of progress in the unraveling of a new PPP reaction sequence in liver that was found to measure 20-30% of total glucose metabolism. The experimental history of these events is fully recorded by Williams et al. in References 42 and 53 and in Reference 1 (pp. 766-798).

First, the foundational experiments that established the F-type PPP were repeated using [L-14C]Rib 5-P and the same rat liver enzyme preparation. However, the reaction mixture was sampled for the labeled Glc 6-P product at a series of much shorter time intervals and right up to the 17-hour termination point that was described for the original work. The results (Table 2) showed a patterned assortment of label distributions in Glc 6-P, which drifted from 8 hours to 17 hours toward the isotope composition in C-1 and C-3 that was noted originally by Horecker et al. (40). However, in this study, the C-1:C-3 ratio at 17 hours was the prized value of 2. Moreover in the seven time samples, which commenced at 1 minute, Glc 6-P was heavily labeled in C-2, C-4, and C-6, whereas C-1 and C-3 only began to accumulate 14C-isotope after 3 hours of reaction. Although this was a study in vitro, it is obvious that liver cells in vivo do not take between 3 and 17 hours to elaborate a path of metabolism and that more enlightening events were being revealed by the isotope distributions in the samples analyzed between 1 minute and 30 minutes of reaction. Second, a carbon balance analysis of all compounds in the various reaction mixtures showed that the intermediates of Fig. 2 only accounted for 80% of the carbon in the Rib 5-P substrate (42). The compounds comprising the missing 20% were identified as sugar phosphates, mostly ketolases (Fig. 4). These new reactants were isolated, shown to be radioactive, and identified as Seh 1,7-P2, D-manno-Heptulose 7-P, D-glycero-d-altro-Octulose 1,8-P2, (D-g-D-a-Oct); D-glycero-d-idulo-Octulose 1,8-P2, (D-g-D-i-Oct) and a small amount of Ara 5-P (42). Octulose monophosphates and bisphosphates and Seh 1,7-P2 were measured in fresh liver by Paoletti et al. (54) and in spinach chloroplasts by Flanagan et al. (33). Last, the structures and order of the reactions of these sugar esters in a new and much modified reaction scheme for the PPP in liver are shown in Fig. 4. The new intermediary compounds were isolated easily from all incubations from 30 minutes to 17 hours. The scheme of Fig. 4 shows the new PPP with prediction 14C-labeling patterns in the intermediates and products of the reactions. The reaction scheme of Fig. 4 was formulated initially from the distributions of 14C in the labeled Glc 6-P (42) and D-g-D-a-Oct 1,8-P2 (55) formed from [L-14C]Rib 5-P during the early intervals of the repeat of the foundational experiment (42). The new pathway, called L-(liver) type PPP, is distinguished from depictions of the classic F-type PPP by the inclusions of Seh 1,7-P2 and octulose-(D-G)-monophosphates and bisphosphates together with Ara 5-P as new intermediates. Aldolase, phosphotransferase (PT), and D-Arabino-s-phosphate ketol isomerase (EC 5.3.1.13) are new enzymes. The effects of mass transfer catalysis by TA were TA-exchange reactions (TA-ex.), which contributed the [4,6-14C]-labeling pattern to Glc 6-P during the first 8 hours were active (1, 42, 56). The TK reaction forming hexose 6-P in the L-type PPP (Fig. 4) used D-g-D-i-Oct 1,8-P2 as substrate. TK and Ald were also very active exchange catalysts (57-59).

A clear demonstration that aldolase is a mandatory catalyst in liver PPP involved the immunochromic evidence of Bleakley et al. (43) who showed the total cessation of hexose 6-P formation when liver aldolase antibody titrated the removal of aldolase from the system where Rib 5-P was reacted with the same rat liver enzyme preparation that established the Fig. 2 scheme. Irrespective of the other contrary data, this evidence alone showed another reaction mechanism involved Ald in liver PPP. The claim that aldolase is an essential enzyme in the PPP was also supported by data of (60) in which, using an in vitro construction of a PPP preparation for the complete oxidation of Glc, noted the formation of Oct-P and the need to include aldolase and sedoheptulose 1,7-bisphosphatase for the construction system to work.

Exposing the Problem of Assigning a Reaction Scheme to the Nonoxidative PPP

The reactions and enzymes shown in Fig. 4 occur in liver cytoplasm and are expected items in the soluble enzyme compartments of most animals. Aldocyclases as well as perhaps the lactating mammary gland and some microorganisms may be exceptions. However, the failure to find incontrovertible evidence for both F-pathways and L-pathways that permitted the placement of C-3 to C-8 glycolyl phosphates in a reaction order and with stoichiometry that satisfied flux demands of a metabolic pathway or cycle is the crux of this 50-year-old enigma. Finding an answer to the problem proved to be quick, obvious, and simple. Uncovering irrefutable evidence and proof for the explanation was a more pressing task that was solved by Flanagan et al. (61). It was the universal use of 14C isotopes in prediction-labeling experiments to both inquire into mechanism and predicate theories for quantifying the F-type and L-type PPP that was the first fundamental error. The second misjudgment was lack of practical attention to the consequences of the glycolyl-group exchange reactions that are actively catalyzed by TK, TA, and Ald. The subscript X is used to distinguish exchange catalysis from the mass transfer activity of these enzymes. Despite some anecdotal warnings by the authors of References (49-51) that both exchange and mass transfer activities could be anticipated in PPP and PS investigations, the warnings were unheeded by the PC measures. A rare exception, in which an exchange-enzyme control accompanied measurements of the L-type PC, is discussed by Williams (1) and Longnecker and Williams (62).

13C-NMR spectroscopy was used by Flanagan et al. (61) for the kinetic investigations of exchange reactions by group-transfer enzymes. Adoption of NMR technology silenced the unfailing criticism of 14C labeling patterns that were measured by wet chemical methods. The maximum catalytic capacities for exchange by the three enzymes were all quantified in reaction mixtures at mass-transfer equilibrium. TA was measured by the
Table 1  Percentage distribution of $^{14}C$ in glucose 6-phosphate formed during the time course of reactions of [1-$^{14}C$]-ribose 5-phosphate with rat liver enzyme preparation (42)

<table>
<thead>
<tr>
<th>Carbon number</th>
<th>1 min</th>
<th>2 min</th>
<th>3 min</th>
<th>30 min</th>
<th>3 h</th>
<th>8 h</th>
<th>17 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.40</td>
<td>0.70</td>
<td>1.30</td>
<td>1.20</td>
<td>28.50</td>
<td>40.50</td>
<td>56.50</td>
</tr>
<tr>
<td>2</td>
<td>44.10</td>
<td>43.80</td>
<td>13.00</td>
<td>11.80</td>
<td>17.70</td>
<td>17.60</td>
<td>12.40</td>
</tr>
<tr>
<td>3</td>
<td>1.10</td>
<td>0.10</td>
<td>3.30</td>
<td>1.10</td>
<td>10.50</td>
<td>16.50</td>
<td>24.40</td>
</tr>
<tr>
<td>4</td>
<td>12.20</td>
<td>2.60</td>
<td>5.10</td>
<td>7.10</td>
<td>5.50</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>7.10</td>
<td>0.40</td>
<td>0.60</td>
<td>0.00</td>
<td>4.60</td>
<td>1.60</td>
</tr>
<tr>
<td>6</td>
<td>40.70</td>
<td>45.70</td>
<td>76.90</td>
<td>78.20</td>
<td>37.80</td>
<td>21.80</td>
<td>3.10</td>
</tr>
<tr>
<td>Recovery (%) of $^{14}C$</td>
<td>105.1</td>
<td>100.8</td>
<td>106.2</td>
<td>103.4</td>
<td>96.9</td>
<td>98.7</td>
<td>100.5</td>
</tr>
</tbody>
</table>

Figure 4  L-type pentose phosphate pathway reactions. The [1-$^{14}C$] ribulose 5-P substrate is the product of reactions shown at Fig. 1. The reaction sequences illustrate the distinctive distribution of $^{14}C$ in hexose 6-P and the following new intermediates: octulose monophosphate, sedoheptulose and octulose bisphosphates. Transaldolase exchange reactions involving ribose and arabinose 5-phosphates are shown. Other detail of the L-type pathway is given in the text. Reprinted from International Journal of Biochemistry, 19. Williams John F., Arora Krishan K. and Longnecker John P. The pentose pathway: A random harvest, 69 Pages, 1987, with permission from Elsevier, http://www.sciencedirect.com/science/journal/13572725

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc. 9 - 1262 -
exchange rate of the 13C-TA dihydroxyacetone group with unlabeled Seh 7-P. TKx was measured by the rate of incorporation of a (2-13C)-TK glycolaldehyde group to unlabeled Fru 6-P and Ald, by the exchange of an unlabeled DHAP-Ald group from D-g-D-a-Oct 1,8-P2 to [1-13C]-Rib 5-P and measurement of the rate of formation of [4-13C]-D-g-D-a-Oct 1,8-P2. A comparison of the exchange capacities of these enzymes with the maximum nonoxidative PPP flux rates in three liver preparations showed that TKx and Aldx exceeded flux by 9-19 times in liver cytosol and acetone powder enzyme preparations in vitro and by 5 times in hepatocytes. TAx was less effective in exchange, only exceeding the flux rate by 1.6 times and 5 times in liver cytosol and acetone powder preparations, respectively. Values for the ratios of the rate of group exchange and pathway fluxes in liver are important because of the feature roles of hepatic tissue and of these preparations in the establishment and status of the Fig. 2 and Fig. 4 schemes in biochemistry. The prevalence of exchange activity was also investigated using the dominant TKx rates relative to the maximum PPP flux rates of normal diemal adipocytes. TKx rates in these preparations exceeded melarona, colonic epithelium, spinach chloroplasts, and epi-
didymal adipocytes. TKx rates in these preparations exceeded PPP flux by 5-600 times (61).

Conclusion

A surfeit of evidence has accumulated slowly during the last 53 years to show that predictions and calculations based on their distributions in PPP and PS products and intermediates are pointless and misleading. The isotope patterns cannot reveal the order of the reaction sequences, which can only be measured by an uncompromised net flow of carbon. Instead the 14C-patterns are used by measurement theoreticians, in the PC reaction mechanism and rigorous chemical procedures. Since its first appearance in biochemical texts. In his monograph (11), Terry Wood correctly concluded that “radioisotopic studies of PPP and some neoplasms and photosynthetic tissue. The case for the L-type pentose pathway. Int. J. Biochem. 1987;29:479-82.


1. Williams JF, Arora K, Longnecker JP. 'The pentose pathway: a random Harvest. Impediments which oppose acceptance of the classical (F-type) pentose cycle for liver, some neoplasms and photoglycotic tissue. The case for the L-type pentose pathway. Int. J. Biochem. 1987;29:479-82.


6. Horecker BL, Smyrniositz pd, Seegmiler Jeff. The enzymatic con-


8. Engelhardt WA, Barakh AP. Oxidative breakdown of phospho-

9. Lehninger AL. Phosphorylation coupled to oxidation of di-
hydrodiphosphopyridine nucleotide. J. Biol. Chem. 1950;190:
345-359.


12. Horecker BL, Smyrniositz pd, Seegmiler Jeff. The enzymatic con-

13. Engelhardt WA, Barakh AP. Oxidative breakdown of phospho-

14. Lehninger AL. Phosphorylation coupled to oxidation of di-
hydrodiphosphopyridine nucleotide. J. Biol. Chem. 1950;190:
345-359.


17. Lehninger AL. Phosphorylation coupled to oxidation of di-
hydrodiphosphopyridine nucleotide. J. Biol. Chem. 1950;190:
345-359.
The Pentose Phosphate Pathway: An Overview

The Pentose Phosphate Pathway: An Overview


Further Reading


See Also

Glucose Homeostasis, Chemistry of
Lipid Homeostasis, Chemistry of
Metabolic Labelling of Sugars
Metabolism, Cellular Organization
Nucleic Acid Metabolism, Chemistry of
Redox Regulation and Signalling: Reactive Oxygen Species (ROS)
The network structure of pathways can be studied from two complementary viewpoints: as networks of enzymes or as networks of chemical compounds. This structure enables more in-depth analysis into metabolic pathways. From these networks, new features regarding pathways on both the local and the global levels can be detected. On the one hand, global features such as the scale-free property of pathways have attracted much attention from the bioinformatics community. On the other hand, local features of networks such as pathway modules can retrieve and characterize subnetworks of related genes that are potentially involved in a particular function of the metabolic pathway. Systems analysis of metabolic pathways must focus not only on existing pathways, but also on reconstructing pathways for new genomes or filling in information regarding missing enzymes. By using the vast amounts of genomic data available, it is possible to reconstruct the metabolic maps of new genomes. Such genomic information has proved useful to refine prediction methods, and they can be complemented with chemical-based information that is inherent in the same network. Overall, a systems approach to metabolism covers the realm of both the genomic and the chemical worlds in an integrated manner. We will show that the concepts of local network features in terms of both these worlds produce modules that can be integrated such that the global view of metabolism can be grasped. The current findings will be described systematically while also involving manual curation such that biologically accurate systems can be produced for analysis.

In bioinformatics, the term “systems approach” is often contrasted to the reductionist approach, in which a large system is broken down into its parts and the parts are studied individually. That is, based on systems theory, a network is studied from the perspective of the organization (relationship) of its parts, from which patterns may emerge. Therefore, we look at the metabolic network in an integrated manner covering the realm of both the genomic and the chemical worlds to identify features that emerge from the network. We will show that the concepts of local network features in terms of both these worlds produce modules that can be integrated such that the global view of metabolism can be grasped. The current findings will be described systematically while also involving manual curation such that biologically accurate systems can be produced for analysis. The recent advancements in systems analysis of metabolic pathways will be introduced.

**Systems Analysis of Metabolic Pathways**

Metabolic pathways have been illustrated using simple diagrams since before the human genome project and related bioinformatics projects had begun. With the involvement of computer science techniques, however, systematic approaches to modeling metabolic pathways have progressed quickly, with various aims that range from metabolite analyses to pathway prediction and reconstruction (1–3). Systems analysis has come to incorporate...
graph-theoretic techniques on the one hand, and physics on the other hand, in the attempt to elucidate the complex functioning of the cellular system. In terms of graph theory, in particular, the network structure of pathways has been studied with complementing views by considering them as networks of enzymes or as networks of chemical compounds to capture more and important information. From these networks, new features regarding pathways on both the local and the global levels have been detected. On the one hand, global features such as the scale-free property of pathways have attracted much attention from the bioinformatics community. These properties have shown that metabolic networks are not so different from other well-known networks such as social networks and the Internet. They also helped to characterize networks in a systematic manner such that particular enzymes that are either undefined (missing) or have important roles in the network could be identified and studied in more depth. On the other hand, local features of networks such as pathway modules can retrieve and characterize subnetworks of related genes that are involved potentially in a particular function of the metabolic pathway. This latter approach of characterizing modules has been supplemented with gene expression information and analyses of chemical reaction patterns not only to infer the function of the genes involved in the particular module, but also to infer the evolution of pathways.

Systems analysis of metabolic pathways needs to focus on existing pathways; it can also be used to reconstruct pathways for new genomes or to fill in information regarding missing enzymes. By using the vast amounts of genomic data available, it is possible to reconstruct the metabolic maps of new genomes. Information of orthologous groups of genes combined with pathway data enables such predictions. Furthermore, the integration of data from a variety of resources such as microarray expression data and localization data can be incorporated in new advanced models to predict and to fill in the gaps in pathways for missing enzymes (3–5). Such genomic information is useful to refine prediction methods, and they can be complemented with chemical-based information that is inherent in the same network (6). Methods for pathway prediction can use a systematic approach to classify chemical reactions based on the specific structures of the chemical compounds involved. Because computer science techniques from graph theory can and have been applied directly for these analyses, some methods will be described later.

**Network Structure**

The analysis of network structure from the viewpoint of computer theory requires an introduction to some background information, which will be provided here. We will introduce the data involved for modeling metabolic pathways and the KEGG pathway database in particular. Furthermore, a basic introduction to graphs as used in computer science will be provided.

**Background: data models**

Several databases for metabolic pathways are available currently from the Internet, and some major representatives are listed in Table 1. KEGG (Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan), BRENDA (Institute of Biochemistry, University of Cologne, Germany), and other database systems have been applied directly for these analyses, some methods will be described later.

| Table 1: Some representative metabolic pathway databases |
|-------------------------------|-----------------|--------------------------------------------------|
| Name                          | Provider         | Description                                      |
| Biocatalysis/ Biodegradation   | University of   | Microbial biocatalytic reactions and biodegradation pathways for xenobiotic and chemical compounds |
| Database                       | Minnesota        |                                                  |
| Biochemical Pathways           | ExPA Sy          | Biochemical pathways                             |
| BioCyc Knowledge Library       | SRI International| Consists of EcoCyc and MetaCyc; collection of metabolic pathways from many organisms, respectively |
| Biomolecular Interaction Network (BIND) | Institute of Biochemistry, University of Cologne, Germany | Interaction, molecular complex, and pathway records |
| BRENDA                        |                  | Collection of enzyme functional data classified according to the Enzyme Commission (EC) list of enzymes |
| Cell Signaling Networks Database | National Institute of Health Sciences, Japan | Signaling pathways of human cells, compiling information on biologic molecules, sequences, structures, functions, and biologic reactions which transmit cellular signals |
| Enzymology Database            | Argonne National Laboratories | Detailed information on a large number of enzymes from the literature |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) | GenomNet | Computerize knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes, providing links from the gene catalogs produced by genome sequencing projects |

© 2008, John Wiley & Sons, Inc.
The reaction formulas (chemical equations) in the ENZYME information, which consists of REACTION, RPAIR, and ENZYME, is the KEGG COMPOUND database of chemical compounds. Compounds have been manually curated from the literature. Such biologic information, but also on pathway information that has orthologous groups of genes that are based not only on sequence similarities but also on pathway information. This results in orthologous groups of genes that are based not only on sequence information, but also on pathway information that has been manually curated from the literature. Such biologic information incorporated into the data means that the resulting catalog of gene groups is truly meaningful.

A another resource that is useful in metabolic pathway analysis is the KEGG PATHWAY database of chemical compounds. This database is supplemented by the database of reaction information, which consists of REACTION, RPAIR, and ENZYME. The reaction formulas (chemical equations) in the ENZYME nomenclature as well as those taken from the KEGG pathways are stored in the REACTION database, which contains, among others, the stoichiometry of substrates and products in an enzymatic reaction. To trace the atomic changes of substrates and products, the RPAIR database is constructed by decomposing each chemical equation into a set of substrate-product pairs.

The RPAIR database contains chemical structure alignments of substrate-product pairs (reactant pairs) and chemical structure transformation patterns, which were generated computationally and curated manually from all known enzyme-catalyzed reactions. These patterns are called RDG patterns, which describe biochemical structure transformations and represent KEGG atom type changes in a reaction. KEGG atom type changes are defined at the reaction center atom (R atom), its neighboring atoms in the different (mismatched) region (D atom), and the matched region (M atom), based on a graph-based alignment of the compounds involved in the reaction. The definition of a graph-based alignment between two compounds is described in the next section. Figure 1 illustrates these RDG atoms. Because these transformation patterns generalize complex enzymatic reactions, given a new set of chemical compound structures, the reactions that could possibly take place between them can be predicted.

**Background: algorithms**

The bioinformatics field has enabled the use of algorithmic techniques from computer science to analyze vast amounts of data efficiently and accurately. For the study of networks, graph models are most appropriate, and numerous algorithms exist for studying graph objects. A graph is defined as a set of nodes connected by edges, in which a node represents a specific object such as a particular chemical compound or a particular enzymatic protein, and an edge represents the relationship between two different nodes, such as the catalysis of one compound into another or a protein–protein interaction. Thus, a graph can be defined as a set of nodes V = {v1, v2, v3, ..., vn} and a set of edges E = {e1, e2, e3, ..., em} in which any edge in E connects exactly two nodes in V and no two edges share the same pair of nodes. A directed graph is a graph whose edges define a source and a target; the direction of the graph is defined, such as the direction of an irreversible reaction from substrate to product. In contrast, an undirected graph does not define any direction on the edges. The degree of a node is defined as the number of nodes with which it shares an edge. A subgraph of a graph is a graph that contains a subset V′ of the nodes in V and all those edges in E that connect those nodes in V′. Consequently, a subnetwork is a subgraph of a network modeled as a graph. We will also define here NP-completeness. A problem is NP-complete if it is known that the problem can be verified quickly, but a solution itself is difficult to find efficiently. For example, the Hamiltonian path problem is a well-known NP-complete problem. Given an undirected graph, the problem is to find a path in the graph that passes through all nodes exactly once. This path is most difficult to find, but given a path, it is easy to verify whether the given path in the graph solves this problem. Efficient methods exist to test whether two graphs are similar (or isomorphic). However, the problem of deciding whether a
One of the earliest features that characterizes metabolic compound networks is the scale-free property, which was derived from the finding that the probability that a node can interact with other nodes, which is the degree distribution $P(k)$ of a metabolic compound network, decays as a power law $P(k) \sim k^{-\gamma}$ with $\gamma \approx 2.2$ in all organisms (12–14). This scale-free property ultimately illustrated that biologic networks were not as different from other nonbiologic networks as thought previously, and that metabolic compound networks of almost all organisms thus exhibited robust and error-tolerant properties as a result. Moreover, an analysis of the scale-free properties of the line graphs of metabolic compound networks (that is, the properties of the metabolic enzyme networks) was performed (15). The network properties of the metabolic enzyme networks are not exactly one-to-one to the metabolic compound networks because several reactions may have common products, which reduce effectively the number of edges in the transformed network. Nevertheless, it was found that the scale-free power-law distribution was still preserved in the metabolic enzyme network, with only a small (less than one) difference between the exponents.

However, it was also found that “hubs” of highly connected nodes, such as pyruvate and coenzyme A, also existed in metabolic compound networks, in which these nodes were highly connected and interacted with many other nodes. It was proposed that these metabolic networks were actually arranged in a hierarchical manner (16), where highly connected modules would be connected to one another in a scale-free manner. These modules would in turn form clusters that would then be connected to other clusters at a higher level, and so on. Such properties can be taken advantage of to infer the function of the genes involved in each corresponding module at various levels of the hierarchy. This method would actually correlate well with the fact that networks of genes are not necessarily working alone, but function in concert with other proteins and complexes at higher levels. This finding in fact correlates surprisingly well with results published recently based on graph-theoretical analysis of gene regulatory networks in Bacillus Subtilis (2). That is, because only a subset of genes is actually active at any one time, the dynamic topology of gene regulatory networks was taken into consideration in this work, as opposed to the full static network. As a result, a hierarchical scale-free network emerged.
Local network features

Commonly occurring patterns in metabolic networks, or network motifs, which can be found using building blocks for finding frequent subgraphs, have shown promise of functional inference (17). Recently, however, critiques have been raised saying that such functional inferences must also take into consideration evolution (18). As such, work on extracting phylogenetic modules from metabolic enzyme networks demonstrated that such functional units are indeed conserved across evolution (19). In this work, phylogenetic profiles were constructed for all the enzymes in the metabolic reference map of KEGG. Using the Jaccard coefficient as a similarity measure, all enzymes were clustered hierarchically based on phylogenetic profiles. Then, edges between the enzymes were added based on the edges in the metabolic network. Finally, clusters were created within each cluster based on these new edges between enzymes. These small clusters were thus defined as phylogenetic network modules, in which enzymes that have similar phylogenetic profiles are close to one another in the metabolic network. In preliminary studies, the enzyme clusters were constructed using only the similarity between phylogenetic profiles differed from those that resulted from the final network modules that metabolic network connectivity, which indicates that phylogeny should indeed be incorporated in metabolic module analysis. These phylogenetic modules also demonstrated that this final network possessed hierarchical network features, such that hubs of important genes exist, but that these hubs are connected by more sparsely linked genes that work as linkers between these hubs to connect the entire network as a whole.

The concept that modules comprise the traditional pathways is gaining more focus as basic functional building blocks (20). Gene expression patterns in pathways and their formation of modules have been an intense topic of study (21, 22). These pathways combined with flux balance analysis have also provided interesting results about the metabolic pathway of yeast (23) and Escherichia coli (24). The latter involves steady-state analysis using reaction stoichiometry information, such as those stored in the KEGG REACTION database, and it is gaining renewed interest for systematic analysis of metabolic networks (6).

Functional Network Inference

In addition to the topological features of networks, other sources of information can and should be incorporated to take a step further into inferring function from the hierarchically organized modules of metabolic networks.

Metabolic reconstruction: genome to pathway mapping

The term metabolic reconstruction refers to the process of linking the genomic repertoire of enzyme genes to the chemical repertoire of metabolic pathways. That is, a metabolic enzyme pathway can be inferred given a set of enzymes (25). This task can be done by first referring to the existing pathway maps in which the involved genes are known. By using the genomic information of multiple (related) species and comparing them against these pathways, ortholog groups involved at specific nodes in the pathways can be identified. This method is the basis of the K.O system. Correspondingly, the entire metabolic pathway of an organism can be inferred given its genome. That is, the K.O system can be used to reconstruct a metabolic enzyme network by first referring to the genes known to be in a particular organism. Once the K.O groups in which these genes are involved are identified, the nodes in the metabolic pathway in which these genes participate can be reconstructed. Thus, new sets of genes can then be compared against the K.O groups to reconstruct the metabolic pathways in which the input genes may be involved.

Integration of heterogeneous datasets

Because the metabolic pathway is in fact a complex process of various degrees of interactions between biomolecules, the integration of the main components and their fundamental interactions are important for the extraction of the functional modules and the identification of their roles in the network. For example, information on cellular components and their interactions can be incorporated to reconstruct metabolic networks more accurately compared with genome annotation and/or sequence information alone (4). This involves the incorporation of data from multiple data sources, such as KEGG (for pathway and genomic data) and PSORTdb (for subcellular localization data). In addition, work has been done to integrate stoichiometric and bibliometric data for reconstructing the human metabolic network (3). To incorporate an even wider variety of biologic data for predicting missing enzymes in metabolic enzyme networks, kernel methods are used. A kernel is a mathematic function that can take as input a variety of data for a specific set of entities and transform it such that the input entities can be classified as distinctly as possible. This method consists of two steps: a training phase and a test phase. The training phase consists of using data for which the properties are known in advance. Then, the test phase can be used to assess the applicability of the properties to new input data sets.

In terms of a metabolic network inference that uses multiple sources of data, as an example, for a given set of genes, expression, genomic context, chemical, and phylogenetic information can be used to train a kernel function to infer a metabolic network. This task is done by developing the kernel function such that a score is obtained for every pair of genes. If this score exceeds a particular threshold, the corresponding genes are considered to be related, and an edge can be drawn between them to form the inferred network. The incorporation of chemical information in this work was attempted in two ways: preintegration and postintegration, to enforce chemical restraints in an indirect and direct manner, respectively. In the indirect manner, all input sources are compared and contrasted with the chemical restraints, whereas in the direct approach, the chemical restraints are applied after an initial network is obtained. This latter approach ensures that chemical compatibility is maintained. A as a result, several enzymes were identified to fill in the missing nodes of the metabolic enzyme network for yeast (26).
Compound scope

Metabolic pathways may also be analyzed from a chemical standpoint, and it has been surmised that the array of concentrations of relatively simple chemicals in pathways may provide information for biologic processing (27, 28). Thus, the study of the metabolic reactions that take place in the metabolic compound network comes naturally.

The idea of the “scope” of a chemical compound was defined recently to characterize metabolic compound networks systematically. This idea developed from the fact that the occurrence of a metabolic reaction generally requires the existence of other reactions that provide its substrates, which generates a series of metabolic reactions. In each step of the corresponding expansion process, those reactions whose substrates are available are incorporated (29). Thus, starting with one or more seed compounds, an expansion can result in a final network whose compounds define the scope of the seed. Using all the metabolic reactions in the reference pathways of the KEGG PATHWAY database, the scopes of all metabolic compounds were calculated, and it was found that large parts of cellular metabolism could be considered as the combined scope of simple building blocks. Analyses of various expansion processes revealed that the incorporation of key metabolites such as adenosine tri-phosphate and coenzyme A would increase the network complexity. It was also shown that the outcome of network expansion is in general very robust against the elimination of a single or few reactions, although the elimination of a key reaction would result in a dramatic reduction of scope sizes. As a result, it was hypothesized that the expansion process displays characteristics of the evolution of metabolism, in that the emergence of metabolic pathways over time could be estimated from the systematic analysis of metabolic compound networks (30).

From this work on compound scope, an interesting analysis of the effect of oxygen on metabolic networks and the evolution of life was made possible (31). Recent evidence suggested that the increasing importance of molecular oxygen to metabolic pathways eventually replaced the enzymatic reactions central to anaerobic metabolism in aerobic organisms (32). Thus, by comparing metabolic compound networks under anaerobic and anaerobic conditions, the effect of the presence or the absence of oxygen on the complexity of specific seed compounds could be determined. Based on the reference pathways for metabolism in KEGG, O_2 was found to be among the most used compounds, superseding even adenosine tri-phosphate. Their analyses revealed four subnetworks of increasing complexity, which form a hierarchy such that certain reactions allow transitions between the subnetworks at different levels. Among these four subnetworks, molecular oxygen was required for transition into the largest network. Furthermore, in another analysis of the enzyme distribution across different organisms, it was found that the distributions of enzymes that catalyze anaerobic networks were not necessarily consistent with the tree of life, which indicates that the adaptation to O_2 had occurred throughout the tree of life. These results were supported by data from geologic and molecular evolutionary analyses indicating that all three domains of life appeared by the time oxygen became widely available (33).

Pathway prediction using RDM patterns

The study of the chemical reactions involved in metabolic compound networks and their scopes can help to predict new pathways. In this case, the RDM patterns defined in KEGG can be used. In fact, an analysis of the RDM patterns in KEGG in the context of their frequency of appearance in the KEGG PATHWAY categories was performed. In particular, the more than 2000 RDM patterns that appear in the metabolic pathways of KEGG were analyzed. Because RDM patterns themselves do not indicate the direction of the reaction, when a reaction in the pathway was defined as reversible, two reactions were generated for the corresponding RDM pattern. The number of unique patterns was counted for each pathway category, and it was found that the reactions in the xenobiotics biodegradation pathways in particular were most distinct compared with the other categories of pathways. In fact, roughly 80% of the RDM patterns were unique to this category. Thus, an attempt was made to use RDM patterns to predict a biodegradation pathway of a new xenobiotic compound.

This task was done by comparing the new compound first against the KEGG COMPOUND database to retrieve a list of candidate compounds that are most similar to the query. The matched compounds are then queried against the RDM pattern library to retrieve a list of putative RDM patterns. In the third step, the query compound is transformed into new possible compounds based on the retrieved transformation patterns. These newly generated compounds are then used iteratively as a new query to repeat the prediction cycle until no new transformations can be found. This approach retrieved successfully the degradation pathway for 1,2,3,4-tetrachlorobenzene (34).

Similar research has attempted to gain insight into protein function prediction based on information hidden in the molecular structure of metabolites (35). Such work may eventually identify the relationship between metabolic structure and protein function, thus possibly improving techniques in the prediction of enzyme function and novel metabolic pathways (36).

Discussion

It may now be generally believed that both metabolic networks can be characterized as hierarchically organized networks of modules that have scale-free properties. Several methods for the analysis of metabolic pathways are actively being developed to understand these functional modules found among them. The results, of course depend greatly on the data being used to find them. It has been shown that the incorporation of genomic and phylogenetic information aids the identification of functionally important modules. In turn, these data can aid phylogenetic analysis and functional annotation of the biologic entities involved.

Work in chemical reaction characterization and analysis enables the prediction of missing enzymes and pathways. The concept of modules of compounds, or compound scopes, defines the extent to which a particular chemical compound plays a role in the metabolic compound network. This research aids
in evolutionary analysis of modules as the importance of specific compounds can be directly analyzed based on its scope and the effect it has on the scope sizes of other compounds in the network.

We have illustrated that metabolic enzyme networks and metabolic compound networks are in fact two sides of the same coin. The global analysis of metabolic networks using the line graph transformation illustrated this point nicely. Thus, it is natural to pursue the relationships between those modules found in the original metabolic enzyme network and in the line-graph-transformed metabolic compound network to ascertain their functions. Such integration of knowledge from various aspects is crucial to gain a true understanding of the biologic processes of life.

We note that metabolic systems are studied often in systems biology using dynamics analysis such as flux balance analysis and differential equations. However, discussion regarding system dynamics is beyond the scope of this current manuscript, and we refer the interested reader to the relevant literature (37, 38). This limitation, however, does not preclude these analyses from the standpoint of integrated systems analysis for understanding the metabolic pathway.

Systems analysis approaches such as those presented in this manuscript have illustrated the importance of systematic and integrated methods of analyzing metabolism. It can be expected that multidimensional data will continue to play an important role in such approaches. Because the consistency of such data will determine the accuracy of the predictions, a balance between the speed of computational techniques and the accuracy of manual curation must be maintained. As long as an over-dependence does not exist on either approach, systems approaches for the study of metabolism should prove to be fruitful.

References

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

Further Reading

Systems Approach to Metabolism
Xenobiotic Metabolism

Costas Ioannides, Molecular Toxicology Group, School of Biomedical and Molecular Sciences, University of Surrey, Guildford, Surrey, United Kingdom

doi: 10.1002/9780470048672.wecb628

Humans are exposed continuously and unavoidably to a myriad of potentially toxic chemicals that are inherently lipophilic and, consequently, very difficult to excrete. To effect their elimination, the human body has developed appropriate enzyme systems that can transform metabolically these chemicals to hydrophilic, readily excretable, metabolites. This biotransformation process occurs in two distinct phases, Phase I and Phase II, and involves several enzyme systems, the most important being the cytochromes P450. The expression of these enzyme systems is regulated genetically but can be modulated also other factors, such as exposure to chemicals that can either increase or impair activity. Paradoxically, the same xenobiotic-metabolizing enzyme systems also can convert biologically inactive chemicals to highly reactive intermediates that interact with vital cellular macromolecules and elicit various forms of toxicity. Thus, xenobiotic metabolism does not always lead to deactivation but can result also in metabolic activation with deleterious consequences.

It is unlikely that the human body could survive and thrive in the chemical environment it lives in if it were not endowed with effective means to protect itself from the adverse effects of the myriad of chemicals to which it is continuously and unavoidably exposed. Although scientists focus on anthropogenic chemicals, humans are exposed to even more naturally occurring chemicals, mostly phytochemicals that, like their man-made counterparts, have the potential to induce toxicity. Both anthropogenic and natural chemicals are referred to as xenobiotics (Gr. foreign to life). Humans are exposed to huge numbers of xenobiotics, in the order of 5000 to 10,000 per day, most of which emanate from the diet. Of the chemicals that humans ingest, 99.9% are natural, largely of plant origin (1).

Biologic Background

Most chemicals to which humans are exposed cannot be exploited by the body either to generate energy or to use as structural blocks to build new tissue; they cannot be used as essential cofactors for enzyme reactions nor as chemical messengers. During the last three decades, however, it has become evident that some dietary phytochemicals possess biologic activity and have the potential to afford protection against major degenerative diseases of high morbidity, such as cancer and cardiovascular disease; this protection may explain the epidemiological findings that populations that consume diets with high vegetable and fruit content are less susceptible to these fatal diseases. Because most of xenobiotics offers no benefit to the human body, its immediate response is to prevent exposure and/or eliminate them. For this objective to be achieved, the body has developed as a first line of defense several transporter systems, such as the P-glycoprotein, which prevent the absorption of chemicals through the gastrointestinal tract by facilitating their efflux from the enterocytes into the intestinal lumen (2). For chemicals that at least partly overcome this obstacle and reach the blood circulation, its immediate response is to excr...
Xenobiotic Metabolism

Phase I and Phase II metabolic pathways. PAPS, adenosine 3′-phosphate 5′-phosphosulphate; UDPGA, uridine diphosphate glucuronic acid.

Table 1 Examples of prodrugs

<table>
<thead>
<tr>
<th>Parent drug</th>
<th>Pharmacological activity</th>
<th>Active metabolite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prontosil</td>
<td>Antibacterial</td>
<td>Sulphanilamide</td>
</tr>
<tr>
<td>Levodopa</td>
<td>Antiparkinson</td>
<td>Dopamine</td>
</tr>
<tr>
<td>Sulindac</td>
<td>Anti-inflammatory</td>
<td>Sulindac sulphide</td>
</tr>
<tr>
<td>Cyclophosphamide</td>
<td>Anti-cancer</td>
<td>4-Hydroxycyclophosphamide</td>
</tr>
<tr>
<td>Terfenadine</td>
<td>Antihistaminic</td>
<td>Fexofenadine</td>
</tr>
<tr>
<td>Codeine</td>
<td>Analgesic</td>
<td>Morphine</td>
</tr>
<tr>
<td>Sulfasalazine</td>
<td>Anti-inflammatory</td>
<td>5-Aminosalicylic acid</td>
</tr>
<tr>
<td>Tamoxifen</td>
<td>Anti-cancer</td>
<td>Endoxifen</td>
</tr>
<tr>
<td>Enalapril</td>
<td>Anti-hypertensive</td>
<td>Enalaprilat</td>
</tr>
<tr>
<td>Glyceryl trinitrate</td>
<td>Anti-inflammatory</td>
<td>Nitric oxide</td>
</tr>
</tbody>
</table>

Phase I metabolism

During Phase I metabolism, also referred to as functionalization, the xenobiotic acquires a functional group such as –OH, –COOH, and –NH₂; alternatively, such functional groups may be unmasked, for example, an alkoxy (–C–OR) group can be dealkylated to unmask a functional group (–C–OH). For example, 7-hydroxycoumarin can be formed either by the hydroxylation of coumarin (insertion of an oxygen atom) or by the deethylation of 7-ethoxycoumarin (unmasking) (Fig. 1). Phase I metabolism involves oxidation, reduction, and hydrolysis reactions and is catalyzed by various enzyme systems (Table 2). The predominant reactions are oxidations, and the most important system catalyzing these is the cytochromes P450 (vide infra). Reductions are catalyzed by reductases, which are not very active in mammalian cells but are very active in gut bacteria that, consequently, may contribute extensively to the metabolism of orally administered drugs and other xenobiotics; the most common reductions are azo- and nitroreductions. The enzymes catalyzing hydrolysis reactions are esterases.
Xenobiotic Metabolism

and amidases that hydrolyze esters and amides, respectively (Table 2).

Phase II metabolism

During Phase II metabolism, also referred to as conjugation, the metabolites generated during Phase I metabolism combine with endogenous substrates, such as sulphate, glucuronic acid, glutathione, and amino acids, to form highly hydrophilic metabolites that are excreted with ease in the urine and feces. As an example, 7-hydroxycoumarin, formed from Phase I metabolism, forms a sulphate and a glucuronide conjugate (Fig. 1). Sulphation is catalyzed by the sulphotransferases and glucuronidation by the glucuronosyl transferases that attach glucuronic acid to the substrate. Chemicals, such as the drug paracetamol (acetaminophen) that already possesses a functional group, may bypass Phase I metabolism and be metabolized predominantly through conjugation (Fig. 2). Conjugation with glutathione represents an effective cellular defense mechanism that neutralizes toxic chemicals that otherwise would cause toxicity. Another Phase II pathway, usually minor, is conjugation with amino acids, the most common being glycine. Finally, methylation and acetylation are unusual pathways in that the generated metabolites are less polar than the parent compound; in this case, metabolism hinders rather than facilitates excretion and additional metabolism is required for eventual elimination.

Enzyme Systems that Catalyze Phase I Metabolism

Although several enzymes contribute to the Phase I metabolism of xenobiotics (Table 2), by far the most prominent are the cytochromes P450, so called because in the reduced state they form a complex with carbon monoxide that is characterized by an absorption maximum at 450 nm.

Cytochromes P450

A widespread enzyme system in nature, cytochromes P450 are found in both prokaryotic and eukaryotic cells and, with the exception of striated muscle and red blood cells, are encountered in every tissue but predominately in the liver, which, as a result, is the principal site of xenobiotic transformations. However, cytochrome P450 enzymes also are active in portals of xenobiotic entry, such as the lungs, the gastrointestinal tract, and the nasal mucosa. Although the cytochrome P450 enzyme system can function as a reductase under anaerobic conditions, its main role is to facilitate the oxidation of a myriad of structurally diverse xenobiotics. It is a haem-containing, membrane-bound enzyme that requires molecular oxygen and NADPH and inserts an atom of oxygen to the xenobiotic (X) while the other is reduced to water.

NADPH + H+ + XH → XOH + NADP+ + H2O

The flavoprotein cytochrome P450 reductase channels the electrons from NADPH to the cytochromes P450, which function as a terminal oxidase (Fig. 3).

Figure 2 Metabolism of paracetamol (acetaminophen).

A principal attribute of the cytochrome P450 system is the unprecedented broad substrate specificity it displays, which explains its pivotal role in xenobiotic metabolism. It catalysts efficiently the metabolism of thousands of structurally diverse chemicals with markedly different molecular shape and size. It achieves this very broad substrate specificity by existing as a "superfamily" of enzymes. Each family is subdivided further into subfamilies that may contain one or more enzymes. Enzymes that share a structural similarity of at least 40% belong to the same family, which is indicated by Arabic numbers, whereas, if the structural similarity exceeds 55%, then they are classified within the same subfamily, which is denoted by capital letters; finally, enzymes belonging to the same subfamily are denoted by Arabic numbers. For example, the CYP family comprises two subfamilies, namely CYP1A and CYP1B; the former consists of two enzymes, CYP1A1 and CYP1A2, whereas within the
Table 2 Enzyme systems catalyzing Phase I and II xenobiotic metabolism

<table>
<thead>
<tr>
<th>Phase I Enzyme system</th>
<th>Principal cellular localization</th>
<th>Phase II Enzyme system</th>
<th>Principal cellular localization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochromes P450</td>
<td>Endoplasmic reticulum</td>
<td>Glucuronosyl transferases</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>Flavin monooxygenases</td>
<td>Endoplasmic reticulum</td>
<td>Sulphotransferases</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Amines oxidases</td>
<td>Mitochondria</td>
<td>Glutathione S-transferases</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Molybdenum hydroxylases</td>
<td>Cytosol</td>
<td>A-cetyl transferases</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Prostaglandin synthases*</td>
<td>Endoplasmic reticulum</td>
<td>Methyl transferases</td>
<td>Cytosol</td>
</tr>
<tr>
<td>Lipoxigenases</td>
<td>Cytosol</td>
<td>Epoxide hydrolases</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>Alcohol/Aldehyde dehydrogenases</td>
<td>Cytosol</td>
<td>Amino acid conjugases</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>Esterases/amidases</td>
<td>Cytosol</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*also referred to as cyclooxygenases

Figure 3 Cytochrome P450 electron transport system. S, Substrate.

The fact that some xenobiotic-metabolizing cytochrome P450 isofoms are expressed polymorphically (vide infra) and that individuals may lack totally a certain isoform without any detriment to health indicates that the role of these cytochromes P450 in endogenous metabolism is far from life threatening. Cytochromes P450 are believed to have evolved from a common ancestor some three thousand million years ago. The function of the earliest forms is thought to have been in the metabolism of essential endogenous chemicals, such as steroid hormones, and then to have evolved to enzymes capable of metabolizing foreign compounds. As a result, their initial narrow substrate specificity toward steroids was lost and, to cope with the new, increasingly chemical environment, they developed into broad-specificity enzymes that could metabolize the diverse chemicals to which they now were exposed. It has been proposed that what forced the evolution of these proteins was the necessity to develop defense mechanisms to protect against plant toxins present in the food chain that were produced to discourage predators (5).
## Enzyme Systems that Catalyze Phase II Metabolism

Conjugation reactions of xenobiotics or their metabolites with endogenous substrates produce highly hydrophilic metabolites; however, to a lesser extent functional groups may be also methylated or acetylated to produce less hydrophilic compounds.

### Glucuronide conjugation

Glucuronide conjugation is a frequently used route of conjugation by xenobiotics where glucuronic acid, in its active form, uridine diphosphate glucuronic acid (UDPGA), is added to the molecule (Fig. 1). Some drugs like oxazepam and morphine are catalyzed predominantly by glucuronidation because they do not require prior Phase I metabolism to generate a functional accepting group. The most readily conjugated functional groups are phenols and alcohols, which yield ester glucuronides, and carboxylic acids, which form ether glucuronides.

### Sulphate conjugation

Conjugation with sulphate, catalyzed by cytosolic sulphotransferases, is also a major route of Phase II metabolism, where inorganic sulphate, made available in the activated form of 3′-phosphoadenosine-5′-phosphosulphate (PAPS), is added to the molecule (Fig. 1). This is the most important pathway in the metabolism of phenols and is a very efficient conjugating system as long as inorganic sulphate is available.

### Glutathione conjugation

Glutathione conjugation is an important pathway of metabolism that allows the cell to defend itself from chemical insult. It uses the nucleophilic tripeptide glutathione, which possesses a nucleophilic sulphur atom, to detoxify chemically reactive metabolites, preventing them from interacting with critical cellular macromolecules with adverse consequences and, thus, preserves cellular integrity. This reaction is catalyzed by the ubiquitously distributed glutathione S-transferases, which transfer a molecule of reduced glutathione to the toxic chemical that results in its neutralization. Glutathione conjugates additionally are processed further metabolically in the body are excreted and result in its neutralization. Glutathione conjugates additionally are processed further metabolically in the body and, usually as N-acetylcysteine conjugates (mercapturates).

### Amino acid conjugation

In this minor route of xenobiotic metabolism, the carboxylic group of organic acids may conjugate with amino acids, glycine being the most common. The carboxylic group of the xenobiotic forms a peptide bond with the α-amino-group of the amino acid. Initially, in an ATP-dependent reaction, the carboxylic group reacts with CoA to form an α-acyl-CoA thiolster derivative, which then interacts with the amino acid to form the conjugate.

### Hydration

Hydration involves the addition of water to epoxides to form dihydrodiols and is catalyzed by microsomal epoxide hydrolase that displays, broad substrate specificity. As epoxides are generally toxic entities, this is a very important route for their detoxification.

### Methylation

Methylation, as well as amino and thiol groups, may be metabolized through methylation, the methyl donor being S-adenosyl methionine, the product of the interaction of ATP with methionine. Usually, it is a minor metabolic route in xenobiotic metabolism, but it plays a major role in the metabolism of endogenous substrates such as noradrenaline (norepinephrine). Methylation is catalyzed by methyltransferases located in the mitochondria.

### Acetylation

Acetylation is an important metabolic route for aromatic and heterocyclic amines, hydrazines and sulfonamides. An amide bond is formed between the amino group of the chemical and the acetate. This reaction is catalyzed by acetyltransferases, the acetyl group being donated by acetyl-CoA.

### The First-Pass Effect

If an orally taken drug is not absorbed, then there will be no pharmacological effect and the drug will be excrated in the feces. However, it is conceivable that a drug is very well absorbed and yet fails to manifest the expected pharmacological effect. Following oral administration, drugs are absorbed through the intestine into the portal circulation that takes them to the liver and then to the systemic circulation. A drug may undergo metabolism either in the intestine and/or through its first passage through the liver to such an extent that very little remains available for distribution to the other tissues, including the site of action, and, consequently, biologic activity is either not manifested or attenuated. In the intestine, metabolism may be catalyzed not only by intestinal enzymes, such as cytochromes P450 and Phase II conjugation enzymes, but also by microorganisms. This phenomenon is known as the first-pass effect or presystemic metabolism. Thus, poor response to drug treatment after oral intake may be due to the fact that, despite complete absorption, only a small fraction of the drug reaches systemic circulation intact because of extensive metabolism in the intestine or liver.

First-pass metabolism, therefore, will decrease the pharmacological effect of the drug, and if the metabolism is extensive, the pharmacological effect may be abolished completely as a result. Drugs that are subject to first-pass metabolism are administered through alternative routes or at higher doses to compensate for the loss during presystemic metabolism. Glyceryl trinitrate, a drug used in the treatment of heart angina, when taken orally undergoes >99% of first-pass metabolism being denitrated in the liver, and, as a result, it is never administered through this
route but is taken sublingually to bypass the intestine. The patient places a tablet under the tongue, and because of the good network of blood vessels, the drug is absorbed very rapidly and the patient benefits from it within a few minutes.

**Factors that Influence Xenobiotic Metabolism**

Several factors can modulate the activity of xenobiotic-metabolizing enzymes, such as age, the nature of diet, and the presence of disease, but the most important appear to be genetic makeup and concurrent or prior exposure to chemicals (enzyme induction and enzyme inhibition).

**Polymeric expression of xenobiotic-metabolizing enzyme systems: clinical implications**

The presence of xenobiotic-metabolizing enzymes and this level of expression are governed by our genes. Xenobiotic-metabolizing cytochromes P450 may be polymorphically expressed, and if this is not appreciated and the necessary steps taken to adjust drug dosage accordingly, it may have a dramatic impact in clinical therapeutics (5, 6). It is recognized now that xenobiotic-metabolizing enzymes, both in Phase I and Phase II, are polymorphically expressed, resulting in inter individual differences in metabolic capacity, so that drug dose regimens are unlikely to be optimal for all patients. For a certain drug, some individuals are poor metabolizers, whereas others are extensive metabolizers. Furthermore, polymorphic expression affects not only drug efficacy but also the appearance of adverse effects. The etiology of adverse effects experienced by some patients exposed to the therapeutic doses of drugs, especially of drugs with a narrow therapeutic index, i.e., drugs whose plasma concentrations must be maintained within a narrow range to achieve the desired pharmacological effect, may be attributed to their individual enzyme profile (7).

One of the first polymorphisms to be identified involved acetylation (N-acetyltransferase) and the antitubercular drug isoniazid. Several people are slow acetylators, the proportion being race related and varying from about 10% in the Japanese to 70% in Caucasians. Acetylation of isoniazid results in loss of pharmacological activity because acetylisoniazid is devoid of antibacterial activity, but it provokes toxicity that occurs more frequently in slow acetylators. Since acetylation is not catalyzed efficiently, isoniazid blood levels are high and the drug interacts with pyridoxal 5-phosphate, the active form of vitamin B₆, resulting in its depletion; this vitamin deficiency causes neuropathy that leads to seizures. This situation can be avoided by the concurrent administration of the vitamin to patients treated with isoniazid. Genetic polymorphism in other conjugating systems that result in metabolic deficiencies also can lead to predisposition to the toxicity of chemicals that rely heavily on these enzymes for their deactivation. Gilbert’s syndrome is a condition where the patient experiences intermittent jaundice because of reduced capacity in eliminating bilirubin through glucuronide conjugation. The produg irinotecan, used in the treatment of advanced colorectal cancer, provokes severe gastrointestinal toxicity in these patients because of suppressed glucuronidation (8, 9). The active metabolite (SN-38) that is generated metabolically cannot be detoxicated by glucuronidation and consequently accumulates. Another early example of polymorphism involves the drug succinylcholine (suxamethonium), a muscle relaxant used primarily during surgery. Its action lasts only a few minutes because it is very efficiently metabolized by cholinesterases present in the liver and plasma. A few people, about 1 in 3000 who genetically lack this enzyme, develop sustained apnoea as a result of paralysis because its effect is prolonged from 30 minutes to hours.

Genetic polymorphism in cytochromes P450 is believed to be responsible for many adverse effects associated with drug intake. The first cytochrome P450 protein to be recognized as being polymorphically expressed was CYP2D6, an enzyme that catalyzes many current psychoactive drugs, such as tricyclic antidepressants; subsequent studies have revealed that polymorphism may involve also the CYP2C9 and CYP2C19 enzymes. Persons lacking an active gene, i.e., a gene that can generate a functional protein, when exposed to drugs relying on this enzyme for their metabolic deactivation will show exaggerated adverse effects as a consequence of accumulation, and ideally should be prescribed lower doses. One such drug is the antihypertensive debrisoquine, which in normal individuals undergoes extensive CYP2D6-mediated 4-hydroxylation that leads to loss of activity. About 10% of European Caucasians and 1% of Japanese have been identified as poor metabolizers because of the lack of a functional CYP2D6. Subjects with poor metabolism inherited two copies of a gene that encodes either an enzyme with low activity or one with no activity. When exposed to dose regimens of debrisoquine developed for normal metabolizers, as a consequence of diminished metabolism, the drug accumulates on repeated administration and adverse effects commensurate with overdose are experienced, such as sustained drop in blood pressure. A number of individuals are classified as intermediate metabolizers, having one copy of the inactive gene; they display CYP2D6 activity but at low level. As a result of CYP2D6 polymorphic expression, the dose required to produce the same plasma levels may differ by as much as 20-fold among individuals. Similarly, the β-blocking agent timolol is metabolized largely by CYP2D6; when exposed to this drug, poor metabolizers experience a prolonged, more intense pharmacological effect (10). If one is dealing with a produg, however, where metabolism leads to the production of the pharmacologically active form, a less intense or a complete loss of the pharmacological effect may occur in poor metabolizers. The analogic effect of codeine is associated with morphine, which is formed because of CYP2D6-mediated metabolism. In poor metabolizers, the drug displays poor efficacy, and the lack of analgesia is due to the active metabolite is not being produced; indeed, morphine plasma levels are extremely low (11, 12).

Polymorphism also can enhance the metabolism of a drug in persons carrying multiple copies of a cytochrome P450 gene, for example, as a result of gene amplification and duplication, so that a drug may be deactivated more rapidly through
metabolism, leading to a less intense or totally absent pharmacological effect because therapeutic plasma levels are not achieved. Such individuals are known as ultrarapid metabolizers and require far higher doses than normal to achieve therapeutic response. However, in the case of intake of a prodrug, the dose should be decreased to prevent an exaggerated effect and toxicity. Ultrarapid metabolizers show symptoms of codeine intoxication, such as severe abdominal pain, when they take a normal therapeutic dose (13).

**Cytochrome P450 regulation by chemicals and clinical consequences**

The levels of cytochrome P450 are not only genetically determined but may be modulated also by factors such as age, nutritional status, and the presence of disease. The most important factor, however, is previous exposure to xenobiotics, either natural or synthetic, that either can induce, largely as a result of enhanced enzyme synthesis, or impair cytochrome P450 activity; both can lead to serious consequences during multiple drug intake. Such drug interactions involve in particular, CYP3A4, as this enzyme is the most active cytochrome P450 in drug metabolism and, moreover, is the dominant form in the liver and intestine, the principal site and the first site of metabolism following oral drug intake, respectively.

Cyclosporin is an immunosuppressant drug taken chronically by organ transplant patients. It has a narrow therapeutic index, and plasma levels have to be maintained within a narrow range to ensure efficacy and to avoid the appearance of serious adverse effects that can be life threatening. The principal catalyst of its metabolism is CYP3A4, and even small changes in the activity of this enzyme can have major impact on its efficacy. The antifungal drug ketoconazole is a potent inhibitor of CYP3A4, and, as a result, it increases plasma levels of cyclosporin leading to toxicity when coadministered with it (14). Similarly, serious interactions can occur with prodrugs that necessitate CYP3A4-metabolism to generate the pharmacologically active form. Terfenadine underdose nearly complete first-pass effect in the liver; it is metabolized by CYP3A4 to generate a metabolite, fexofenadine, which possesses antihistaminic activity. Normally, terfenadine is never seen in the blood of patients because of the extensive first-pass metabolism. When, however, it is taken together with ketoconazole, its metabolism is impaired and terfenadine can escape into the systemic circulation causing cardiotoxicity (14). To avoid such complications, the active metabolite fexofenadine has replaced the original formulation that necessitate CYP3A4-metabolism to generate the pharmacologically active form.

Cyclosporin is an immunosuppressant drug taken chronically by organ transplant patients. It has a narrow therapeutic index, and plasma levels have to be maintained within a narrow range to ensure efficacy and to avoid the appearance of serious adverse effects that can be life threatening. The principal catalyst of its metabolism is CYP3A4, and even small changes in the activity of this enzyme can have major impact on its efficacy. The antifungal drug ketoconazole is a potent inhibitor of CYP3A4, and, as a result, it increases plasma levels of cyclosporin leading to toxicity when coadministered with it (14). Similarly, serious interactions can occur with prodrugs that necessitate CYP3A4-metabolism to generate the pharmacologically active form. Terfenadine underdose nearly complete first-pass effect in the liver; it is metabolized by CYP3A4 to generate a metabolite, fexofenadine, which possesses antihistaminic activity. Normally, terfenadine is never seen in the blood of patients because of the extensive first-pass metabolism. When, however, it is taken together with ketoconazole, its metabolism is impaired and terfenadine can escape into the systemic circulation causing cardiotoxicity (14). To avoid such complications, the active metabolite fexofenadine has replaced the original formulation that necessitate CYP3A4-metabolism to generate the pharmacologically active form.

Cyclosporin is an immunosuppressant drug taken chronically by organ transplant patients. It has a narrow therapeutic index, and plasma levels have to be maintained within a narrow range to ensure efficacy and to avoid the appearance of serious adverse effects that can be life threatening. The principal catalyst of its metabolism is CYP3A4, and even small changes in the activity of this enzyme can have major impact on its efficacy. The antifungal drug ketoconazole is a potent inhibitor of CYP3A4, and, as a result, it increases plasma levels of cyclosporin leading to toxicity when coadministered with it (14). Similarly, serious interactions can occur with prodrugs that necessitate CYP3A4-metabolism to generate the pharmacologically active form. Terfenadine underdose nearly complete first-pass effect in the liver; it is metabolized by CYP3A4 to generate a metabolite, fexofenadine, which possesses antihistaminic activity. Normally, terfenadine is never seen in the blood of patients because of the extensive first-pass metabolism. When, however, it is taken together with ketoconazole, its metabolism is impaired and terfenadine can escape into the systemic circulation causing cardiotoxicity (14). To avoid such complications, the active metabolite fexofenadine has replaced the original formulation that necessitate CYP3A4-metabolism to generate the pharmacologically active form.
be critical. Pulmonary cytochrome P450 activity is more likely to be increased when exposure to the inducing agent occurs by inhalation; smokers, for example, display higher pulmonary cytochrome P450 activity (CYP1) compared with nonsmokers, which is ascribed to the inhalation of polycyclic aromatic hydrocarbons.

**Bioactivation of Xenobiotics**

A striking paradox that was first recognized in the 1970s was that the xenobiotic-metabolizing enzyme systems, long regarded as being exclusively involved in the deactivation and elimination of chemicals, such as the cytochromes P450, could in fact assume the reverse role, i.e., they could convert innocuous, chemically inert xenobiotics to highly reactive and toxic metabolites with deleterious consequences to the body. This process is referred to frequently as “metabolic activation” or “bioactivation.” Although the chemicals that undergo activation are structurally diverse, the basic mechanism of activation appears to be quite similar. Activation involves primarily oxygenation, although reduction is also important with some chemicals (Fig. 5). The reactive intermediates that are formed are electrophiles, having electron-deficient atoms, and so are chemically very highly reactive. Since these reactive species are generated intracellularly, they can interact readily and irreversibly with vital cellular macromolecules, such as DNA, RNA, and proteins, to provoke various types of toxicity; thus, in this case, metabolism confers to the chemical adverse biologic activity. Alternatively, these reactive intermediates may interact with tissue oxygen, giving rise to toxicity indirectly through redox cycling, by acting as radical generators. These free radicals are capable of inducing cellular damage similar to that resulting from the covalent binding of electrophiles to cellular constituents. Simultaneously, a chemical will be subject to metabolism through pathways that lead to deactivation so that the extent of toxicity will be dependent on the balance of activation and deactivation. An extensively studied example of such drug bioactivation involves the drug paracetamol (acetaminophen), which is primarily metabolized through conjugation with sulphate and glucuronic acid leading to deactivation. However, to a very small extent, it undergoes cytochrome P450-dependent oxidation to form a reactive quinoneimine that has the potential to cause hepatotoxicity following covalent interaction with proteins (Fig. 2). The low levels of the quinoneimine are, however, effectively neutralized by conjugation with glutathione. As a result, paracetamol is a very safe drug at therapeutic doses, despite the formation of a cytotoxic metabolite, and only becomes unsafe when this fine balance of activation/deactivation is disturbed. Some groups of people, such as the chronic alcoholics, are vulnerable to the toxicity of paracetamol. Alcohol is the prototype inducer of CYP2E1, one of the cytochrome P450 enzymes that catalyzes the oxidation of paracetamol to the quinoneimine. Consequently, alcoholics, because of the higher CYP2E1 activity, form more of the toxic intermediate of paracetamol and, therefore, are sensitive to the hepatotoxicity of this drug. Thus, in this case, toxicity ensues because of increased activation that leads to the enhanced production of the toxic metabolite.

Most chemical carcinogens also rely on bioactivation to genotoxic metabolites that readily interact with DNA and set into motion the processes that eventually lead to tumorigenesis. The metabolic pathways once again are catalyzed by both
Phase I and Phase II enzymes, and the chemical is concurrently subject to deactivation pathways. Although cytochromes P450 frequently catalyze the first step in the bioactivation of most chemicals, other enzyme systems, such as the sulphotransferases and acetylases, also are essential in the generation of the ultimate toxic species, the entity that interacts with the cellular macromolecules. For example, the heterocyclic amine 2-amino-3-methylimidazo-(4,5-f)quinoline (IQ), a carcinogen formed during the grilling or frying of meat and fish, requires bioactivation to express its carcinogenicity. The activation pathway involves CYP1A2-catalyzed N-hydroxylation, followed by esterification of the hydroxylamine with sulphate and acetate to generate the sulphatoxy and acetoxy esters, respectively, that break down spontaneously to yield the nitrenium ion, the presumed ultimate carcinogen (Fig. 6). Ring hydroxylation at the 5-position and direct conjugation of the parent compound with glucuronide or sulphate are strictly deactivation pathways; the ring-hydroxylated metabolite of IQ is excreted eventually in conjugated form.

**Figure 6** Metabolic activation and deactivation of IQ. IQ, 2-amino-3-methylimidazo-(4,5-f)quinoline.
Conclusions

The human body is equipped with an array of enzyme systems that enable it to transform the myriad of chemicals to which it is inevitably exposed to metabolites that are readily and efficiently eliminated. In this way, the residence time of chemicals in the body is minimized and their accumulation prevented. Any factor that modulates the enzymes involved in the metabolism of a certain chemical will influence also its biologic activity, including toxicity. Obviously, the amount of reactive intermediate produced, and hence incidence but is determined by the nature of the enzymes active in the body at the time of exposure. A chemical is subject to several metabolic pathways, the majority of which will bring about its activation and facilitate its excretion. However, some routes of metabolism will transform the chemical to a metabolite capable of inducing toxicity and carcinogenicity. Obviously, the amount of reactive intermediate produced, and hence incidence of toxicity, will be largely dependent on the competing pathways of activation and deactivation, and whatever metabolic pathways, the majority of which will bring about its activation and facilitate its excretion. However, some routes of metabolism will transform the chemical to a metabolite capable of inducing toxicity and carcinogenicity. Obviously, the amount of reactive intermediate produced, and hence incidence of toxicity, will be largely dependent on the competing pathways of activation and deactivation, and whatever factor influences this delicate balance of activation/deactivation of a chemical will impact also on its biologic activity and safety.

References

6. Ingelman-Sundberg M, Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics 2005;6:5–13.

Further Reading


See Also
ADMET Properties of Drugs
Cytochrome P450 Monooxygenases, Chemistry of
Drug Metabolizing Enzymes, Chemistry of
Enzyme Inhibition, Chemistry and Mechanisms of
Polymorphisms, Detection of
Cytochrome P450 monooxygenases represent a large superfamily of heme enzymes, which require a dioxygen molecule and two electrons supplied by a NAD(P)H-dependent protein redox partner to form catalytically active high-valent ferryl-oxo intermediate. These heme enzymes are found in most organisms from all biological kingdoms and catalyze various oxidative chemical transformations important for the biosynthesis of steroid hormones, vitamins, signaling molecules, and many other compounds, as well as for the metabolism of xenobiotics.

Cytochromes P450 represent a superfamily of heme enzymes with a common fold and mechanism that requires an atmospheric dioxygen molecule, two electrons from NAD(P)H, and one or two protons available from solvent water to form an active catalytic ferryl-oxo intermediate (1). All cytochromes P450 have one heme b (protoporphyrin IX) buried deeply in the protein globule, with molecular masses in the range 45–60 kDa. In the resting state, heme iron is oxidized (Fe$^{3+}$) and is coordinated with the thiolate sulfur atom of cysteine as the fifth axial ligand with the sixth position vacant or occupied by water. The iron serves as a sole redox center in the catalytic cycle, which accepts two electrons sequentially and channels them to coordinated dioxygen to form a (hydro)peroxo-ferric intermediate. Second protonation of the distal oxygen atom and transient formation of the iron coordinated oxo-water, Fe–OOH$_2$, results in the heterolytic scission of O–O bond and formation of the ferryl oxo porphyrin-cation radical, which is termed Compound I, by analogy to the same active intermediate first documented in peroxidases and catalases (2).

Reactions Catalyzed By Cytochromes P450

Although they are termed "monooxygenases," cytochromes P450 catalyze multiple chemical transformations that include hydroxylation, demethylation, epoxidation, C–C bond cleavage, deamination, and many others (3, 4). Most reactions performed by cytochromes P450 can be catalyzed effectively by Compound I, which is formally the same high-valent intermediate in all isozymes (5, 6). Substrate binding at the distal site of the heme brings the target chemical group into the close proximity of the reactive iron-oxygen catalytic intermediate, often in the regiospecific and stereospecific mode, to provide the selectivity of the chemical transformation. Thus, the broad range of metabolized substrates and the unprecedented variability of the chemical reactions catalyzed by cytochromes P450 can be attributed largely to the differences in substrate binding properties of different isozymes.

With respect to their functional roles in living organisms, cytochromes P450 can be separated tentatively into the enzymes essential for biosynthetic pathways and the xenobiotic metabolizing enzymes (1). The former group includes cytochromes P450 involved into the biosynthesis of steroid hormones, antibiotics, and the transformation of vitamins and cofactors such as retinoids, carotenoids, eicosanoids, and fatty acid derivatives as well as those involved in endogenous intracellular and extracellular signaling. Most reactions are catalyzed with narrow substrate specificity, and they afford products with high regioslectivity and stereoselectivity. As a result, some cytochromes P450 from this group are critical for normal life cycle, and deficiencies can lead to serious diseases (7). In contrast, most cytochromes P450 that belong to the latter group can metabolize a variety of substrates of different sizes through multiple chemical mechanisms and even can form several different products from one substrate. Some enzymes may also bind simultaneously two or three substrate molecules, same or different, and various cases of the resulting activation or inhibition of the product turnover were documented both in vivo and in vitro (4, 7). Cytochromes P450 from both groups are represented in all biological kingdoms; although in many cases, the information about their biological functions and underlying chemical mechanisms is incomplete and fragmented at best.

Despite the unprecedented variability of chemical reactions catalyzed by cytochromes P450, all the important features of their mechanism can be attributed to the oxidative transformations driven by the main catalytic intermediate, which is a...
Cytochrome P450 Mono-oxygenases. Chemistry of ferryl-oxo heme complex with the porphyrin pi-cation radical, or Compound I (3, 4, 8). Some important exceptions involve the hydroperoxoferric complex, Compound 0, and perhaps the one-electron reduced Cpd I, ferryl oxo heme complex Compound II (1, 3, 4, 9). Nonclassic reactions catalyzed by some cytochromes P450, which do not use oxygen, and thus deviate from the main P450 mechanism, include direct reductive processes driven by the heme iron (conversion of nitric oxide to nitrous oxide by CYP55, reductive dehalogenation, reduction of azo compounds and quinines, etc.). Another important role of cytochrome P450 activity is that they can react with peroxides with no consumption of external redox equivalents and perform peroxidase-type chemistry with the O–O bond scission, or afford rearrangements of fatty acid hydroperoxides.

Hydroxylation of hydrocarbons

Oxygen atom from Cpd I is inserted into the C–H bond of saturated hydrocarbons (Scheme 1a) by means of hydrogen atom abstraction followed by recombination of the transient hydroxyl with the carbon radical [the so-called “oxygen rebound” mechanism proposed by Groves in 1976 (8, 10)]. Another possibility can be the concerted oxygen insertion into the C–H bond. Both pathways are rationalized by the two-state mechanism developed by Shaik et al. (6, 9), which describes different reactivities of Cpd I in the high-spin and low-spin states. The possibility of involvement of Cpd 0, which is the ferric-hydroperoxo intermediate, into the carbon hydroxylation is also debated in the literature, but this species is much weaker oxidant and may participate in catalysis only in rare cases when the barrier for hydroxylation by Cpd 0 is lower than the barrier for the Cpd I formation. Carbon hydroxylation is one of the most common reactions for all cytochromes P450, with substrates that include steroid hormones, fatty acids, vitamins, and antibiotics (5, 11).

Oxidation and dealkylation of heteroatoms

Oxidative transformation of the heteroatom attached to the carbon (Scheme 1b) may result in direct electron abstraction from the former and formation of oxide (-NO, -SO, -PO, etc.) or may proceed through the hydrogen abstraction from the latter with subsequent hydroxylation as in the previous section. The second pathway is often followed by the heteroatom elimination and formation of carbonyl group. Substrates that contain nitrogen (i.e., amines and amides) are easier to oxidize and give more various products in P450 catalyzed reactions than compounds that contain oxygen. An example is the reaction of N-demethylation of caffeine catalyzed by the human CYP1A2.

Olefin oxidation (epoxidation)

Insertion of oxygen atom from Cpd I into the carbon-carbon double bond with formation of epoxide (Scheme 1c) reveals features characteristic for a concerted process, although formation of radical intermediates is possible in many cases. A unified description of this alternative is also provided by the two-state mechanism of catalysis by Cpd I (see the section on Hydroxylation of hydrocarbons). Essentially, the concerted oxygen insertion represents a low-spin reaction surface, whereas the distinct radical intermediate is formed on the high-spin reaction pathway. In the latter case, the carbon radical may attack the nearby heme nitrogen and modify the heme covalently. This reaction is also an important inactivation pathway of cytochromes P450 during oxidative transformations of terminal double and triple bonds.

Aromatic oxidation and hydroxylation

Aromatic compounds may be oxidized either through epoxidation or via addition of oxygen atom from Cpd I (Scheme 1d). Usually, both pathways afford more stable phenols or quinones as the end product through rearrangements and/or addition of another nucleophile. Direct abstraction of an electron from aromatic moiety is viable in the presence of strong electron-donating substituents. The oxidative metabolism of polycyclic aromatic compounds is represented by CYP1A1 in humans.

Dehydrogenation and desaturation

Desaturation, or formation of double bonds after hydrogen abstraction (Scheme 1e), is an important class of transformations of steroids, flavones, and drugs with heterocyclic compounds. These reactions do not involve insertion of oxygen into the
In the resting state (1) of cytochrome P450, the low potential heme iron (−400 through −250 mV) is mostly in the low-spin state (S = 5/2) with the water coordinated as the sixth ligand trans to the proximal thiolate. When the substrate binds in the vicinity of the heme, water ligation is destabilized, and the heme iron may turn to the tetrapyrrolic, pentacoordinated configuration 2 with the concomitant shift of the spin equilibrium to the high-spin (S = 3/2). As a result of the loss of the sixth ligand by the heme iron, the cytochromes P450 saturated with substrates are reduced much faster than in the substrate-free state because of the positive shifts of the redox potential by ∼100 mV. Such control over the rate of the first electron transfer by the substrate binding represents an important feature of the overall steady-state kinetics of many cytochromes P450; although some isozymes exist in a predominantly high-spin state, even without substrates, and they presumably lack this switch.

Some cytochromes P450 can bind two or three molecules of substrate or inhibitor simultaneously (CYP107, CYP158A1, CYP158A2, CYP2B4, CYP2C8, CYP2C9, CYP3A4, CYP245) and as a result may reveal non-Michaelis turnover kinetics. Functional homotropic and heterotropic cooperativity (i.e., the perturbation of the activity of the enzyme with respect to one substrate molecule by the binding of another molecule of the same or different type, respectively) typically is observed for such cytochromes P450. This allosteric behavior of xenobiotic metabolizing cytochromes P450 constitutes the important aspect of the drug–drug interaction problem in the pharmaceutical and medicinal chemistry and industry.

First electron transfer—reduction of ferric heme

Reduction of the heme iron of cytochromes P450 to the ferrous state is necessary for the binding and subsequent activation of atmospheric dioxygen. Initially, two electrons are derived from NAD(P)H by flavin adenine dinucleotide (FAD)-containing prostaglandin peroxidases (CYP2J and CYP108 families), which usually do not involve NADPH and oxygen consumption.

Reaction Cycle and Intermediates

Substrate binding and spin shift

In the resting state 1 (Fig. 1) with no substrate at the active site of cytochrome P450, the low potential heme iron (−400 through −250 mV) is mostly in the low-spin state (S = 5/2) with the water coordinated as the sixth ligand trans to the proximal thiolate. When the substrate binds in the vicinity of the heme, water ligation is destabilized, and the heme iron may turn to the tetrapyrrolic, pentacoordinated configuration 2 with the concomitant shift of the spin equilibrium to the high-spin (S = 3/2). As a result of the loss of the sixth ligand by the heme iron, the cytochromes P450 saturated with substrates are reduced much faster than in the substrate-free state because of the positive shifts of the redox potential by ∼100 mV. Such control over the rate of the first electron transfer by the substrate binding represents an important feature of the overall steady-state kinetics of many cytochromes P450; although some isozymes exist in a predominantly high-spin state, even without substrates, and they presumably lack this switch.

Some cytochromes P450 can bind two or three molecules of substrate or inhibitor simultaneously (CYP107, CYP158A1, CYP158A2, CYP2B4, CYP2C8, CYP3A4, CYP245) and as a result may reveal non-Michaelis turnover kinetics. Functional homotropic and heterotropic cooperativity (i.e., the perturbation of the activity of the enzyme with respect to one substrate molecule by the binding of another molecule of the same or different type, respectively) typically is observed for such cytochromes P450. This allosteric behavior of xenobiotic metabolizing cytochromes P450 constitutes the important aspect of the drug–drug interaction problem in the pharmaceutical and medicinal chemistry and industry.
bacterial cytochromes P450 and mitochondrial P450 systems. In these systems, the electron transfer from the FAD-containing reductase to the heme enzyme cytochrome P450 is mediated by the soluble iron-sulfur protein ferredoxin. In bacteria, all three proteins are soluble, whereas in mitochondria of eukaryotic cells, cytochromes P450 and reductases are associated with the inner membrane. The systems that belong to the second class include only two proteins, and the cytochrome P450 is reduced directly by the cytochrome P450 reductase (CPR), which contains both FAD and flavin mononucleotide. Variations of these systems include fusion enzymes in which a single polypeptide chain folds into two or three domains that correspond to the heme protein, iron-sulfur protein, and flavinprotein components of cytochrome P450 catalytic system (12). Many CPR flavoproteins strongly favor the usage of either NADH or NADPH, whereas some flavoproteins can use both efficiently.

Electron transfer in Class I systems is studied in detail for CY P101 and its reduct partner putidaredoxin (2Fe-2S), which contains protein with molecular mass 12.6 kDa. Binding of puti- daredoxin at the heme proximal site of CY P101 and formation of the transient electron transfer complex is accompanied by subtle conformational changes of both proteins and concomitant changes in their redox properties that stimulate the reduction of the heme iron to the ferrous state. The complex of CY P101-Pd is not very tight (Kd is in micromolar range), because facile dissociation is necessary to maintain the fast overall turnover. However, this complex is highly specific and can be perturbed significantly by single point mutations at the protein-protein interface. The apparent first-order rate of the first electron transfer in this system measured in kinetic experiments is in the range of 50–100 s⁻¹.

Oxygen binding and autoxidation—first uncoupling

Ferrous cytochromes P450 bind dihydrogen as the sixth ligand to the heme iron, and the resulting diamagnetic oxy-ferrous complex 4 is similar to that in myoglobin and hemoglobin. Quantum chemical studies show that the wave function of this complex predominantly is a mixture of approximately similar fractions of Fe²⁺−Ο₂ (closed shell, Pauling configuration) and Fe⁷⁺−Ο₂⁻ (open shell singlet, Weiss configuration), with the minor contribution of other configurations. Results of vibrational spectroscopy reveal that both 0−0 and Fe–O bonds are slightly weaker in oxyP450 (1128–1139 cm⁻¹ and 539–543 cm⁻¹) than in oxygenated myoglobin and hemoglobin (1148 and 572 cm⁻¹, respectively) (13). This difference is caused by more electron-rich thiolate ligand and the excess de- nuration of electron density to the antibonding orbitals of Fe-0-0 moiety in cytochrome P450. Similar effects were observed in nitric oxide synthase, which has the same thiolate ligand for the heme iron.

A comparison of available X-ray structures of oxy-ferrous complexes in CY P101, CY P107, and CY P158 reveals very similar structural arrangements of the heme iron and axial ligand with major differences concerning the detailed stereochemistry of the amino acid side chains and water molecules in the immediate vicinity of the bound dihydrogen (14, 15). The main difference is caused by the replacement of the mostly conserved alcohol functionality of T252 residue in CY P101 by methyl groups of A245 (CY P107) and A245 in CY P158. The absence of T252 hydroxyl within the hydrogen bond distance to the distal oxygen atom of the coordinated dihydrogen is compensated in the latter two enzymes by the specific positioning of hydroxyl groups on the bound substrates, which in turn stabilize water molecules assumed to be the ultimate source of protons for the oxygen activation (see below).

Usually, oxygen binding is faster than other steps in the enzymatic cycle, and it is not rate limiting in the overall substrate turnover. However, the lifetime of oxy-ferrous complex in cytochromes P450 at physiologic temperatures is in the range of seconds, because it decomposes spontaneously back to the resting ferric state with the formation of superoxide ion in an autoxidation reaction. Usually, the rate of autoxidation is lower by one or two orders of magnitude for the enzymes saturated with substrate. Higher stability of substrate-bound oxy-ferrous cytochromes P450 is explained in part by their higher redox potentials (i.e., by thermodynamic stabilization of the oxygenated heme enzyme in the enzyme-substrate complex). In addition, the presence of substrates in the active site was shown to impose dyna- namic constraints on the superoxide escape and to increase the kinetic stability of oxy-ferrous P450. Inhibition of autoxidation is an essential part of the regulatory role played by substrates in the overall optimization of the P450 cycle via minimization of this uncoupling reaction.

Second electron transfer and protonation, compound 0

Direct observation of the reduction of oxy-ferrous cytochrome P450 and formation of peroxy-ferric (5a) or hydroperoxo-ferric (5b) complex was never accomplished at ambient conditions because this intermediate, termed Compound 0, undergoes additional transformations and disappears faster than it is formed. However, it was isolated and studied by electron paramag- netic resonance (EPR), optical absorption, and resonance Raman spectroscopy at low temperatures using radiolytic reduction of oxy-ferrous cytochromes P450. As expected, the hydroperoxo-anion forms a relatively strong bond with the heme iron, and the resulting complex reveals the typical fea- tures of low-spin heme-thiolate complexes with the narrow span of g-values in EPR spectra and red-shifted Soret band (436–440 nm) (20). Spectroscopic and theoretical studies reveal that the length and strength of the 0–0 bond in Compound 0 are similar to those observed in the low-spin oxygen activating nonheme (hydro)peroxo-ferric complexes. The presence of the distinct hydroperoxo-ferric heme intermediate in the frozen solu- tions and in crystals of cytochromes P450 and several other heme proteins indicates that the spontaneous homolysis or het- erolysis of 0–0 bond is unfavorable. Thus, the efficient for- mation of the main active intermediate Compound I requires catalytic delivery of the second proton to the distal oxygen atom.
Second protonation and O-O- scission.

peroxide dissociation–second uncoupling

The second protonation of Compound 0 at the distal oxygen atom catalyzes fast scission of O–O bond with the leaving water molecule. In cryoradiolytic experiments, Compound 0 is stable below the glass transition temperature, 180–190 K, which means that the second proton delivery requires sufficient mobility as well as possibly relaxation and diffusion of the solvent molecules. Experiments with the native CY P101 and D251N mutant proved that at higher temperatures, Compound 0 disappears with formation of Compound I (6) and concomitant product formation (16). For the T252A mutant CY P101, in which the native proton transfer mechanism is perturbed, the dissociation of peroxide with no product formation is the dominant path of Compound 0 decomposition. The latter reaction is considered as the main source of reactive oxygen species in the poorly coupled P450 systems, in which the redox equivalents from NADPH oxidation are not used efficiently for the product formation. In general, the coupling efficiency measured by the ratio of the product molecules formed per one consumed NADPH molecule can be very different for the same cytochrome P450 with different substrates. The efficient proton delivery requires specific positioning and stabilization of water molecules near dioxygen moiety, which can be perturbed significantly by variations in the structure of substrate.

Compound I—properties and activity

Compound I (6) is the main catalytically active intermediate. A description of the main chemical reactions is given above. Compound I has yet to be characterized on the natural P450 pathway, although the transient formation of the short-lived ferryl-oxo intermediate with porphyrin-cation radical has been documented using reaction with peroxides. The electronic configuration of this intermediate, as evaluated by quantum chemical methods, represents it as a triradicaloid with two possible spin states \( S = \frac{1}{2}, \frac{3}{2} \), with closely spaced energies. In addition to the almost degenerate ground state with doublet and quartet relative stability regulated by the protein matrix, the sextet \( S = \frac{5}{2} \) can also be involved. The presence of high spin and low spin states of Cpd I provides variability in the reaction pathways and product distributions. This concept of two-state and multistate reactivity of Cpd I in cytochromes P450 developed by Shaik et al. (6, 9) provides basis for the unified explanation of the plethora of apparently contradictory experimental results in P450 chemistry.

Structures—Common Fold and Variations

All cytochromes P450, from soluble enzymes in archaea and bacteria to the mammalian integral membrane proteins, have essentially the same fold, which is specific for this superfamily despite their functional diversity (17) (Fig. 2). Only several enzymes with this fold do not follow classic P450 mechanism based on activation of dioxygen (i.e., nitric oxide reductase, CYP55 and CYP152, which uses hydrogen peroxide). Approximately 50% of the residues are located in helical fragments, which are designated by letters A to L with the addition of B', F', G', and J' in some isozymes, and 7–10% are found in small beta-sheets interspersed with flexible loops. As compared with the soluble bacterial enzymes, membrane-bound proteins usually have an extra N-terminal fragment that is thought to be incorporated into the membranes and longer loops with one or two small additional helical fragments.

The heme (protoporphyrin IX, heme b) is buried in the protein matrix between the long I-helix on the distal side and L helix on the proximal side. The loop at the N-terminus of L-helix contains absolutely conserved cysteinate with the deprotonated sulfur atom that serves as the proximal axial ligand of the heme iron. Three hydrogen bonds with this sulfur atom are formed by the amide groups of the highly conserved neighboring residues. These bonds are necessary to raise the redox potential of cytochrome P450 to the range accessible for the ultimate physiologic reductant NAD(P)H. Hydrogen bonds between propionic carboxyls and the side chains of protein amino acids prevent the loss of the heme and stabilize the native conformation.

A rather important and highly conserved feature in all structures of cytochromes P450 with only a few exceptions is the
presence of the acid-alcohol pair (i.e., sequential acid and alcohol side chains at the heme) just next to the dioxygen coordinated to the heme iron. An acidic side chain is provided by an Asp or Glu residue (n), whereas alcohol (n + 1) is from Thr or Ser. The hydroxyl of the latter is very often hydrogen bonded to the peptide carbonyl of the Gly (n − 3) forming a kink, which interrupts the regular α-helical structure and is thought to be essential for the efficient oxygen activation. Structures of the oxo-ferrous complexes resolved for CYPI01, CYPI07, and CYF158 reveal the appearance of specifically positioned water molecules, which are stabilized by the appropriate rearrangement of the side-chains of the acid-alcohol pair described above. These water molecules play an important role in the directed proton delivery provided by protein for the formation of Compound I and Compound II, although the details of the mechanism and specific pathway of this proton delivery are still debated.

Helices B, F, G, and G and connecting loops are involved in substrate binding and in general are much more variable and flexible (upper part of the substrate binding cavity shown in Fig. 2). These parts have different length in different classes of P450, with longer disordered insertions typical for eukaryotic enzymes and shorter and more compact structures characteristic for the proteins from extremophilic organisms. In all classes, however, the protein has to experience a large-scale conformational change to open access to the substrate-binding cavity. Comparison of the X-ray structures of cytochromes P450 with and without substrates and the results of molecular dynamics studies suggest that these conformational changes involve concerted movement of the fragments, which form the substrate-binding cavity that includes helices F and G and fragments between them, as well as B and C helices. Such an “opening” movement was first observed in the X-ray structures of CYPI01 with the “wired” substrates (19). In these structures, the diaminealkane chain traced the tentative substrate-binding cavity. However, in many cases, the conformations of cytochromes P450 with the bound substrates or inhibitors are even more compact than in substrate-free state, in which the flexible fragments rearrange to provide additional interactions with the hydrophobic molecule bound in the active center. In general, the high plasticity of P450 fold provides basis for the flexible fragments rearrangement of the side-chains of the acid-alcohol pair described above. These water molecules play an important role in the directed proton delivery provided by protein for the formation of Compound I and Compound II, although the details of the mechanism and specific pathway of this proton delivery are still debated.

Helices B, F, G, and G and connecting loops are involved in substrate binding and in general are much more variable and flexible (upper part of the substrate binding cavity shown in Fig. 2). These parts have different length in different classes of P450, with longer disordered insertions typical for eukaryotic enzymes and shorter and more compact structures characteristic for the proteins from extremophilic organisms. In all classes, however, the protein has to experience a large-scale conformational change to open access to the substrate-binding cavity. Comparison of the X-ray structures of cytochromes P450 with and without substrates and the results of molecular dynamics studies suggest that these conformational changes involve concerted movement of the fragments, which form the substrate-binding cavity that includes helices F and G and fragments between them, as well as B and C helices. Such an “opening” movement was first observed in the X-ray structures of CYPI01 with the “wired” substrates (19). In these structures, the diaminealkane chain traced the tentative substrate-binding cavity. However, in many cases, the conformations of cytochromes P450 with the bound substrates or inhibitors are even more compact than in substrate-free state, in which the flexible fragments rearrange to provide additional interactions with the hydrophobic molecule bound in the active center. In general, the high plasticity of P450 fold provides basis for the ability to bind multiple organic molecules with different sizes and chemical structures, which is critically important for the efficient metabolism of xenobiotics.

References


Further Reading


See Also

Drug Metabolizing Enzymes
Hemes in Biology
Inorganic Chemistry in Biology
Metallo-Enzymes and Metallo-Proteins, Chemistry of
Oxygen-Activating Enzymes, Chemistry of
Hemes in Biology

Paul R. Ortiz de Montellano, University of California at San Francisco,
doi: 10.1002/9780470048672.wecb221

Heme (iron protoporphyrin IX) is a fundamental factor in biology with many physiologic roles. Its biosynthesis and degradation are regulated tightly because heme in excess of that required for incorporation into the available apoproteins is toxic. Heme itself is the dominant form in biology, but it is modified additionally to satisfy specific biochemical requirements. Both the biosynthesis of heme and its distribution into the various compartments of the cell and organism require the intervention of heme transporters. In contrast to the passive transport afforded by proteins like albumin, the transfer of heme across membranes is mediated by energy-dependent transporters. One role of heme is to function as a regulatory molecule whose concentration controls the translational expression of multiple genes directly. As the prosthetic group of receptors that respond to oxygen, nitric oxide (NO), and other iron ligands, it is also involved in signaling. Finally, it is the prosthetic group of diverse families of hemoproteins, including the globins, cytochromes, P450 monooxygenases, NO synthases, dioxygenases, peroxidases, peroxygenases, and catalases.

Iron is an essential element for life because its oxidation-reduction properties are in the biologically accessible range, are readily modulated, and can be used to satisfy a diversity of functions. These functions include electron transfer, ligand binding and sensing, ligand transport, and catalytic functions that involve the activation of molecular oxygen or peroxides. However, iron itself is highly insoluble in the ferric state, and in the ferrous state it supports the formation of deleterious species such as the hydroxyl radical. Therefore, nature has evolved strategies for the chelation and encapsulation of iron that allow the exploitation of its useful features but minimize the associated toxicity. The most effective and ubiquitous of these solutions is to incorporate the iron into a porphyrin, producing heme. Incorporation of the iron into heme not only allows better control of its solubility and spatial localization, but also enables fine-tuning of its redox properties for specific tasks.

The field of hemes in biology is enormous and it is possible only to touch on the salient points in this article, which focuses primarily on the biosynthesis, degradation, transport, and regulatory role of heme. The catalytic functions of heme are diverse and fascinating, but many of the relevant enzymes are summarized only briefly here.

Heme Structures and Nomenclature

Heme (heme B) is a highly conjugated cyclic tetrapyrrole in which an iron is coordinated to the four nitrogen atoms of a protoporphyrin IX framework. In heme, four methyls, two vinyls, and two propionic acid substituents are distributed asymmetrically around the porphyrin periphery (Fig. 1). The four pyrrole rings are labeled A to D, the four meso-carbons are denoted by Greek letters, and the positions of the eight substituents are numbered. Crystallographers sometimes use a different A–D order for the four pyrrole rings. The formal IUPAC nomenclature for heme (Fig. 1) is used by organic chemists but is not often used in biology. The compound is called heme when the iron is in the ferrous state and is called hemin when it is in the ferric state, but chemists do not always adhere strictly to this distinction.

Heme is the most common biologic form, but additional elaborations of the porphyrin skeleton are required for certain purposes. Among these more baroque iron porphyrins are heme A, the structure found in cytochrome c oxidase; heme C, the prosthetic group of cytochrome c in which the covalently bound cysteines are part of the protein structure; and less common structures such as the heme D in the catalase from Escherichia coli (Fig. 2). The 5-methyl group of heme is modified additionally by the formation of an ester link to an active-site carboxyl group in most members of the CYP4 family of cytochrome P450 enzymes (1), and by both the 1-methyl and the 5-methyl groups in the mammalian peroxidases, including myeloperoxidase, lactoperoxidase, eosinophil peroxidase, and thyroid peroxidase. In myeloperoxidase, a third covalent bond is formed in addition to these ester links that links a methionine to the 2-vinyl group. Other modifications are known that tailor the heme for specific biologic functions.
Heme in Biology

Figure 1 The structures of heme (Fe^{2+}) and hemin (Fe^{3+}). The substituents and pyrrole rings are numbered (A) as most commonly used in biology and (B) according to the IUPAC nomenclature convention.

Figure 2 The structures of heme A, heme C, heme D, and the modified heme found in the CYP4 family of P450 enzymes.

Heme Biosynthesis

Mammalian heme biosynthesis traverses eight enzymatic steps and requires shuttling of the first intermediate, 5-aminolevulinic acid from the mitochondrion into the cytosol, and later transferring coproporphyrinogen III back into the mitochondrion for the three final biosynthetic steps (2).

Pre-porphyrin steps

The first and rate-limiting step in heme biosynthesis is the condensation of glycine and succinyl-CoA to give 5-aminolevulinic acid by aminolevulinic acid synthase (ALAS) (Fig. 3). Two mammalian forms of this enzyme are known: ALAS1, which is ubiquitously expressed, and ALAS2, which is specifically expressed in erythroid precursors (3, 4). Both enzymes are homodimers and use pyridoxal 5-phosphate as a cofactor. Although the crystal structures of mammalian ALAS1 and ALAS2 have not been determined, the structure of a highly homologous enzyme from Rhodobacter capsulatus has been (5). The condensation reaction catalyzed by these enzymes involves the formation of an imine bond between the pyridoxal and glycine followed by deprotonation, the displacement of the CoA from succinyl-CoA, decarboxylation, and the hydrolytic release of 5-aminolevulinic acid from the pyridoxal cofactor. ALAS1 has three putative
heme-binding motifs (HRMs), two in the N-terminus of its mitochondrial targeting domain and one in the N-terminus of the mature protein. Binding of heme to one or more of these HRMs inhibits mitochondrial import of the protein and thus provides end-product feedback regulation of heme biosynthesis (6). This regulation complements the direct heme-dependent downregulation of ALAS expression at the transcriptional level. Mutations in ALAS result in X-linked sideroblastic anemia (5).

In plants and in most bacteria, 5-aminolevulinic acid is produced by an alternative pathway involving tRNA-bound glutamate and two enzymatic steps catalyzed by glutamyl-t-RNA reductase and by glutamate-1-semialdehyde-2,1-aminomutase (7, 8).

The 5-ALA formed in mammalian mitochondria is transferred to the cytosol, where aminolevulinate dehydratase (ALAD) catalyzes the condensation of two monomers to give porphobilinogen (Fig. 3), the monopyrrole building block of the porphyrin skeleton. ALAD is a zinc-dependent tetramer of homodimers with one catalytic site per dimer (9, 10). The condensation reaction requires the formation of an enamine link, tautomerization of the enamine, closure of the five-membered ring by an aldol-like reaction, and dehydration to give the pyrrole ring. Mutations in ALAD are associated with the rare recessive autosomal ALAD porphyrias. Deficiencies in ALAD activity are also involved in hepatorenal tyrosinemia and lead poisoning, the former because the defect produces succinyl acetone that...
inhibits ALAD and the latter because lead displaces essential zinc ions from the protein complex.

The next stage of heme biosynthesis is the formation of the linear tetrapyrrole hydroxymethylbilane by porphobilinogen deaminase (PBGD). The crystal structure of PBGD shows that a dipyrrole cofactor is covalently within the active site of the enzyme (11). This dipyrrole is condensed with four additional porphobilinogen monomers to form a hexameric polymer. Cleavage of the tetrapyrrole unit from the dipyrrole cofactor produces hydroxymethylbilane (Fig. 3). Mutations in the PBGD gene are responsible for the disorder known as acute intermittent porphyria (12). A cute episode of this disorder are treated by the administration of hematin (the Fe–OH complex of heme), which leads to restoration of heme levels, the downregulation of ALAS1, and a decrease in the synthesis of toxic porphyrin precursors.

**Porphyrens and heme**

Hydroxymethylbilane is converted by uroporphyrinogen III synthase into uroporphyrinogen III in a remarkable ring-closing reaction that inverts the orientation of pyrrole ring D of hydroxymethylbilane (Fig. 3). This ring flip exchanges the positions of the propionic and acetic acid side chains and introduces the substituent asymmetric character of heme. The crystal structure has not yet provided a clear understanding of how this transformation is achieved (13). Mutations in uroporphyrinogen III synthase cause a rare disease known as congenital erythropoietic porphyria (14). This disease results in the accumulation of uroporphyrin I, the abnormal porphyrin without the pyrrole ring D. Fig. Uroporphyrin I is formed by an abnormal, nonenzymatic closure of hydroxymethylbilane and cannot be processed further. The accumulated porphyrin acts as a photosensitizer that causes the cutaneous lesions associated with the disease.

Uroporphyrin III is the branching point in the pathway at which heme synthesis diverges from those of chlorophyll and the corrin. In heme biosynthesis, uroporphyrinogen III is next decarboxylated by uroporphyrinogen III decarboxylase to give coproporphyrinogen III (Fig. 3) (15). This enzyme promotes the decarboxylation of the four acetic acid side chains to give the methyl groups at positions 1, 3, 5, and 8 of the heme framework. Mammals use a form of this enzyme that is oxygen-dependent and releases CO2 and H2O2, but an oxygen-independent form is found in bacteria. The crystal structures of the human and the yeast enzymes have led to both acid-base and free radical mechanistic proposals (18, 19).

This decarboxylation reaction occurs in the mitochondria and thus requires the translocation of coproporphyrinogen III from the cytosol into this organelle. Hereditary coproporphyria, the disorder in humans that is caused by mutations in the coproporphyrinogen oxidase gene, is treated by the administration of hematin.

Coproporphyrinogen oxidase converts protoporphyrinogen IX to the fully deuterated porphyrin in a reaction that uses O2 as the terminal electron acceptor (Fig. 3). The crystal structure of the homodimeric enzyme shows it has one FAD per monomer, which presumably mediates the porphyrin oxidation reaction (19). Like the decarboxylation mediated by coproporphyrinogen oxidase, this reaction also occurs in the mitochondrion. Mutations in the protoporphyrinogen oxidase gene are responsible for variegate porphyria (21). A cute attacks of this disease can be effectively treated by intravenous administration of hematin.

The synthesis of heme (Fig. 1) is completed in the mitochondrion by insertion of iron into the protoporphyrin IX framework by ferrochelatase. Ferrochelatases from various organisms have been crystallized and their structures determined. The human enzyme contains one 2Fe-2S cluster in each of the two subunits of the functional dimer (22); possibility as a mechanism to link heme synthesis to iron availability. Erythropoietic protoporphyria, which is characterized by cutaneous photosensitivity, is caused by mutations in the ferrochelatase gene (23).

**Additional elaboration of the heme framework**

The formation of heme C requires the addition of a cysteine thiol to each of the two vinyl groups of heme B (Fig. 2). The thiol links are usually formed with a cysteine in a CXXXCH sequence, but small variations of this motif are known (24).

Three systems catalyze the formation of these bonds and are known imaginatively as systems I, II, and III (25). System I occurs commonly in Gram-negative bacteria and in plant mitochondria, system II in Gram-negative and Gram-positive bacteria as well as in thylakoids, and system III in fungi, invertebrates, and vertebrates. In all three systems, an addition of the cysteine to the vinyl group is mediated enzymatically, with a supporting cast of several proteins required in systems I and II but not in system III. Despite its ubiquitous nature, the catalytic or structural advantage of the heme C covalent linkage remains unclear (26). Interestingly, a Synechocystis hemoglobin is known in which the heme is covalently linked through a vinyl group to a histidine residue (27).

The synthesis of heme A (Fig. 2) involves the initial addition of the farnesyl moiety to the heme 2-vinyl group by heme O synthase, which generates heme O that only has this modification. In a second step, heme A synthase oxidizes the B-methyl of heme O to an aldehyde, which generates heme A. An electron transfer mechanism (rather than double hydroxylation) has been proposed for this final biosynthetic step (28).

In contrast to the biosyntheses of heme A and heme C, which require dedicated protein catalysts, the ester groups that link methyl groups of heme B with protein carboxy groups are generated autocatalytically. As has been clearly shown...
for lactoperoxidase (29) and for several CYP4 cytochrome P450 enzymes (30), the ester links are generated in the initial catalytic turnovers of the proteins in question. Investigation of the reaction details, and mimicry of the reaction in horseradish peroxidase (31), a protein that normally does not form such bonds but does upon introduction of an active site carboxylic acid residue, strongly supports a carboxyl radical/hydrogen abstraction mechanism (Fig. 4).

Heme Catabolism

The only physiologic (as opposed to pathologic) mechanism for the degradation of heme in mammals is catalyzed by heme oxygenase. Analogous heme oxygenases are found in plants, bacteria, and fungi. Two heme oxygenases, known as HO1 and HO2, are present in mammals. HO1 is distributed widely, is induced by a diversity of agents, and is critical for the removal of the heme released by erythrocyte lysis and the degradation of hepatic hemoproteins. The levels of HO1 expression are regulated in a feedback manner at the transcription level by heme. HO2, although widely distributed, is concentrated in tissues such as the testes and the brain, is only induced by hormones, and has been postulated to fulfill physiologic roles beyond the simple removal of excess heme. For example, HO2 has been reported to function as an oxygen sensor that controls a potassium channel in the carotid body (32). The structures of rat and human HO1 (33, 34), and of heme oxygenases from several microorganisms (35), have been determined, but HO2 has not been crystallized. One of the significant differences between the sequences of HO1 and HO2 is the presence in HO2 of two HRMs analogous to those in 5-aminolevulinate synthase. The role of these HRMs in HO2 is unknown.

The oxidation of heme by heme oxygenase requires O2 and seven electrons. The electrons in mammals are provided by cytochrome P450 reductase, but alternative electron donors are employed in the plant and bacterial systems. However, the actual transformation is similar in all cases. The first step is hydroxylation of the meso-carbon at which the porphyrin ring will be cleaved (Fig. 5). In mammals, this position is the α-meso carbon, but alternative meso-positions can be oxidized in bacteria. The resulting α-meso-hydroxyheme is further oxidized by O2 to α-verdoheme and CO. Additional catalytic turnover produces ferric biliverdin that, after reduction of the iron, loses the iron atom and dissociates from the protein. In mammals, the biliverdin is further reduced to bilirubin by biliverdin reductase and is then excreted as a glucuronic acid conjugate. In situations where the relevant glucuronyl transferase activity is low, as in some neonates or in Crigler–Najar syndrome, bilirubin accumulates and causes jaundice or, at higher levels, causes neurological toxicity. Agents such as tin protoporphyrin IX can be used to inhibit heme oxygenase and thus to decrease the formation of biliverdin and bilirubin.

All the products of the oxidation of heme by heme oxygenase are important physiologically. Biliverdin and its reduction product bilirubin are powerful antioxidants and, at nontoxic concentrations, contribute to cellular protection. CO, the second product, also has potent biologic activities, although it is often...
Figure 5 The multistep conversion of heme to α-biliverdin catalyzed by heme oxygenase. The electrons from NADPH are transferred to heme oxygenase by cytochrome P450 reductase or other electron donor proteins.

unclear how these activities are mediated. Finally, the ferrous iron that is released causes upregulation of the iron-binding protein ferritin and can also have an antioxidant effect. These products, singly or in combination, have been shown to confer protection against oxidative injury and cellular stress. Induction of HO1, or administration of CO or bilirubin, inhibits apoptosis, whereas inhibition of HO1 stimulates apoptosis. Therefore, the inhibition of heme oxygenase may be relevant to the treatment of cancer (36). In another context, oxidized low density lipoprotein (LDL) is a key player in the development of atherosclerotic plaques, and LDL oxidation is inhibited by HO1 and its antioxidant products. The manipulation of the HO1 system is therefore of potential interest in cardiovascular disease (37). The heme oxygenases and/or their products are relevant to the amelioration of reperfusion injury, avoidance of transplant rejection, and a variety of inflammation-based disorders (38).

In bacteria, heme oxygenases play a role in iron acquisition. A greater variety is seen in the bacterial heme oxygenases than is seen in man, but the essential catalytic features are preserved and the mechanism of the oxidation is the same (35, 39). The presence of two heme oxygenases in Pseudomonas aeruginosa, one that cleaves the heme at the α-meso-carbon and the other at the δ-position, may reflect the fact that one of the two provides the tetrapyrrole for a bacterial phytochrome response regulator (40, 41). It may be desirable to generate a different regioisomer of biliverdin in the iron acquisition pathway so that it does not interfere with the regulatory system. Under oxidative-stress conditions, heme can also be degraded nonenzymatically to nonporphyrin monopyrrole and dipyrrole products, although the mechanistic details of this degradation remain obscure (42).

**Figure 5**

Heme Transport

Heme (hemin) at physiologic pH is relatively insoluble and is found bound to proteins or membranes except for a low concentration (<10⁻⁷ M) of so-called “labile” heme (43). Unbound heme is potentially cytotoxic and its levels are controlled tightly by feedback regulation of its biosynthesis at the level of 5-aminolevulinic acid synthase, and of its degradation by heme oxygenase, which is a heme-inducible enzyme. As heme is lipophilic, it is able to diffuse into and through cell membranes. Nevertheless, the efficient trafficking of heme involves protein transporters, some of which have been identified. Transporters of heme in the circulatory system, such as albumin, can be classified simply as heme-binding proteins. However, proteins that mediate the energy-dependent transmembrane transport of heme have been identified. Subtractive suppression hybridization has led to the identification of heme-carrier protein 1 (HCP1) in the duodenum of mice (44). This protein also transports Zn protoporphyrin IX, which indicates that the iron is not essential for recognition. Interestingly, homologous proteins are found in man and other mammals. HCP1, which is upregulated in hypoxia, has 459 amino acids and is predicted to have nine
transmembrane domains (44). It has a GxxSxxGxGRR motif that is found in bacterial tetracycline transporters, with which it has significant (22%) similarity, but it does not have the CP motif of the HRMs in proteins such as HO2.

Heme must be exported from the mitochondrion, where the final steps of heme synthesis occur, and a system may exist to import heme into the nucleus, as diverse genes are regulated directly at the transcriptional level by heme. The transporters that mediate this trafficking of heme have not been unambiguously identified. Recent work has shown that the mitochondrial ATP-binding cassette transporter ABCB6, a homodimeric protein that is located on the outer mitochondrial membrane, is required for porphyrin uptake into the mitochondria during heme biosynthesis (45). Heme can be transported by this protein, but it transports coproporphyrinogen III more efficiently, which implicates it as the protein involved in importation of this heme precursor. A 60-kDa protein known as FLVCR, which serves as the feline leukemia virus C receptor (thus the name), has been shown to export cytoplasmic heme from developing erythroid cells and is postulated to do so to protect them from heme toxicity (46). FLVCR has the same GxxSxxGxGRR motif found in HCP1. A second heme efflux protein denoted as ABCG2, also known as breast cancer resistance protein (BCRP), has been reported (47). This 70-kDa protein, a member of the ABC transporter family, has six predicted transmembrane domains and is functional as a homodimer.

Heme that is released into the blood stream by lysis of red cells, catalysis of haptoglobin-hemoglobin complexes, or other mechanisms binds to albumin ($K_d \sim 10^{-10} M$) and hemopexin ($K_d \sim 10^{-8} M$) (48). The hemopexin–heme complex is taken up in the liver by a receptor-mediated process. A candidate for the hemopexin–heme receptor has been identified (49). Other heme-binding proteins are scavenging and degrading heme. Gram-negative bacteria have receptors that interact directly with hemoproteins, such as hemoglobin and hemoglobin–haptoglobin, or that interact with a hemophore excreted by the bacterium that scavenges the heme and delivers it to the cell (50). The receptors use the TonB/ExbB/ExbD complex to transport the heme through the outer membrane. The best understood of these hemophores is Serratia marcescens, for which a crystal structure is partially determined by control of Bach1 nuclear localization. Bach1 has six HRMs that are critical for regulation of its binding to DNA, whereas seven HRMs are involved in its response to heme.

In mammals, Bach 1 is a heme-activated leucine zipper protein that acts as a transcriptional repressor (54). Bach1 associates with members of the Maf-related oncogene family. The resulting heterodimers bind to the Maf recognition element of target genes, including HO1, globin genes, and ALAS2. Bach1 has six HRMs that are critical for regulation of its activity by heme. Heme binding seems to act as a signal for nuclear export of the protein, so that heme regulation is partially determined by control of Bach1 nuclear localization. A concentration of $-1 \mu M$ heme almost completely inhibits the binding of the Bach1-Maf dimer to DNA.

The maturation of reticulocytes to erythrocytes requires a massive synthesis of heme to enable the assembly of hemo
globin. It is important to couple the synthesis of heme to that of the hemoglobin α-chains and β-chains, as both the unassembled apoproteins and excess heme are toxic to the cell. The HRI kinase is responsible for coordinating heme and protein synthesis in reticulocytes. HRI inhibits globin synthesis at the translation initiation level under limited heme conditions. Under these conditions, HRI is activated and phosphorylates Ser51 of the eIF2 α-subunit. The phosphorylated protein binds tightly to eIF3B and ties up this protein, preventing it from catalyzing the replacement of GDP by GTP in eIF2α required for its role in initiating protein synthesis (55). HRI has three HRMs, one of which binds heme reversibly and is thought to couple the activity of the protein to the heme concentration and another that binds heme stably and is thought to be involved in the response of HRI to gases: NO increases HRI activity, whereas CO decreases it. The molecular chaperones Hsp70 and Hsc70 play important roles in this process, as immature HRI has no activity but becomes responsive to heme deprivation when transiently complexed to the chaperones.

Neuronal differentiation has been found in a model system to be heme-dependent (58). Inhibition of heme synthesis at the level of ALAS reduces the number and the length of neurites to and regulates histidine kinases, heme-responsive transcription factors, cyclic nucleotide phosphodiesterases, and factors such as the eIF2α heme-regulated HRI kinase. Through these interactions, heme contributes to cellular functions such as cell growth and differentiation, oxygen and nitric oxide sensing, cell respiration, and globin gene activation. Furthermore, as illustrated by its control of heme synthesis through binding to ALAS1 and inhibition of its migration into the mitochondrion, it can influence protein maturation directly.

5. cerevisiae, a facultative aerobic, responds to changing oxygen levels by altering the expression of a battery of genes. This coordinated response to oxygen levels in S. cerevisiae is mediated by the heme activator protein Hap1 (53). The activation of Hap1 increases in parallel with the concentration of heme and reaches maximum activation at micromolar heme concentrations. The binding of heme allows Hap1 to bind to regulatory motifs in nuclear DNA and, thus, allows it to promote the transcription of genes encoding proteins that are essential for respiration and the control of oxidative damage. A repressor (ROX1) of genes involved in anaerobic growth is also regulated by Hap1. A zinc cluster and a dimerization domain are involved in its binding to DNA, whereas seven HRMs are involved in its response to heme.

Heme as a Regulatory Molecule
Heme is not only a protein prosthetic group but also it controls the activities of diverse regulatory systems directly. Heme binds
induced by nerve growth factor (NGF), and this effect can be overcome by the addition of exogenous heme. These consequences of heme deficiency reflect the associated inactivation of the NGF-dependent Ras-ERK 1/2 signaling pathway.

Heme as an Enzyme Prosthetic Group

Heme is a ubiquitous prosthetic group in proteins with oxidation-reduction functions. The proteins that incorporate heme as an essential component include respiratory proteins, such as cytochrome c and cytochrome c oxidase. The oxygen carriers myoglobin and hemoglobin are members of a second class of hemoproteins that have some relationship to gas sensor hemo-proteins, as represented for NO by guanylate cyclase; for O2 by Hap1, FixL, and HemA T; and for CO by CooA. The largest class of hemoproteins is made up of catalytic enzymes and includes the cytochromes P450, peroxidases, peroxygenases, catalases, NO synthases, prostaglandin synthases, and thromboxane/prostacyclin synthases.

A general scheme of the catalytic manifold of hemoproteins is outlined in Fig. 6. In this abbreviated mechanism, the reaction of the ferric heme with H2O2, or with oxygen and two electrons derived from NAD(P)H, produces a “Compound I” ferryl (FeIV = O) intermediate. This intermediate can catalyze a single two-electron oxidation (monooxygenase or peroxygenase activity) or two one-electron oxidations (peroxidase activity) (57, 58). In addition to the dominant role of the ferryl intermediate in hemoprotein catalysis, oxidations can also be mediated by intermediates that precede the ferryl species. For example, the iron peroxy anion (Fe-OO-), a precursor of the ferryl intermediate, is thought to be involved in the carbon–carbon bond cleaving reactions of cytochrome P450 enzymes and the formation of NO by NO synthases. The function of these and other catalytic hemoproteins, all of which illustrate the virtuoso catalytic power and flexibility of the heme group, are reported in more detail in other chapters of this volume.

References

11. Hellwell JR, Nishi YP, Habash J, Faulder PF, Rafferty J, Cianci M, Wu M, Hadner A. Time-resolved and static ensemble...


Further Reading

Latunde-Dada GO, Simpson RJ, McKee AT. Recent advances in mammalian haem transport. Trends Biochem. Sci. 2006;31:182–188.

See Also

Hemoglobin and Myoglobin, Chemistry of
Metal Transport Through Membranes
Metalloenzymes, Chemistry of
Nitric Oxide Signaling
Oxygen Activating Enzymes, Chemistry of
Sensing and Adapting, Bacterial Mechanisms for
Human hemoglobin binds four oxygen molecules with positive cooperativity in a binding cascade of eight partially ligated intermediates whose oxygen ligands are distributed in different combinations between the hemoglobin’s two αβ dimers. Traditionally, it had been assumed that, given their close association in forming the hemoglobin tetramer, the two αβ dimers would always respond to oxygen binding in a synchronous and symmetric manner. Using linkage thermodynamics, the intermediate binding constants are evaluated via dimer–tetramer assembly with ligand configurations within the tetramer fixed through the use of hemesite analogs. It is observed that the free energy contribution of the asymmetrically ligated intermediate composed of one fully oxygenated αβ dimer plus one unoxygenated dimer is not the same as other doubly ligated intermediates, which contain at least one bound oxygen on both dimers. Therefore, the dimers within the tetramer respond to oxygenation differently, and cooperativity is dependent on the distribution of ligands between the two αβ dimers.

Introduction

The structural and functional properties of human hemoglobin (Hb) have been the subject of study for decades, stimulated by the intriguing characteristic of positive cooperativity. How do the four subunits that compose the Hb tetramer communicate with one another? The answer to this question has been sought primarily through the comparison of deoxy with oxy Hb. However, to understand the molecular mechanism of a chemical reaction, it is necessary to characterize the intermediate(s) of the process, and the reaction of Hb with O2 is no exception.

The binding of four O2 ligands by human Hb occurs through a series of 14 partially ligated intermediates, of which eight are unique in their combinatorial arrangement of bound O2 among the two α-subunits and two β-subunits. The well-known sigmoidal binding curve that results (see Fig. 2 for an example) is indicative of a strong positive cooperativity of oxygenation, thus, the binding constants for each Hb intermediate are changing as the O2 binding process continues.

The individual microscopic binding constants cannot, however, be measured from the binding curve: Only four average, macroscopic binding constants can be directly observed. This result is a result of several factors, foremost of which is the high cooperativity of O2 binding itself, which suppresses the concentrations of the intermediates. Thus, the binding curve is dominated by the properties of the two end-states, i.e., the fully deoxygated tetramer and its fully oxygenated counterpart. Other factors that contribute to the low resolution of the binding curve are the lability of the bound O2 and the continuous dissociation of the tetramer to its constituent αβ dimers.

Therefore, in a system that binds O2 close to equilibrium to begin with, the rearrangement of bound O2 among the heme binding sites acts only to mask the individual properties additionally (such as a microscopic binding constant) of a given intermediate. This classic problem of disproportionation can be solved experimentally through the use of hemesite analogs that either block O2 binding or O2 dissociation in specific subunits (α1, β1, α2, β2) within the tetramer. The dissociation of tetramer to free dimer, and the resulting dimer rearrangement among tetramers, cannot be blocked but can be measured. The dimer → tetramer assembly free energy, $\Delta G_{asm}$, can then be applied as a constraint that permits the determination of the Gibbs free energy, $\Delta G_{ij}$, of each intermediate binding reaction by employing thermodynamic linkage analysis.

The microscopic O2 binding constants thus determined reveal a particularly strong energetic coupling between the subunits.
within each αβ dimer of the tetramer (2). This intradimer cooperativity is evident particularly in the intermediate composed of one oxy dimer and one deoxy dimer, or the asymmetric doubly ligated species. Identification of this intermediate provided the first direct experimental evidence of intradimer cooperativity, which challenged the commonly held two-state model of cooperativity, in which the two dimers within the tetramer are assumed to maintain the same structural and energetic properties throughout the binding process. Rather than maintaining dimer–dimer symmetry, the αβ dimers each exhibit a unique O2 affinity and continue to modulate the O2 affinity of each other.

Background
Since the determination of its crystal structure almost five decades ago, the study of cooperativity and allostery in human Hb was focused primarily on the properties of the two end-states, the deoxy and oxy tetramers. The approach to mechanistic questions of subunit–subunit coupling within the tetramer is now shifting to the characterization of the partially ligated intermediates. Although crystal structures are not yet available for the intermediates, their individual O2 binding constants are now determined for one set of solution conditions.

Structural elements of the hemoglobin tetramer
The human hemoglobin tetramer is composed of two types of polypeptide chains, designated α (with 141 amino acid residues) and β (with 146 residues). Both subunit types exhibit a high degree of α-helical content with no β structure, and each contains a noncovalently associated b-type Fe heme to which O2 binds. As a tetramer, the four subunits are organized structurally as two αβ dimers held together by a polar, water-filled dimer–dimer interface (Fig. 1). Although the dimer–dimer interface dissociates readily under physiologic conditions to produce free αβ dimers, the intradimer interface is hydrophobic and only dissociates appreciably in the presence of certain metal ions or under denaturing conditions. Therefore, the αβ dimers are shared constantly and redistributed among the tetramers. When the Hb tetramer binds O2, a large change in quaternary structure is observed in which the two αβ dimers reorient relative to one another. From the deoxy or T structure, this reorientation can occur in either one of two major forms, which yields the R or the R2 structure. The R structure is observed by crystallization of oxyHb under high salt conditions, whereas the R2 structure is observed in low salt crystals. Nuclear magnetic resonance analysis has demonstrated that the R and R2 structures can coexist in solution (3). Additional crystal structure conditions have revealed that multiple oxy or R as well as deoxy or T forms are possible (6–8). All structural forms of the tetramer are ligated symmetrically (or unligated); i.e., the two dimers within the tetramer are always observed as structurally equivalent in available crystal structures. Structural changes that take place in the αβ dimers themselves are referred to in the Hb literature as “tertiary” and include the movement of the heme Fe into the plane of the heme when oxygenation begins; subsequent movement of helices are close to the heme and to the dimer–dimer interface. A significant structural change in the intradimer interface is not observed in crystal structures, which has led to the conclusion that oxygenation-induced tertiary structural changes are not communicated between the subunits within a dimer, i.e., between α1 and β1 or between α2 and β2. Although this belief has spanned the course of several decades, more subtle structural changes in the intradimer interface have not been ruled out. Recently, A one et al. have pointed out that the intradimer structure has not been analyzed thoroughly in modern crystal structures of Hb.

Figure 1. The cascade of O2 binding to the four subunits of the human Hb tetramer. The polar dimer–dimer interface is composed of J1 plus J2 plus J3 contacts. Two intradimer interfaces exist, the a1b1 and the a2b2, both are nonpolar. Each tetramer is assigned a species designation, which begins with deoxy Hb (species 01) and ending with oxy Hb (species 41). The first O2 can bind to any one of four subunits; however, because oxygenation of the α1 subunit is indistinguishable from oxygenation of the α2 subunit, the two isomeric tetramer species that result are designated “11a” and “11b.” This labeling is likewise the case for the β subunits. Similar isomeric oxygenation microstate tetramers are also generated in the second and third binding steps. Crystal structures are from the Arnone laboratory 14, 15.
Very few attempts to crystallize the partially ligated intermediates of Hb have been reported, and structural information is still not available. When O2 binding occurs, the quaternary reorientation of the two dimers forms the primary basis for the popular two-state model of Hb cooperativity. In this model, O2 binding to a heme Fe causes significant structural change only in the dimer-dimer interface. The bonds of the deoxy or T interface, held by the model to be significantly weaker than those of the oxy or R interface, maintain the tetramer in the low-affinity conformation. Oxygenation of deoxy Hb causes bonds in the T dimer-dimer interface to break, which weakens the low-affinity T state relative to the high-affinity R state. Therefore, the O2 affinity of the tetramer is controlled by the strength of the dimer-dimer interface, as modulated by the number of O2 ligands bound to the subunits. The particular configuration of the bound O2 among the four hemepools is not significant in this model. For example, in a tetramer that bears two bound O2, six possible configurations of the bound ligands exist among the four hemepools (see Fig. 1). In the two-state model, all six of the doubly ligated tetramers have the same O2 affinity because differences between the two ligands results in the same number of bonds broken in the dimer-dimer interface, regardless of the exact distribution of the ligands among the four hemepools.

**Relationship of macroscopic to microscopic binding constants**

Oxygen-binding curves can be analyzed directly to yield four macroscopic binding constants: $K_5$, $K_6$, $K_7$, and $K_8$. Usually, the microscopic constants are defined as product constants, i.e., products of the microscopic constant $K_{i,j,k,l}$ (where $i$ is the number of bound O2):

$$K_1 = K_{0,1} = K_{0,1}$$
$$K_2 = K_{0,2} = (K_{0,2})(K_{0,1})$$
$$K_3 = K_{2,0} = (K_{0,2})(K_{0,2})(K_{1,2})$$
$$K_4 = K_{2,0,2} = (K_{0,2})(K_{2,0})(K_{2,2})(K_{2,4})$$

The constant $K_{i,j,k,l}$ is composed of microscopic constants, as each O2 binding step is composed of multiple microscopic reactions, which is illustrated by the reaction arrows in Fig. 1. Thus, $4$ ways exist to bind the first O2, $12$ ways to bind the second O2, $12$ ways to bind the third O2, and $4$ ways to bind the fourth O2. Each microscopic constant is designated by the notation $i,j,k,l$ or the species formed in the binding process (Fig. 1, Table 1). For each binding step $i = 1,2,3$, and $4$, the microscopic constant $K_{i,j,k,l}$ is defined. For example, the reaction arrows in Fig. 1 illustrate the binding process.

**Chemistry**

The dimer-tetramer assembly constant $K_{0,0}$ is very sensitive to O2 binding by human Hb, which ranges over 130,000-fold among the intermediates, in comparison with the 400-fold change in O2 binding constant under conditions of this study (pH 7.4, 21.5°C) (9). The equilibrium between free dimer and assembled tetramer is an integral property of Hb in solution, and has a marked impact on the O2 binding curve observed experimentally.

**The concentration-dependent oxygen isotherm**

The sigmoidal shape of the O2 binding isotherm, i.e., the cooperativity of O2 binding, is dependent on the concentration of the Hb solution (Fig. 2). As the solution is diluted, the relative concentration of free dimer increases, and unlike the tetramer, the free dimer binds O2 noncooperatively with high affinity. Thus, the true tetramer-binding curve is observed only at the highest Hb concentrations. At lower concentrations, the experimental isotherm reports a mixture of tetramer and free dimer (1). The thermodynamic scheme that links the ligation of the free dimer to the ligation of the assembled tetramer (Fig. 2) shows all reaction equilibria that contribute to the concentration-dependent isotherms. Binding to the free dimer is designated by $\Delta G_{exo}$, which denotes the change in intrinsic (noncooperative) free energy. To solve the linkage scheme experimentally, the assembly free energy change for the deoxy tetramer, $\Delta G_{as}$, is determined in an independent kinetic measurement using haptoglobin trapping of the free dimer. In addition, the correspond-
The macroscopic thermodynamic linkage scheme

The concentration dependence of the O₂ binding curve is a result of thermodynamic linkage between O₂ binding and dimer → tetramer assembly. Consider the first binding step as illustrated in the linkage scheme in Fig. 2. Conservation of free energy dictates that the change in free energy during assembly followed by ligand must equal the change in free energy during ligand followed by assembly:

\[ \Delta G_{\text{asm}} + \Delta G_1 = \Delta G_{\text{tet}} + \Delta G_{\text{asm}} \]  (3)

Therefore, the change in the tetramer assembly free energy during O₂ binding is equal to the change in the O₂ binding constant during tetramer assembly:

\[ 1\Delta G_{\text{asm}} - 2\Delta G_{\text{asm}} = \Delta G_1 - \Delta G_{\text{tet}} \]  (4)

Each stepwise microscopic binding reaction follows the same formula, as thermodynamic linkage holds for all binding steps.

Forming hybrid tetramers from parent tetramers

The equilibrium between free dimer and tetramer can be exploited to provide a means of forming partially ligated hybrid tetramers by mixing any two parent tetramers. However, it is necessary to fix the hemestesite ligand to prevent disproportionation caused by ligand rearrangement among the hemestesites. Hemestesite analogs employed for either the deoxy heme (which replaces Fe^II) or the oxy heme (which replaces Fe^III) O₂ are:

<table>
<thead>
<tr>
<th>Native</th>
<th>A analog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fe^II/Fe^II/Fe^II/O₂</td>
<td>Zn^II/Fe^II/Fe^II/O₂</td>
</tr>
<tr>
<td>Fe^II/Fe^II/CN</td>
<td>Fe^II/Fe^II/CN</td>
</tr>
<tr>
<td>Co^III/Fe^III/Co</td>
<td>Co^III/Fe^III/Co</td>
</tr>
<tr>
<td>Zn^II/CN</td>
<td>Fe^II/Mn^III</td>
</tr>
</tbody>
</table>

Each hemestesite analog perturbs the Hb tetramer in some manner. The Fe^II/CN and Mn^III analogs are susceptible to electron exchange over very long incubation periods (10). The Co^III analog exerts a specific effect on α-subunit binding constants (11), and the use of Zn^II imparts a light sensitivity to the solution (12). However, the relative relationship between each measured microstate-binding constant is found to be invariant among the analog species (9).

Using the Zn^II/Fe^II/O₂ analog as an example, when deoxy ZnHb (species 01) is mixed with an equimolar amount of native FeHb (species 41), a mixture is formed that contains the asymmetrically doubly ligated species 21 (Fig. 3A). Likewise, species 11 or 12 are formed by mixing species 01 with 23 or 24, respectively (see Fig. 1 for illustrations of each species). Species 22 is formed by mixing species 23 with 24.

### Table 1: Relationship between macroscopic and microscopic O₂ binding constants

<table>
<thead>
<tr>
<th>Overall binding step</th>
<th>Microstate binding step</th>
<th>Average microstate binding constants</th>
<th>Macro Kᵢ₋→₁⁺</th>
<th>Micro kᵢ₋→₀⁺⁺⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 → 1</td>
<td>01 → 11</td>
<td>k₉₁₋→₁¹ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td>k₀₋₁ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>01 → 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 → 21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 → 2</td>
<td>12 → 21</td>
<td>kₙ₋₁₋→₁²₁ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td>K₁₋₂ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 → 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 → 23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 → 3</td>
<td>22 → 31</td>
<td>kₙ₋₁₋→₂²₁ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td>k₂₋₃ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>21 → 32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 → 4</td>
<td>31 → 41</td>
<td>kₙ₋₁₋→₃²₁ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td>k₃₋₄ = ( \frac{K_{11} + K_{12}}{2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>32 → 41</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each binding step is composed of multiple microstate binding steps, and each microstate binding step represents an average of isomeric forms of the microstates, yielding the average microstate binding constants. The macrostate binding constant is then the average of the binding constant for each microstate.
species 31 or 32 are formed by mixing species 41 with 24 or 23, respectively. In this way, all possible combinatorial forms of the partially ligated intermediates can be formed. Only the parent tetramers 01, 23, 24, and 41 are present in pure form in solution: All other species are present in equilibrium with their respective parent tetramers. A assembly of tetramers from free dimers occurs very rapidly with a rate constant \( k_{on} \) of \( 1.1 \times 10^7 \text{M}^{-1}\text{s}^{-1} \). This assembly is referred to as the “consensus on constant,” which is not dependent on the number of bound ligands or their configuration, the presence of hemestie analogs, or the presence of mutations. Therefore, it is in the tetramer->free dimer dissociation constant, \( k_{off} \), that the sensitivity of the assembly constant is manifest, because

\[
\Delta G_{asm} = -RT \ln \left( \frac{k_{on}}{k_{off}} \right)
\]

**Key Experiments and Observations**

Two experimental approaches are taken to measure the assembly free energy of partially ligated Hb intermediates: an equilibrium method and a kinetic protocol. In the equilibrium method, symmetrically ligated tetramers are mixed to generate asymmetrically ligated hybrid tetramers. Then, the relative stability of the hybrid to its parents is measured, which permits the hybrid assembly free energy to be calculated from the independently measured \( \Delta G_{asm} \) of the parents. In the kinetic approach, the tetramer->dimer dissociation constant is measured by trapping free dimers kinetically with the plasma protein haptoglobin.

**Assembly free energy of hybrid tetramers**

**Low-temperature isoelectric focusing**

Species 21 represents a unique halfway point in oxygenation of Hb in that one of its dimers is fully ligated and the other is fully deoxygenated. The 21 hybrid is formed in vitro by mixing species 01 and 41, as in the example in Fig 3. One of the two parent Hbs carries an electrophoretic tag to enhance separation based on charge, typically the HbS variant (Hb Glu→Val). At equilibrium, the assembly free energy of species 21 is related to the assembly free energies of each parent by

\[
\Delta G_{asm} = \frac{1}{2} (\Delta G_{asm}^{41} + \Delta G_{asm}^{21}) + b_{21} \tag{6}
\]

where \( b_{21} \) is the free energy deviation from the average of the parent tetramer assembly free energies. The deviation free energy is measured directly from the fraction of each tetramer at equilibrium:

\[
b_{21} = -RT \ln \left( \frac{f_{21}}{f_{41}} \right) \tag{7}
\]

The relative fractions of hybrid and parent tetramers are measured by quenching the dissociation of tetramer to free...
dimer dissociation constant, \( k_{\text{off}} \), to the assembled tetramer. The tetramer dissociation reaction to the right:

\[
\text{Hb} \xrightarrow{k_{\text{off}}} 2\text{a} \beta \text{dimers} + \text{Hp} \xrightarrow{k_{\text{kon}}} \text{Hb}_2 + \text{Hp} \text{(dimers)}_2
\]

Because both \( k_{\text{kon}} \) and \( k_{\text{off}} \) are very rapid (essentially diffusion-controlled) processes, the overall rate-limiting step in Equation 8 is \( k_{\text{kon}} \).

The \( k_{\text{off}} \) for the asymmetrical ligated species 21 was measured first by forming the unligated version of the hybrid by mixing native Fe-heme deoxy Hb with 2\(n\) Hp (Fig. 4). Because both parents and hybrid (the unligated species 21 or 21L) have the same assembly free energy, \( \Delta G_{\text{asm}} \), mixing the parent tetramers in a 1:1 ratio generates an equilibrium hybrid mixture that contains approximately 50% hybrid 21L. Because of the slow \( k_{\text{off}} \) for both parents and hybrid (7.5 hours), equilibrium is attained after 3 days of incubation. In practice, the amount of hybrid present after 24 hours is sufficient for detection in the reaction with Hp.

The anaerobic hybrid mixture is mixed with an oxygenated solution of Hp in a stopped-flow instrument. The absorbance is monitored for 20 seconds, and the resulting observed rate constant is measured at 0.20 ± 0.02 s⁻¹. This results in an assembly free energy, \( \Delta G_{\text{asm}} \), of –0.3 ± 0.1 kcal/mol when combined with the consensus on constant in Equation 6 (9).

This value is in excellent agreement with the results of the equilibrium low-temperature isoelectric focusing experiment.

Model-independent distribution of \( \Delta G_c \) among the hemoglobin intermediates

The measurements described here show that binding \( O_2 \) ligands to only one \( \alpha \beta \) dimer within the Hb tetramer occurs with a
An extensive kinetic analysis of ligand binding in normal Hb carried out by Goldbeck et al. has demonstrated agreement with the unique binding constant for the asymmetric doubly ligated Hb (14). Thermodynamic experiments from the Ackers laboratory that employs asymmetrical modified human Hbs have confirmed the asymmetric character of Hb cooperativity (15). This discovery generates critical energetic and structural questions, particularly with respect to the relationship of interdimer to cross-dimer cooperativity, in a classic system that was once thought to be well understood.

References

Human Hemoglobin: Identification of a Key Intermediate

Table 2 The distribution of binding free energy among the individual binding steps in Hb

<table>
<thead>
<tr>
<th>Binding steps</th>
<th>Stepwise binding free energy, $\Delta G_{i \rightarrow i+1}$ kcal/mol</th>
<th>Stepwise microscopic binding constant, $k_{i \rightarrow i+1}$ M$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygenation through the asymmetric doubly ligated tetramer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 01 + O$_2$ → 11 or 12</td>
<td>$-5.5 \pm 0.3$</td>
<td>1 e + 4</td>
</tr>
<tr>
<td>2 (11 or 12) + O$_2$ → 21</td>
<td>$-6.6 \pm 0.4$</td>
<td>3 e + 4</td>
</tr>
<tr>
<td>3 21 + O$_2$ → 31 or 32</td>
<td>$-6.6 \pm 0.4$</td>
<td>8 e + 4</td>
</tr>
<tr>
<td>4 (31 or 32) + O$_2$ → 41</td>
<td>$-8.9 \pm 0.3$</td>
<td>400 e + 4</td>
</tr>
<tr>
<td>Oxygenation through the symmetric doubly ligated tetramer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 01 + O$_2$ → 11 or 12</td>
<td>$-5.5 \pm 0.3$</td>
<td>1 e + 4</td>
</tr>
<tr>
<td>2 (11 or 12) + O$_2$ → 22,23 or 24</td>
<td>$-4.8 \pm 0.5$</td>
<td>1 e + 4</td>
</tr>
<tr>
<td>3 (22,23 or 24) + O$_2$ → 31 or 32</td>
<td>$-7.9 \pm 0.5$</td>
<td>70 e + 4</td>
</tr>
<tr>
<td>4 (31 or 32) + O$_2$ → 41</td>
<td>$-8.9 \pm 0.3$</td>
<td>400 e + 4</td>
</tr>
</tbody>
</table>

The binding cascade shown in Fig. (2) is essentially composed of two different pathways, passing through either an asymmetrically ligated species at the second binding step or a symmetrically ligated species.


Further Reading


See Also

Bioenergetics of Self-Assembly
Energetics of Protein Folding
Ligand-Operated Membrane Channels
Protein-Protein Interactions
Thermodynamics in Living Systems
Metal Homeostasis
F. Wayne Outten and Benjamin S. Twining,
University of South Carolina, Columbia, South Carolina
doi: 10.1002/9780470048672.wecb324

Transition metals are a key component of biological systems. Because of their special properties, they are incorporated into proteins functioning in dioxygen transport, electron transfer, redox transformations, and regulatory control. The metals used in biological systems have been selected throughout evolution based on their availability in the environment and their kinetic lability, resulting in preferential use of first-row transition metals in biology. These essential metals must be obtained from the environment and concentrated within the cell for use in biochemical pathways. Once appropriated, metals must be directed to metalloenzymes or metal storage proteins within the cell. In addition, organisms must be able to distinguish between essential and toxic metals and must have mechanisms for minimizing the toxicity of both essential and toxic metals that are present in excess. Metal homeostasis is broadly defined as the metal uptake, trafficking, efflux, and sensing pathways that allow organisms to maintain an appropriate (often narrow) intracellular concentration range of essential transition metals. This review will introduce several unifying concepts of metal homeostasis with brief illustrative examples for each concept.

Transition metals are key nutrients for nearly all organisms because of their role in critical biochemical pathways such as respiration, photosynthesis, and nitrogen fixation. However, metals cannot be synthesized like other biomolecules and must be obtained from the environment. Once acquired by organisms, metals must be routed to the correct intracellular destination while preventing deleterious side reactions or nonspecific chelation by other cellular components. Metals that are difficult to obtain may be stored for future use by the organism. As intracellular metal concentrations increase, organisms must have the ability to remove (efflux), sequester, or detoxify the excess metal. Finally, organisms must have some ability to distinguish between essential and nonessential metals, despite their similarities, in order to prevent poisoning by nonessential metals. The genetic and biochemical pathways that are used by organisms to acquire, traffic, store, and detoxify metals are collectively known as metal homeostasis systems. Maintenance of intracellular transition metal concentrations within an optimal range has posed a major challenge for biological systems throughout evolution. The study of metal homeostasis has been a key part of the field of bioinorganic chemistry since its inception, providing numerous insights into how transition metals are integrated into biological systems.

Concepts in Metal Homeostasis

There are many functions for transition metals within biological systems (Table 1). The choice of metal used for each function is determined by the chemical characteristics of that metal, including its size as well as its thermodynamic stability and kinetic lability when complexed with biological ligands. Biological systems have mostly incorporated the first-row transition metals. The divalent forms of these metals are particularly common in biology because they have higher ligand exchange rates than their M³⁺ counterparts. Kinetically labile transition metals are required to allow assembly and disassembly of metal centers and for rapid binding and release of substrates in metal-catalyzed reactions. The relative abundance and availability of the transition metals in the environment has also dictated their use in biology because they have higher ligand exchange rates than their M³⁺ counterparts. Kinetically labile transition metals are required to allow assembly and disassembly of metal centers and for rapid binding and release of substrates in metal-catalyzed reactions. The relative abundance and availability of the transition metals in the environment has also dictated their use in biology. An excellent discussion of these issues in the context of metal selection for metalloenzyme use is presented in the technical article "Chemistry of Metalloenzymes" by R.J.P. Williams.

To ensure an ample supply of transition metals for incorporation into biomolecules, organisms have been selected to contain metal homeostasis systems.

Note: Although zinc is not considered a transition metal based on the IUPAC definition, zinc has often been included with the transition metals as distinct from the alkali and alkaline earth metals. For the purposes of this review, we will consider zinc as part of the transition metal group with the caveat that it does not strictly meet the IUPAC definition.
**Metal Homeostasis**

**Table 1** Environmental and intracellular metal concentrations and select biological functions for essential metals

<table>
<thead>
<tr>
<th>Metal</th>
<th>Environmental concentrations&lt;sup&gt;a&lt;/sup&gt; (M)</th>
<th>Cellular concentrations&lt;sup&gt;b&lt;/sup&gt; (M)</th>
<th>Examples of select biological roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>$10^{-8}$</td>
<td>$1.4-6.9 \times 10^{-7}$</td>
<td>Nitrogen fixation</td>
</tr>
<tr>
<td>Cr</td>
<td>$10^{-9}$</td>
<td>$1.7 \times 10^{-4}$</td>
<td>Insulin signaling</td>
</tr>
<tr>
<td>Mn</td>
<td>$10^{-6}-10^{-10}$</td>
<td>$2.4 \times 10^{-6}, 1.0 \times 10^{-4}$</td>
<td>Photosynthesis</td>
</tr>
<tr>
<td>Fe</td>
<td>$10^{-5}, 10^{-12}$</td>
<td>$1.6-5.0 \times 10^{-4}$</td>
<td>Dioxygen transport, electron transfer, nitrogen fixation</td>
</tr>
<tr>
<td>Co</td>
<td>$10^{-6}, 10^{-11}$</td>
<td>$1.0 \times 10^{-5}$</td>
<td>Alkyl group transfer</td>
</tr>
<tr>
<td>Ni</td>
<td>$10^{-6}-10^{-8}$</td>
<td>$1.0 \times 10^{-4}$</td>
<td>Hydrogenase, hydrolyase</td>
</tr>
<tr>
<td>Cu</td>
<td>$10^{-7}-10^{-9}$</td>
<td>$0.7-4.3 \times 10^{-5}$</td>
<td>Dioxygen transport, electron transfer</td>
</tr>
<tr>
<td>Zn</td>
<td>$10^{-9}, 10^{-10}$</td>
<td>$1.0-1.9 \times 10^{-4}$</td>
<td>Structural stabilization, hydrolyase</td>
</tr>
<tr>
<td>Mo</td>
<td>$10^{-7}-10^{-8}$</td>
<td>$3.0-4.8 \times 10^{-6}$</td>
<td>Nitrogen fixation, oxo transfer</td>
</tr>
<tr>
<td>Cd</td>
<td>$10^{-10}-10^{-12}$</td>
<td>$1.0 \times 10^{-5}$</td>
<td>Carbonic anhydrase</td>
</tr>
<tr>
<td>W</td>
<td>$10^{-11}$</td>
<td>n.d.</td>
<td>Carbon dioxide reduction/fixation</td>
</tr>
</tbody>
</table>

<sup>a</sup>Range shown is for both ocean and freshwater aquatic systems.


**The metal quota**

A key concept in metal homeostasis is the "metal quota." This quota is simply the amount of a given metal required for normal cellular function under a specific growth condition. In theory the metal quota can vary considerably. For example, metal requirements may change in response to growth stage (quiescence or active growth), oxygen availability, carbon source, and a variety of other environmental factors. Similarly, different organisms or different cell types within an organism may have different metal quotas (1). For example, photosynthetic cyanobacteria have been reported to have higher Mn and Fe quotas than other nonphotosynthetic prokaryotes (2). In this case, the higher metal quotas stem directly from increased amounts of specific metalloproteins (Photosystem I and Photosystem II). At the other extreme, the obligate parasite that causes Lyme disease, Borrelia burgdorferi, seems to have no requirement for Fe and lacks many metalloproteins found in other microbes (3). These unusual adaptations may allow B. burgdorferi to survive in the iron-limited environment within the host. In general, however, direct measurement of cellular metal concentrations has shown remarkably small differences between widely divergent organisms, often less than an order of magnitude (Table 1). Thus there does seem to be a well-conserved optimum range of intracellular metal concentrations for most essential metals. The quota of a specific metal is largely dictated by the concentration of cellular proteins that require the metal for their function and by the importance of the metalloproteins (i.e., does the metalloprotein play a critical role or can another protein functionally substitute?). However, metal bioavailability and metal toxicity also strongly influence the metal quota and the nature of each metal-specific homeostasis system.

**Metal bioavailability**

A critical factor in metal homeostasis is the bioavailability of the specific metal. The "bioavailability" of a metal can be distinct from the overall abundance of the metal in the environment. For example, iron is the fourth-most abundant element in the Earth’s crust. However, in the current-day oxygen-rich atmosphere of Earth, iron is largely present in the ferric (Fe<sup>3+</sup>) form. In aqueous, aerobic environments at neutral or basic pH, ferric iron forms nearly insoluble iron hydroxides. Because of the insolubility of Fe<sup>3+</sup>, iron is one of the least bioavailable of the essential transition metals. The concentration of iron in seawater (3 × 10<sup>-5</sup> ppm) is nine orders of magnitude lower than the crustal concentration (5.10<sup>10</sup> ppm). In contrast, zinc (Zn<sup>2+</sup>) is present at only 70 ppm in the crust but is found at 1 × 10<sup>-3</sup> ppm in seawater (4, 5). The increased solubility of zinc means that it is actually more bioavailable than iron, despite its overall lower abundance. As discussed, iron homeostasis requires strategies for mobilizing iron in the environment, whereas homeostasis of metals such as zinc largely begins with cellular uptake.

The bioavailability of a given metal is influenced by its chemical speciation in the ambient environment. Although some metals occur predominantly in their "free" or "aquo" form (that is, the inner coordination sphere of the metal ion is occupied solely by water molecules), most bioactive metals occur as complexes in the natural environment. Hydroxide, carbonate, and chloride anions can all bind transition metals to a significant extent. For example, Fe<sup>2+</sup> forms hydroxide complexes.
complexes [e.g., Fe(OH)$_3^+$ or Fe(OH)$_2$] in natural waters at pH > 7 and Cu tends to form carbonate complexes (CuCO$_3^-$) under similar conditions. Metals may also bind to organic molecules, which may be expressly produced by resident biota to influence metal availability (such as siderophores for iron) or result from the degradation of cells. It has been shown that most iron, copper, cobalt, and zinc ions in the marine environment are bound to unidentified organic molecules (6-9). Metals in terrestrial, freshwater, and near-shore environments are often bound to humic and fulvic acids produced via decomposition of terrestrial organic matter (10). As a general rule, only the free metal ion can react with uptake or transport proteins (11). However, the ligand exchange kinetics of most inorganic complexes are fast enough that these forms are often considered bioavailable as well. Organic complexes are often not immediately available (except in some cases such as lipophilic complexes); however, some cells have evolved specific biochemical mechanisms for obtaining required metals from organic ligands. For example, iron-siderophore complexes may be directly transported into the cell, and iron bound to nonspecific ligands may be obtained with ferric chelate reductases that reduce Fe$^{3+}$ to Fe$^{2+}$ for subsequent transport across the cell membrane (12).

In environments where metal bioavailability is limited by biological or geochemical factors, normal cellular growth and functioning may be impaired. For example, nearly 40% of the limited availability of iron is a fascinating and fully developed field. Readers are referred to the following sections for more information on this topic.

The "labile metal pool"

It is currently routine to quantify the total metal content of cultured cells or organisms using sensitive analytical techniques. For example, inductively coupled plasma mass spectrometry (ICP-MS) enables simultaneous detection of most transition metals in bulk cell culture samples at parts per trillion concentrations, and high-resolution magnetic sector instruments are capable of separating all transition metals from common matrix plasma interferences. These approaches have been used to establish the metal quota under various conditions in various organisms (17, 18). Most of any given metal within the cell will be stably incorporated into metalloenzymes and storage proteins. However, a small but physiologically critical pool of the metal has to be available for incorporation into newly synthesized metalloproteins. This "labile pool" of metal is not "free" in the chemical sense as the metal is likely bound to metallochaperones or other carrier proteins or to small molecules like citrate or glutathione. The labile metal pool is an important aspect of homeostasis because intracellular metal sensors that regulate
Metal Homeostasis

Metal homeostasis likely respond to changes in this subpopulation of metal. Also, the labile metal pool is more likely to undergo spurious side reactions under adverse conditions than metal safely incorporated into metalloenzymes or storage proteins. Thus it is critical to define the concentration of the labile metal pool and to understand how it is maintained as part of overall metal homeostasis. As is evident in the “Chemical Tools and Techniques” section, it is not trivial to establish a value for the labile metal pool.

Cellular Mechanisms to Maintain Metal Homeostasis

**Figure 1** shows a generic scheme for a typical metal homeostasis system. We now will consider each component in detail.

**Acquisition of metals from the environment**

Some metals may need to be mobilized from the environment to make them bioavailable. Iron in particular must be rendered more soluble to be accessible for uptake. Microorganisms and some plants have evolved with secreted ligands known as siderophores (or phyto-iron transport proteins). These ligands bind Fe³⁺ with extraordinary affinity. For example, a complex of the siderophore enterobactin with ferric iron has a formal stability constant of 10⁴⁹ (19). Once siderophores compete with other environmental ligands for iron, the ferric iron–siderophore complex then binds to specific transport proteins at the microbial cell surface and is taken into the cell. Most microorganisms can synthesize or use multiple siderophores as a source of iron. A similar strategy is used to move iron through the blood of multicellular organisms like mammals. Transferrin is an iron-binding plasma protein that preferentially binds ferric iron. Circulating Fe³⁺–transferrin complex is recognized by cell-surface receptors for uptake via receptor-mediated endocytosis (20). Because of tight regulation of iron transport throughout the body, pathogenic microbes are faced with the same problem of limited iron availability as their microbial counterparts in other environments. In response to this pressure, many microbial pathogens contain transporters and enzymes that allow them to obtain iron from host sources, such as hemoglobin and transferrin (21). Some plants use an alternative strategy for iron acquisition from the environment. Plant cells in the roots (where most iron is obtained) excrete protons in order to acidify the soil. Lowering the pH in the microenvironment around the roots increases the solubility of iron and allows it to be acquired by transporters (16). A similar pH-dependent process allows mammals to render dietary iron more soluble during digestion. In addition, some organisms contain ferric reductases that reduce ferric iron to the more soluble ferrous form. Reduction of ferric iron can occur at the cell surface or within the cytoplasm to release iron from ferric chelates like siderophores (22).

**Concentrating metals within cells**

Once the metal has been removed from the environment, it must be transported into the cell for use. In most cases, environmental metal levels are significantly lower than the cellular metal quota (Table 1). This dichotomy requires organisms to concentrate...
metals inside cells via energy-dependent processes. Transport of metals across the lipid bilayer by transport proteins is often linked to ATP hydrolysis or to the proton motive force (PMF) in order to provide the energy for concentrating metals. In some cases, the soluble metal is directly transported into the cell. In other cases (such as iron), metals complexed with acquisition molecules are transported. Once inside the cell, these complexes are disrupted to release the metal. In addition to transmembrane transporters, uptake of some metals requires other accessory proteins. For example, ferrous iron must be oxidized by a multi-copper oxidase enzyme at the cell surface in order to be transported by some eukaryotic transporters (23).

Intracellular metal trafficking

Once concentrated within the cell, specific metals must be routed to the proper metalloprotein. This problem is not trivial since many biological ligands are capable of binding essential metals. Nonspecific interaction of metals with inappropriate ligands could prevent the metal from reaching the necessary target metalloprotein. Spurious side reaction of metals with other molecules in the cell may also generate free radicals. For instance, the Fenton reaction between Fe^{2+} with H_2O_2, which is a normal byproduct of aerobic respiration, can generate highly dangerous hydroxyl radicals leading to cell damage. The routing of a metal to its correct target protein may occur as a result of the affinity of the metalloprotein for its specific metal. However, this mechanism seems inadequate as the intracellular concentration of a specific metalloprotein may be quite low compared with the concentration of competing, nonspecific ligands such as glutathione, citrate, or nucleic acids. In addition, the presence of multiple transition metals within the cell would make it difficult for a metalloprotein to partition a specific metal in a single step. It has become clear that several metalloproteins require a metallochaperone carrier protein to donate the correct metal to their active sites. For example, the Cu-Zn superoxide dismutase (SOD1) enzyme requires a copper metallochaperone, known as CCS, to donate copper for assembly of its active site (24).

Metal efflux and detoxification

If metals accumulate to high intracellular concentrations (for example, because of environmental excess), organisms must have homeostasis mechanisms for removing, sequestering, or detoxifying the metals. In the case of toxic metals such as silver, removal of the metal via efflux transporters may be sufficient. Increased expression of metal storage proteins or novel metal sequestration proteins can also protect the cell from excess metal. This strategy seems to predominate for essential metals under most conditions. For example, the model organism Escherichia coli contains multiple ferritin homologues that are differentially regulated to provide excess storage capacity for iron under adverse conditions (28). In eukaryotes and some prokaryotes, the cysteine-rich metallothionein protein family functions to store or buffer excess copper and zinc (27).

Regulation of metal homeostasis

Expression of the systems mentioned must be carefully coordinated within the cell in order to maintain metal homeostasis. Consequently most organisms contain metalloregulatory proteins to regulate metal homeostasis. Metalloregulatory proteins are transcription factors that sense cellular metal levels and either activate or repress transcription of metal homeostasis genes in response to changes in metal levels. Typically the metal in question binds directly to the cognate metalloregulatory protein and acts as an allosteric switch for activating or inhibiting the transcription factor. In addition to regulation at the transcriptional level, some metal homeostasis systems are also regulated post-transcriptionally. For example, bacteria and mammals regulate iron homeostasis at the mRNA level. In bacteria, a small regulatory RNA, RyhB, binds target mRNAs involved in iron homeostasis and metabolism and regulates their stability or their translation. The ryhB gene itself is under the control of a metalloregulatory protein, Fur, that senses cellular iron (29).
helped to identify new genes that encode metal homeostasis to characterize metal responsive gene transcription and have been used to isolate novel antioxidant factors. In addition, gene resulting phenotypes. For example, the Atx1 copper metallochaperone was originally identified in yeast as part of a genetic screen to isolate novel antioxidant factors. In addition, gene reporter constructs and DNA microarray analysis have been used to characterize metal responsive gene transcription and have helped to identify new genes that encode metal homeostasis proteins. Global transcriptional analysis has also led to a deeper understanding of how perturbations in metal homeostasis impact other cellular pathways. For example, DNA microarray analysis in E. coli has revealed that extensive remodeling of iron metalloproteins occurs when cells shift from an iron-rich environment to an iron-poor environment. Overall iron metalloprotein content is reduced in order to conserve available iron. However, this remodeling has profound effects on cellular metabolism since many metabolic proteins contain iron in the form of Fe-S clusters or heme. Similar processes have also been observed in eukaryotic organisms. Thus, genetic techniques are allowing investigators to determine how metal homeostasis is integrated with overall cellular metabolism and physiology.

Chemical Tools and Techniques

Genetics

Genetic approaches have been used with considerable success to study metal homeostasis in genetically tractable model organisms like E. coli and S. cerevisiae. These approaches involve deleting genes that encode metal homeostasis proteins (such as metal transporters or metallochaperones) and analyzing the resulting phenotypes. For example, the A31 copper metallochaperone is a small, representative sample from a large body of literature). Herein are only a small, representative sample from a large body of literature). Herein are only a small, representative sample from a large body of literature).

Biochemical characterization of metal homeostasis components

Purification and in vitro characterization of metal homeostasis proteins has provided a wealth of information about protein function. Specific metal binding sites have been characterized in metal transporters, metallochaperones, and metalloregulatory proteins using physical inorganic techniques such as extended x-ray absorption fine structure (EXAFS) spectroscopy, Mossbauer spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and electroparamagnetic resonance (EPR) spectroscopy. For example, spectroscopic analysis of the C001–A31 complex revealed the coordination chemistry of the copper binding site and led to a chemical exchange model for step-by-step copper donation from A31 to the target metalloprotein C002. Three-dimensional crystal structures of metal homeostasis components have also given researchers insight into the mechanisms of homeostasis. Proteins involved in metal transport and storage proteins by directly binding to 5’ or 3’ untranslated regions within the mRNA and altering translation initiation or mRNA stability. IRP directly senses cellular iron levels by virtue of an Fe-S cluster present in the protein. A detailed discussion of these metalloregulatory proteins is presented in the technical article “Metalloregulatory Proteins” by Deborah Zamble.

Measurement of the “labile metal pool”

As mentioned, accurate determination of the “labile metal pool” is a key goal of metal homeostasis research. Most recent efforts have focused on measuring the labile metal pool in yeast without disrupting the cell. This trend toward nondisruptive approaches stemmed from the realization that metal localization and speciation could be greatly perturbed by diluting cellular reductants, exposing the intracellular milieu to oxygen, disrupting subcellular organellae, and other adverse consequences of cell breakage. The general approach to define the labile metal pool relies on chelation of the labile metal pool in situ followed by detection and measurement of the chelator-metal complex. For example, to measure labile iron in E. coli, the ferric iron chelator desferrioxamine can be added to cells resulting in stabilization of all labile iron in the ferric form. The ferric iron can then be measured using EPR spectroscopy. Similarly, metal-specific fluorescent probes can be added to cells and the concentration of “available” metal measured by changes in fluorescent signal. Probes such as calcein and zinquin have been used to quantify the labile pools of iron and zinc, respectively. However, interpreting the results from these approaches can be difficult. If the metal binding affinity of the chelator or probe is sufficiently higher than the metal binding affinities of cellular metalloproteins, the probe may strip metal ions from metalloproteins and artificially increase the amount of “labile” metal (for a thorough discussion of troubleshooting the measurement of labile iron, see Reference 40). In addition, the cell membrane permeability, intracellular localization, and toxicity of the probe must be considered when evaluating its use.

To avoid these sorts of problems, some investigators have relied on the metal binding affinities of metalloproteins to estimate the labile metal concentration. For example, two zinc metalloregulatory proteins, Zur and ZntR, control expression of zinc homeostasis genes in E. coli by directly binding Zn to sense cellular zinc levels. Zn-Zur represses zinc uptake systems when cellular zinc levels are adequate, whereas Zn-ZntR activates zinc efflux systems when zinc levels rise too high. Thus, the two regulators are thought to sense zinc at the lower and upper concentration limits of the labile zinc pool in E. coli. Measuring zinc-dependent DNA binding by Zur and ZntR allowed the investigators to establish the threshold zinc concentrations for activation of each regulator. These values were then reported as the upper and lower concentration limits of the labile zinc pool in E. coli (18). This approach assumes that both metalloregulatory proteins are in equilibrium with the labile zinc pool. It is not clear if this assumption applies in vivo since regulation of homeostasis proteins at all levels (from gene transcription to protein degradation) necessarily requires a lag between metal sensing and the desired change in the activity of the target homeostasis components, such as transporters and storage proteins. This lag may prevent a true equilibrium from forming under typical conditions where metal availability is not constant. Intermediate
proteins may also act as carriers to load Zn into Zur or ZntR in vivo, which further complicates the labile zinc measurement since the additional protein–protein interactions required in vivo could significantly alter the zinc concentrations at which the regulators respond.

Measurements of labile iron range from about 0.9 to 12.3 \( \times 10^{-6} \) M in the cytoplasm of a range of organisms from bacteria to mammals (40, 41). In contrast, some approaches have estimated cytoplasmic pools of labile zinc and copper to be 10\(^{-15}\) and 10\(^{-18}\) M, respectively (18, 45), whereas other studies have measured labile zinc at around 5 \( \times 10^{-12} \) M (43). Although the absolute measurements vary somewhat, these studies clearly demonstrate that labile pools of copper and zinc are tightly controlled in vivo to avoid accumulation of these metals. The identity of the biological ligands that maintain such small labile pools of zinc and copper remains a topic of some controversy. As we can see, measurement of the labile metal pool is strongly influenced by the technique used and final determination of labile metal pools will likely require a convergence of approaches.

**In vivo localization of cellular metals**

The specific localization of transition metals within subcellular compartments has been characterized by a variety of methods. Subcellular organelles, such as mitochondria, purified by traditional centrifugation methods have been directly analyzed to identify subcellular populations of metals like copper (46, 47). Use of fluorescent probes coupled with microscopy allows investigators to visualize the subcellular location of metals, especially in eukaryotic cells because of their larger size. Studies of metal homeostasis in vivo have also benefited greatly from recent advances in analytical instrumentation. Subcellular mapping and quantification of transition metals within cells is possible using electron, proton, and X-ray microbeam techniques. Electron microprobes are capable of a spatial resolution of several nanometers (48) but require that cells be sectioned before analysis. Although widely available, electron microprobes also lack sensitivity to transition metals. Much higher sensitivity is possible when X rays are used to excite characteristic fluorescence, and impressive advances have been made with synchrotron X-ray fluorescence (SXRF) microprobes (49) (Fig. 2). Through the use of Fresnel zone plate optics at third-generation synchrotron facilities, spatial resolutions approaching 50 nm are now possible, and detection limits on the order of 10\(^{-18}\) mol per cell are routinely achieved for the transition metals (50). SXRF has been used to determine subcellular localization of transition metals in cardiomyocytes (52), human leukemia cells (52), and plant seeds (53). SXRF can also be combined with immunofluorescence probes (54) or metal-specific fluorescent sensors. For example, Yang et al. (55) visualized the localization of intracellular “labile” copper in the mitochondria and Golgi apparatus of...
mouse fibroblast cells using both fluorescent probes and X-ray analysis.

Future Research Directions

The metallome and metallome homeostasis

Initial study of metal homeostasis tended to focus on a single metal. However, proper functioning of most organisms requires careful balancing of multiple essential transition metals. The sum of all essential metals used by an organism for cellular function is known as the "metallome." The metallome consists of the total metal content of a cell but also includes all specific metal–biomolecule complexes that are present in a given cell. Characterizing this global entity is one of the most exciting future research directions in the field of metal homeostasis. Measuring the metal quota for each essential metal is still a key requirement for defining the metallome. Recently it has become possible to simultaneously measure all essential transition metals in an organism. For example, ICP-MS has been used to measure metal quotas in a diverse range of prokaryotic and eukaryotic organisms under controlled growth conditions. If the number of cells in the sample and the cell volume are known, ICP-MS measurements can be converted into total cellular concentrations to facilitate the comparison of metal quotas in organisms with a wide range of cell sizes. The power of approaches like ICP-MS is that they allow investigators to simultaneously measure cellular concentrations of essential transition metals in different organisms, under different environmental conditions, and in different genetic backgrounds to more fully define the metal quota. A further dimension of the metallome concept is the specific location of transition metals within the cell (as opposed to their total concentrations in the entire cell). In this context, localization can refer to the protein or biomolecule component where the metal is bound as well as the subcellular localization of the metal in various organelles. One approach to define metal localization will be to combine metal analysis techniques with proteomics in order to identify all metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57). Global analysis of the metalloprotein portion of the metallome will also lead to the identification of previously unknown metalloproteins, including their specific transition metal content, within an organism (56, 57).
Metal Homeostasis


Further Reading

See Also
Metal Transport Through Membranes
Chemistry of Metallochaperones
Chemistry of Metalloenzymes
Metalloregulatory Proteins
Metalloregulatory Proteins
Deborah B. Zamble, University of Toronto, Ontario, Canada
doi: 10.1002/9780470048672.wecb330

Metalloregulatory proteins are proteins that bind metals and modulate gene expression through direct interactions with DNA or RNA. The genes under this metal-dependent control encode a variety of proteins involved in the cellular homeostasis of both essential and toxic metals. Metalloregulators are present in all types of organisms, and extensive information exists about their mechanisms, although many unanswered questions remain. The global activities of these metal-responsive factors require overcoming complex challenges, such as the manner in which the proteins regulate gene expression, the mechanisms of the metal-dependent protein conformation transformations, and the ability of the proteins to recognize the designated metal(s). An understanding of these key biomolecules draws from disciplines such as cell biology, protein chemistry, and inorganic chemistry and provides molecular insight into one fundamental aspect of life.

Although a variety of biologic processes are regulated by metals, for the purposes of this review metalloregulators are defined as proteins that act as metal-responsive genetic switches (1). These proteins are sensors that monitor the cellular levels of one or more metal ions and then respond to changes in availability by modulating the expression of a variety of metal pathways. A few examples of metalloregulatory pathways exist in which the two responsibilities, sensing and regulating, are divided up between separate protein components. Learning how metalloregulators function entails defining the cell biology of what they do and the bioinorganic and bioorganic chemistry of how they do it. This article considers some of these aspects in general and then addresses them more specifically in the context of examples. The focus is on how these proteins bind and respond to metal ions. Other properties, such as the details of specific protein contacts with nucleic acids or RNA polymerase, will not be discussed. For more information on that area, the reader is referred to the article entitled “Transcription Factors.” Furthermore, the scope of this article includes protein families that respond to transition metals, both essential and poisonous, as well as the environmental toxins arsenic and lead. The alkali and alkaline earth metals will not be included here. Finally, the examples described are predominantly from prokaryotic organisms because these systems are more clearly defined than the eukaryotic versions. Several reviews on eukaryotic regulators are listed at the end of this article under “Further Reading.” Unfortunately, space limitations prohibit citing many of the primary references of the information discussed below, so the references are limited to a small number of publications. The reader is directed to the comprehensive review articles that are listed at the beginning of each section, which contain all of the appropriate references.

Biology
Many metal ions that can get into cells, such as mercury or lead, are poisonous and must be neutralized and exported as quickly as possible. However, even metals that have an essential cellular role, such as zinc, iron, or copper, can be toxic in excess. An organism must ensure an adequate supply of these nutrients while keeping the concentrations under tight control to prevent accumulation and cellular damage. Furthermore, each essential metal performs distinct cellular functions and seems to be regulated independently. Consequently, the metalloregulators are not only sensitive to changes in metal availability, caused by variations in the external supply or the nutritional needs of the organism, but they must be able to discriminate between the various types of metals present in the biological system. Metalloregulators contribute to the maintenance of this delicate balance by controlling the expression of metal uptake and export pathways, detoxification and storage/sequestration systems, as well as proteins that employ the metals such as metalloenzymes. The activity of the metalloregulators is linked intimately to the operations of these metal-centered pathways. For more information, see the articles on “Metal Complexes, Assembly of,” “Metal Homeostasis, intercellular,” “Metal Homeostasis: An Overview,” “Metal Transport through Membranes,” and “Metalchaperones, Chemistry of.”

The most common genetic control point of the metalloregulators is transcription, with one major exception in iron regulation discussed below. The response elicited by the metal can be repression, derepression, and/or activation of transcription (Fig. 1). Proteins are classified into families based on sequence homology (Table 1), and usually they respond in
the same manner as other family members to metal ions, although they may differ in metal selectivity. In many cases, the mechanism of genetic control is fairly straightforward. Metal binding to members of the ArsR/SmtB, DtxR, Fur, and NikR families of metalloregulators either activates or inhibits DNA binding. The DNA recognition sequences of these proteins are close to or are overlapping the transcription start sites in the promoters of the genes that they regulate, so it is likely that DNA binding by these proteins (either the apo or holo, depending on the system) sterically blocks transcription initiation by the RNA polymerase (2, 3, 4, 5).

However, as described in this section, the regulation by some other metalloregulators is more complicated. Several examples are discussed, each showcasing a possible mechanism of genetic control. A thorough understanding of the cellular biology of how these systems work, the biological chemistry is an integral component of their activities and thus is a part of the discussion.

**Mercury regulator**

One of the first metalloregulatory proteins to be characterized extensively is the prokaryotic MerR transcription factor (1, 6, 7), which acts either as a repressor (apo-protein) or an activator (holo-protein) of the mer operon encoding mercury resistance proteins (Fig. 1c). The −35 and −10 sequence elements of the mer promoter, binding sites for the RNA polymerase initiation complex, are separated by an unusually long distance that results in poor constitutive transcription. A apo-MerR binds to the DNA between these sequences and blocks the DNA, which results in a slight increase in repression on the suboptimal promoter. It also recruits the RNA polymerase to the transcription start site where it waits in a stalled complex. Upon binding of
**Table 1** Families of prokaryotic metalloregulators that function as transcription factors that are discussed in the text

<table>
<thead>
<tr>
<th>Family</th>
<th>Examples</th>
<th>Metal Systems regulated</th>
<th>Response to Metal</th>
<th>Metal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ArsR/SmtB</td>
<td>ArsR SmtB</td>
<td>Zn(II)/Co(II)/Cd(II)</td>
<td>Release of DNA</td>
<td>Dep reression of metal resistance proteins, efflux transporters, metal storage</td>
</tr>
<tr>
<td></td>
<td>CadC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MerR</td>
<td>MerR</td>
<td>Zn(II)/Co(II)/Cd(II)</td>
<td>DNA distortion</td>
<td>Activiation of metal resistance proteins, efflux transporters</td>
</tr>
<tr>
<td></td>
<td>CueR</td>
<td>Ni(II)/Co(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DtxR/IdeR</td>
<td>DtxR/IdeR</td>
<td>Fe(II)</td>
<td>DNA binding</td>
<td>Repression of uptake transporters, virulence factors</td>
</tr>
<tr>
<td></td>
<td>MntR</td>
<td>Zn(II)/Cd(II)/Pb(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fur</td>
<td>Fur</td>
<td>Fe(II)</td>
<td>DNA binding</td>
<td>Repression of: uptake transporters, metal scavengers, virulence factors, other cellular functions</td>
</tr>
<tr>
<td></td>
<td>Zur</td>
<td>Zn(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nur</td>
<td>Ni(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mur</td>
<td>Mn(II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NikR</td>
<td>NikR</td>
<td>Ni(II)</td>
<td>DNA binding</td>
<td>Repression of uptake transporters Activation of: nickel enzymes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*This table is not an exhaustive list of metal-dependent members of these families. Furthermore, proteins in these families that respond to factors other than metals that are not listed here.*

**Hg(II).** MerR undergoes a conformational change that causes the DNA to straighten out and to unwind; these distortions place the −10 and −35 sites in optimal positions for productive initiation of transcription. MerR can activate fully the mer promoter over a small range (less than an order of magnitude) of mercuric salt concentrations that range around $10^{-8}$ M in vitro and slightly higher in cell culture experiments. This sensitive and cooperative sensor provides a rapid and robust response that only is turned on when needed and is activated fully before cytoplasmic Hg(II) concentrations reach levels that affect cell growth.

Once the mercury has been eliminated from the cell, it would be wasteful energetically to continue to synthesize the resistance proteins. However, it is unclear how the Hg(II)-MerR complex could disengage from the promoter in a timely manner. It is possible that apo-MerR can displace the metal complex at the promoter because it binds the recognition sequence with slightly higher affinity (7), or that the Hg(II) is released eventually from the MerR-DNA complex. However, a recent study suggests that another protein encoded by the mer operon, MerD, may play a role in switching off the induction (8). MerD, which shares sequence homology with MerR, can form a ternary complex with apo-MerR and the DNA recognition complex. At the addition of mercury, the presence of MerD causes a fraction of the DNA to be released, which allows the expression of the divergently transcribed mer operon.

**Post-transcriptional iron regulation**

In mammals, as well as in certain other species, iron metabolism is regulated at the posttranscriptional level (for reviews see References 9-12). Two homologous iron-regulatory proteins (IRP1 and IRP2) bind with high affinity to specific RNA sequences called iron-responsive elements (IREs), which are present in the untranslated regions of the mRNAs that encode many of the proteins involved in iron metabolism. The IREs are conserved hairpin structures, but subtle differences in the sequences of the IREs and the surrounding mRNA affect IRP binding and fine-tune the strength of the interaction.

The IREs are bound by the IRPs at low iron concentrations (Fig. 1d). Two different effects of IRP binding exist, depending on the location of the IRE in the mRNA. In the case of proteins that would not be useful under limiting iron conditions, such as ferritin (iron storage), ferroportin (iron efflux), and aminolevulinate synthase (heme biosynthesis), the IRE is near the translation start site at the 5′ end of the mRNAs and IRP binding blocks translation initiation and protein production. In contrast, under the same iron-deficient conditions, the mRNA for the transferrin receptor (iron uptake) is protected from degradation by IRP binding to multiple IREs at the 3′ end of the mRNA, which enhances protein production and leads to an increase in iron intake. In an iron-replete situation the IRPs do not bind to the mRNA so the effects are reversed: The genes for iron-using proteins are translated, and the transferrin receptor mRNA is degraded.

The RNA-binding activities of the two IRPs have different mechanisms of inhibition by iron, and not all of the details are defined clearly. The binding of IRP1 to mRNA is blocked by the formation of an [4Fe-4S] cluster, which allows the protein to bind to the mRNA. The binding of IRP2 to mRNA is blocked by the formation of an [4Fe-4S] cluster, which allows the protein to bind to the mRNA.
to function as a cytosolic aconitase enzyme. IRP2 does not have an iron cluster, but it is targeted for proteosomal degradation in an iron-dependent process. The IRPs are regulated by a variety of additional factors such as heme, oxidative stress, nitric oxide, and phosphorylation by intracellular signaling factors, indicating that IRPs provide a bridge between iron metabolism and other cellular pathways.

**Other links to translation**

Another link to posttranscriptional regulation of metal homeostasis is the global iron regulator Fur (4). Fur controls the transcription of a large number of genes (more than 90 in E. coli), most of which encode proteins involved in iron acquisition as well as other essential metabolic pathways. The Fe(II)–Fur complex represses the transcription of these genes when Fe(II) is limiting, the Fur repression is alleviated when iron reaches critical levels, the internal kinase domain of the Fur protein is coupled to a protein conformational change such that it alters the DNA or RNA complex. In several examples, high-resolution structural studies have shed some light on the mechanisms of this allosteric response.

**Chemistry**

**Allosteric regulation**

The binding of the metal ion coregulator to a metalloregulator protein is coupled to a protein conformational change such that it alters the DNA or RNA complex. In several examples, high-resolution structural studies have shed some light on the mechanisms of this allosteric response.

**ArsR/SmtB family**

In the absence of metal, members of the ArsR/SmtB class of transcription factors bind to their respective DNA promoters and inhibit transcription; metal binding to the proteins decreases the affinity for DNA and allows transcription to proceed (derepression, Fig. 1b (2, 7)). All members of this family have a conserved helix-turn-helix (HTH) DNA-binding motif, with the same overall dimeric "winged" helix structure. Phylogenetic analysis suggests that this family evolved from a common evolutionary ancestor to sense specific types of metals (2, 17). However, the mechanisms of metal-induced DNA release do not seem to be the same. Furthermore, substantial variability exists in the metal-binding sites, which roughly can be divided into two distinct types based on their location and the nature of the amino acid ligands. A few family members have both metal sites, although in these cases, the metal bound to one of the sites may act as a structural cofactor, and at least one other member possesses a divergent third site (18).

Several proteins in this family respond to metal binding in a cysteine-rich site with at least some of the ligands from the N-terminal helix (α3) that is a part of the HTH DNA-binding motif. This site, often referred to as α3, α3N, or site 1, controls a response to thiophilic ions such as cadmium, lead, or arsenite. For example, the arsenic-bound by ArsR is coordinated by 2–3 cysteines (cys32, cys34, cys37) from the same α3 helix, so clustering the cysteines around arsenite would cause a large distortion of the helix that is proposed to disrupt DNA binding (19).

SmtB is regulated by the second type of metal-binding site called α5, α5C, or site 2. This site, which bridges the interface between α5 helices of the dimer, is composed of carboxylate and imidazole ligands and regulates the response to harder metal ions such as Cu(II), Ni(II), and Zn(II). The structures of SmtB revealed that when both α5 sites of the dimer are filled with zinc, a significant change occurs in the 3' structure of the protein that compacts the molecule (Fig. 2a (20)). This zinc-dependent conformational switch likely is controlled by a hydrogen-bond
Metalloregulatory Proteins

DNA-Binding Helix

D104
H106
E120'
H117'
R87
L88
L83

DNA-Binding Helix

Figure 2

(a) Metal binding to SmtB. (Top) The holo-SmtB structure (dark gray, pdb 1R22) is significantly more compact than the apoprotein (light gray, pdb 1R1T). The zinc ions (spheres) are only observed in the θ5 sites because the ligands of the θ3 sites were mutated to generate a protein that is still functional in vivo but only binds one zinc ion per monomer. The putative DNA-binding helix of the HTH motif is indicated. (Bottom) The hydrogen-bonding network between one of the θ5 zinc ions and the DNA-binding helix (L83) is highlighted. This network is not observed in the structure of the apoprotein. For clarity, the holo-protein is rotated slightly from the above view, and the two monomers are colored with different shades of gray. Putative hydrogens are indicated by thin bonds. (b) Structure of the E. coli nickel-responsive repressor NikR. The apoprotein (top, pdb 1Q5V), holo-protein (middle, pdb 2HZA), and DNA complex (bottom, pdb 2HCV) are shown with each monomer in the tetramer drawn in a different shade of gray. Two ribbon–helix–helix DNA-binding dimers flank the central core of four metal-binding domains. The nickel ions (smaller spheres) are coordinated in a square-planar site by H87, H89, C95, and H76' of the opposing monomer (shown in inset, site rotated for clarity). The larger spheres in the DNA complex are best modeled as potassium ions. The sections of the metal-binding domain that contact the DNA are circled, and these regions are not well ordered in the apo-structure and could only be modeled in one of the four monomers. A recent structure of apo–CadC suggests a different allosteric mechanism (21). CadC has both metal-binding sites, but the second site is probably structural because mutagenesis of the ligands in site 2 does not affect the metal-dependent response. The inducer site is u3N, and only two of the ligands (cys58 and cys60) are in the u3 HTH helix; the other two (cys7 and cys11) are donated by the N-terminal of the opposing subunit. Although the crystal structure does not have metal bound to the

u3N site, the structure suggests that binding of the metal would pull the N-terminal strand toward the DNA-binding motif and block access to the DNA sterically block. However, mutants of the u3N ligands do not release the DNA even though they still bind metals tightly, albeit in altered coordination sites, which indicates that the allosteric response must be more complicated (22). Structures with metals bound to the regulatory site or in a complex with DNA will help to clarify the details of this system.

NikR

NikR is the only known metalloregulator with a ribbon-helix-helix DNA-binding motif (5). This protein is a nickel-responsive repressor (Fig. 2b) that functions as a tetramer (Fig. 2b). It binds four nickel ions in square-planar sites that bridge the protein subunits, with the two flanking DNA-binding dimers linked to the central core of four metal-binding domains by flexible linkers. Several structures of E. coli NikR (apo-
Metalloregulatory Proteins

DtxR/Idr family

The homologous iron-dependent regulators DtxR and IdeR control the expression of virulence factors as well as proteins involved in iron homeostasis in pathogenic and nonpathogenic bacteria (3, 26, 27). The DNA-binding domains are N-terminal HTHs, and the dimeric iron complexes bind to palindromic sequences in the promoters of regulated genes and repress transcription (Fig. 1a and Fig. 3). Two distinct metal-binding sites exist in each monomer. One site is referred to as ancillary because metal-dependent activity is much less sensitive to mutating this site than of the primary-site ligands (See Reference 3 and references therein), although this site does have a role in enhancing the metal sensitivity of the repressor (28). Biochemical studies support a multistep, metal-activated DNA-binding mechanism that includes dimerization (29–32). Multiple X-ray crystal structures of the proteins without metals, in complexes with a variety of divalent metals bound in one or both sites, or bound to DNA, revealed that two dimers bind independently to opposite sides of the DNA (Fig. 3). Also, they suggest how the metal could influence DNA binding in addition to stabilizing the active dimer (for example, see References 33–36). The corepressor causes a shift in the DNA-binding domains in relation to the metal-binding domains, which closes the distance between these motifs and rotates them with respect to the rest of the molecule. One link between these two domains is Met10, which is a ligand of the primary-site metal from an N-terminal helix that contacts the DNA-recognition helix. The conformational change that allows Met10 to serve as a ligand also promotes hydrogen bonding between Glu9 and the imidazole of His106, another metal ligand. A hydrogen bond between His79 and Glu105, ligands of the ancillary and primary sites, respectively, support communication between the two metals. Furthermore, in the DtxR–DNA complexes, it was observed that the N-terminal helix is unwound so that it can be moved out of the way and avoid a steric clash with the DNA (34). This change in secondary structure may be mediated through hydrogen bonding with a metal-bound water.

Metal selectivity

Organisms employ a variety of transition metals, each for distinct functions. Although indirect connections exist between the metal pathways, it seems that the direct cellular control of individual metals, both nutrients and environmental toxins, functions independently. Each metalloregulator is dedicated to one metal or a subset of metal ions, even though it may belong to the same family as other metalloregulators. It is becoming clear that to differentiate between the available metals these proteins can take advantage of the inorganic chemistry of the metals’ ions by a variety of means (18).

Metal selection by protein sites

The high-resolution structures of the metal-binding sites of two MerR homologs, CueR and ZnR, reveal very distinctive coordination environments, in part controlled by the number of cysteine residues (Fig. 4) (37). CueR responds to Cu(I), Ag(I), and Au(I), and it binds all three metals in a linear, two-coordinate site composed of two cysteine residues that are conserved in all members of this family. This low coordination number is preferred by metals in the +1 oxidation state, and the site is shielded to prevent expansion of the coordination
Metalloregulatory Proteins

Cys114
Cys124
Cys79
Cys115
His119
PO₄³⁻
Cys112
Cys120
Ser77'
CueR
ZntR

Figure 4
(Top) Structure of CueR (pdb 1Q05) with two copper ions bound. One metal site is blown up in the picture below and has been rotated slightly for clarity. Ser77' is on the opposite subunit from the copper ligands and is replaced by a cysteine in ZntR. (Bottom) Metal-binding site of the dizinc cluster in ZntR (pdb 1Q08). The two subunits are drawn in light and dark gray ribbons.

number with external ligands. In addition, the authors suggest that CueR can discriminate against Hg(II), which is one of the few divalent metals that binds favorably to linear dithiolate sites, because it is optimized to provide charge compensation for a +1 metal ion but not +2. ZntR, which responds to Zn(II), Cd(II), and Pb(II), has the same overall dimeric structure as CueR, but it has four cysteines in each metal-binding site that contains a dinuclear zinc cluster. Additional ligands include a histidine and a phosphate ion (or sulfate). One of these cysteines, Cys79, is from the opposite end of the dimerization helix of the opposing subunit, which is linked to the DNA-binding domain of the monomer and suggests a mechanism of communication between the two domains. However, this cysteine is conserved only in the homologs that respond to divalent metals, which prefer higher coordination numbers and require larger charge neutralization than the monovalent metals.

Although the structure of Hg(II)-MerR has not been reported yet, a sequence alignment indicates that only three of the four cysteine ligands of ZntR are conserved in MerR. Furthermore, spectroscopic and mutagenesis analysis demonstrated that the metal is bound to these three cysteines (7). This intermediate number of cysteines suggests how MerR can select against the monovalent ions, which prefer the dithiolate sites and would not provide as much charge neutralization, as well as the divalent metals such as Zn(II) that prefer higher coordination numbers.

Metal selectivity through the metals

Given that the coordination geometry of a metal complex is a determinant of metal selectivity, it is predicted that all metalloregulators would have a metal-binding site preorganized to accept the appropriate metal(s) while excluding all others. However, this clearly is not always the case because many examples of metalloregulators exist in which the metal ligands could accommodate multiple different metals. Furthermore, if this model were correct, the affinity of the different metals would parallel the selectivity for the DNA-binding response, which is not always observed. Instead, in some cases, the metal-binding sites are clearly flexible and the presence of the correct metal is recognized by the distinct coordination geometry imposed on the protein.

For example, NmtR from Mycobacterium tuberculosis is a member of the ArsR/SmtB family that responds in vivo to nickel, and to a lesser extent to cobalt, which binds in an α₅ site. Zinc is a poor allosteric inducer both in vitro and in vivo, even though it binds more tightly to the protein than nickel and cobalt (38). An explanation for these observations was provided by spectroscopic analysis that revealed the Zn(II) ion bound in a tetrahedral 4-coordinate site, whereas N(III) was bound in a 6-coordinate octahedral site (38, 39). In contrast, CzrA from Staphylococcus aureus responds well to zinc, not nickel, but this protein binds the different metal ions in an α₅ site with the same type of geometries as in NmtR (39). Thus, these two proteins accommodate each metal ion in the preferred coordination geometries of the metals, but they have evolved such that only one activates the allosteric response of each protein: 4-coordinate activates CzrA and 6-coordinate activates NmtR.

One question that is raised by these studies is whether a metal ion that binds tightly to the protein but is a poor allosteric effector in vitro will compete with the inducer and act as an inhibitor. This issue is currently under investigation (40), and one factor that clearly must be addressed is whether all of the possible metals are even available in vivo.

Metal availability

When NmtR from M. tuberculosis was transplanted to a cyanobacterial host, a response from the protein was observed when extra cobalt was added to the growth media but not to...
operons that encode proteins that are clearly involved in metal specific metalloregulators can be found through genetic experiments that demonstrate a change in expression when the organism is grown in the presence of extra metal. Then, the expression of genes controlled by a metalloregulator can come from these operons. The first clue that a gene or an operon is controlled by a metalloregulator, i.e., how much metal is necessary to activate the genetic switch? For example, in an in vitro study of a pair of E. coli zinc sensors, ZnRE and Zur, revealed a sensitivity that correlated with physiological function (42). In response to zinc, Zur shuts off transcription of zinc uptake genes and ZnRE turns on transcription of efflux. They have a graded response such that Zur responds to lower zinc concentrations than ZnRE, turning off uptake before efflux is activated, which prevents both uptake and efflux from working against each other concurrently in a futile cycle. It is also interesting to note that the responses of ZnRE and Zur both occur over a very small gradient of zinc concentrations (≈10 orders of magnitude), which reveals the tight window of optimal zinc levels that falls between starvation and toxicity. However, the metal-buffered conditions suggest that this optimum is on the order of femtomolar free zinc, which is far less than one zinc ion per cell. CusR, which activates CusEs export, may be even more sensitive (37). Essentially, such sensitivities imply that no free copper or zinc exists in an E. coli cell under healthy growth conditions and invite questions about the availability of the essential metals for the destination biomolecules that use the metals as cofactors, many of which have much weaker thermodynamic affinities for the metals than the metalloregulators. It has been proposed that in a cellular context, metal delivery is under kinetic control (43), possibly through the activity of intracellular trafficking factors. In addition, many factors exist in a cellular milieu with varying degrees of metal-binding capabilities, such as amino acids, carbohydrate metabolites, thiol-containing molecules, and even weak nonspecific sites on protein surfaces, which would soak up any "free" metal ions and would provide pools of readily accessible ions if needed. The metalloregulators must be tuned to this buffering capacity to respond appropriately and to maintain a healthy balance of metal homeostasis.

Tools and Techniques
To define the complete mechanisms of the metalloregulators requires the use of a broad spectrum of methods, some of which are mentioned below. The first clue that a gene or an operon is controlled by a metalloregulator can come from in vivo experiments that demonstrate a change in expression when the organism is grown in the presence of extra metal. Then, the specific metalloregulators can be found through genetic experiments. A novel approach, facilitated by the many complete genome sequences now available, is to search for a gene that encodes a homolog of a known metalloregulator, or to examine operons that encode proteins that are clearly involved in metal homeostasis. Once a gene is identified as a possible metalloregulator, the assignment must be confirmed by in vitro experiments. A flip recombinant expression and purification of the protein, several different methods are used to examine metal binding and how this affects DNA (or RNA) binding. The protein might be purified with some metal bound, but whether that metal is physiologically relevant or one that was available in the expression host is a tricky question and should be resolved by in vivo experiments. Some metal–protein sites can be observed by electronic absorption spectroscopy (UV/visible spectroscopy), which can be used to determine the stoichiometry if the affinity is tight enough, as well as the dissociation constant, possibly through the use of small-molecule chelators as competitors. Metal binding and stoichiometry can also be examined by treating the protein with excess metal, removal of unbound metal with a method such as dialysis or gel filtration chromatography, and then direct metal analysis with inductively coupled plasma atomic emission spectroscopy ICP-AES, ICP-MS, atomic absorption spectroscopy, or other techniques. If any standard procedures to measure the strength of a protein–ligand interaction can be used to determine the metal affinity; in addition to UV/visible spectroscopy, these procedures include fluorescence spectroscopy, isothermal titration calorimetry, and equilibrium dialysis.

Conserved residues, particularly cysteine, histidine, and aspartate/glutamate, can signal likely metal ligands, which is an assignment that can be tested by mutagenesis. Detailed information on the coordination sphere can also be provided by spectroscopic techniques such as X-ray absorption spectroscopy, as well as UV/visible spectroscopy, or electronic paramagnetic resonance spectroscopy for some metals.

To examine DNA binding, typically a pair of complementary oligonucleotides that contain the DNA recognition sequence is used. If the binding site is not known, a method such as DNA se footprinting on a longer fragment of DNA that contains the whole promoter region will reveal the location of the binding site. The recognition sequence can be confirmed by using in vivo reporter assays. DNAase footprinting, mobility shift assays, and fluorescence anisotropy are some common techniques used to monitor DNA binding in the presence or absence of metal(s). Variations on these methods can provide information on whether the protein binds or unwinds the DNA, in the case of metalloregulators that function as transcription factors, in vitro transcription assays can sometimes be used to examine directly how the protein and metal influence transcription.

Finally, biophysical studies can examine how metal binding influences protein conformation and can serve as the basis for a hypothesis about the connection between the two activities, metal binding and DNA binding. Furthermore, it is clear from the discussion that a high-resolution structure, either from X-ray crystallography or nuclear magnetic resonance spectroscopy, is indispensable. A structure of the apo protein provides information on likely metal sites and the conformation of the DNA–binding domain. Structures of the metal-bound and/or DNA-bound complexes, in comparison with the isolated protein, can illuminate the molecular details of the structure-function relationship and can serve as a key reference point in understanding all of the pieces of information provided by solution studies.


33. Pohl E, Holmes RK, Hol WGJ. Motion of the DNA-binding domain with respect to the core of the diphtheria toxin repressor (DtxR) revealed in the crystal structures of apo- and holo-DtxR.

Further Reading

See Also
Metal Complexes, Assembly of
Metal Homeostasis, Intracellular
Metal Homeostasis: An Overview
Metal Transport Through Membranes
Metallochaperones, Chemistry of
Transcription Factors
Carbohydrate–Carbohydrate Interactions
Nicole Seah and Amit Basu, Brown University, Providence, Rhode Island
doi: 10.1002/9780470048672.wecb199

Carbohydrate–carbohydrate interactions have been suggested as mediators of cell adhesion and aggregation. Studies of four different interactions—sponge cell aggregation, embryo and myelin compaction, and melanoma cell adhesion—have provided insights into the role of the saccharides in these events. The biological context of these associations as well as the results of experiments using biophysical and chemical model systems are described.

Cell-surface glycans in the form of glycoconjugates, glycoconjugates, and glycolipids play an important role in biological recognition processes. Although most known receptors for cell-surface carbohydrates are proteins (i.e., lectins), several studies over the past few decades have clearly established that carbohydrates can function as binding partners for each other. These investigations of carbohydrate–carbohydrate interactions (CCIs) are broadening our view about the function of cell-surface glycans and may lead to a more thorough understanding of the principles of carbohydrate recognition. Although noncovalent interactions between individual residues of polysaccharides such as glycosaminoglycans and hyaluronic acid have long been recognized (1), the role of CCIs in cell adhesion and signaling is not fully understood. Some types of CCIs, such as those involved in sponge cell adhesion, are thought to have primarily structural roles. In other instances, such as melanoma cell adhesion, the CCI is believed to initiate signal transduction.

Chemical structures of carbohydrate epitopes from four well-studied CCIs are shown in Fig. 1. CCIs can occur via homotopic self-recognition, or they can be heterotopic. A homotopic interaction of the sulfated disaccharide 1 mediates sponge cell adhesion, whereas self-recognition of the Lewis x triaccharide 2 has been identified in embryonic development. Divalent cations are required for these two homotopic CCIs. The interaction between the glycosphingolipid GM3 (3) and the glycolipids LacCer (4) or GalCer (5) mediate the adhesion of melanoma cells to lymphoma or endothelial cells and initiate signal transduction. Carbohydrate epitopes involved in CCIs can also be as small as a monosaccharide, as is the case with the glycolipid GalCer (6) and cerebroside sulfate (CBS, 7) found in the myelin sheath. These heterotopic CCIs also use calcium, but the metal does not seem to be essential for association. Other carbohydrates that engage in CCIs have been reported for the aggregation of human embryonal carcinoma cells and trout sperm fertilization (2, 3).

All structures shown in Fig. 1 are involved in a trans-interaction, which requires that sections of the apposing cell membranes be brought in close proximity so that membrane-bound carbohydrates can interact with each other. This type of cell junction has been referred to as a glycosynapse, an analogy to immunologic and neurological synapses. The existence and characterization of glycosynapses and the elucidation of their mechanisms of action remain important challenges for the field of cell-surface carbohydrate recognition.

Alternatively, preliminary evidence has hinted at the occurrence of CCIs between carbohydrates that are present on the same membrane, i.e., interactions in cis. In these cases, a glycolipid has been shown to perturb the function of an N-glycosylated membrane-bound protein (5–7). Although the perturbation may be a result of a CCI between the glycolipid and the N-glycan, a specific and conclusive demonstration of a CCI in these systems awaits further studies.

The following sections provide a biological context for each of the four well-characterized CCIs shown in Fig. 1 and describe in greater detail the results of studies using a variety of chemical model systems and biophysical tools. It is notable that despite the very different contexts for these interactions (e.g., glycolipid vs. proteoglycan), in each case the minimal carbohydrate residues alone are necessary and sufficient for association, as long as the sugars are presented in a multivalent format. This ability to reproduce CCIs in model systems (e.g., glycosylated polymers, surfaces, proteins, vesicles, etc.) has greatly facilitated studies of this phenomenon. Although biomimetic models necessarily do not capture all components of a CCI present in the cellular context, they provide insights into carbohydrate molecular recognition. Additionally, these model systems offer tools that can subsequently be used to study the biological functions of CCIs in their more complex native environments.
Carbohydrate–Carbohydrate Interactions

Figure 1 Glycan structures that participate in carbohydrate–carbohydrate interactions.

CCI in Sponge Cell Adhesion

Biological Context

One of the first examples of a CCI was identified in various sponge species, where interactions between cell-surface proteoglycans mediate the self-association of sponge cells (8). Sponges that have been dissociated into their component cells are capable of reassociating in a species-specific manner. This phenomenon has been extensively studied using the marine sponge *Microciona prolifera* as a model system. The selective self-association is mediated by a proteoglycan referred to as the *Microciona* aggregation factor (MAF). MAF contains an N-linked sulfated glycan (g200) composed of fucose, glucuronic acid, mannose, galactose, and N-acetyl-glucosamine (9). MAF extracted from sponge cells does not bind to other sponge cells until it is rendered multivalent by glutaraldehyde cross-linking (10). MAF-mediated cell aggregation can be inhibited by several antibodies that bind to specific carbohydrate epitopes within the glycan. The Block 1 antibody reactive epitope is a carbohydrate containing a novel pyruvate ketal (11). A second epitope, reactive with the Block 2 antibody, was identified as GlcNAc(3S03)β1–3Fuc (12). The identity of these oligosaccharide epitopes was determined by degradation of the polysaccharide followed by nuclear magnetic resonance (NMR) and microscopic (MS) analysis. Other oligosaccharide components of proteoglycan aggregation factors from additional sponge species have also been identified (12).

Model Studies

Early studies of sponge cell aggregation demonstrated that CCI s were still observed when the glycans were removed from their native context. A striking visual example of the species specificity of CCI is evident in the aggregation of proteoglycan-attached microbeads, shown in Fig. 2 (13). Figure 2a shows three different sponges, each with a distinct coloration. Adhesion factors from the sponges were removed and covalently attached to beads color-coded pink, yellow, or white for each sponge. When the three beads were mixed together in the presence of calcium, they self-sorted into singly colored clusters based on the origin of the glycans on each bead (Fig. 2b). In contrast, if the pink, yellow, and white beads were all coated with the same glycan, no color sorting was observed (Fig. 2c). These glycan-coated beads are also competent binding partners for whole cells, as well as immobilized glycans 14.

To demonstrate the minimal structural requirements for aggregation, sulfated disaccharide 1 was attached to a gold nanoparticle via a thiol linker (15). Particles coated with α-linked fucose residues rapidly aggregated in the presence of Ca2+. Minor structural modifications to the sugar prevented aggregation, and although the β-linked sugar was capable of forming aggregates, the association was weaker than the α-anomer.
Sponge cell CCI has also been studied using bovine serum albumin (BSA) conjugates of the disaccharide 1 (16). A solution of 3BSA aggregated rapidly in the presence of Ca\(^{2+}\) but not Mg\(^{2+}\), which is consistent with previous results observed with the intact proteoglycans. 3BSA was immobilized on a dextran-coated surface for adhesion studies using surface plasmon resonance (SPR). When a solution of 3BSA was flowed over the chip, formation of glycoconjugate multilayers on the surface was observed. Similar results were obtained from SPR measurements using intact MAF or g200 (17). Analysis of binding isotherms from the adhesion studies with 3BSA provided an association constant of 10\(^{10}\) M\(^{-1}\) for the interaction of the carbohydrate units. The adhesive forces between glycans involved in sponge CCI have been examined using atomic force microscopy (AFM) (14, 17, 18). AFM tips and mica surfaces were modified with intact MAF or the extracted g200 glycan. Pull-off adhesion forces ranging between 100 and 300 pN have been measured for these interactions. Examination of g200 adhesion from different individuals exhibit distinctions between self–self-interactions and self–non-self-interactions, consistent with the allogenicity of sponge cell adhesion (17).

### CCI in Embryo Cell Adhesion

#### Biological Context

Embryonic development is a process in which cell–cell adhesion is crucial for proper growth of the nascent organism. After fertilization of an egg, the embryo begins to divide into a tightly packed cell cluster known as the morula. The Lewis x trisaccharide, also referred to as the stage-specific embryonic antigen 1 (SSEA-1), mediates cell adhesion and compaction of the morula during the 8–16 cell developmental stages. Lactofucopentaose III (LNF-III), a Lewis x-containing oligosaccharide, was identified as a carbohydrate inhibitor of fully compacted 8–16 cell embryos. Multivalent inhibitors were constructed by conjugating LNF-III to a lysyllysine peptide (19). Decompaction of embryo cells was observed only with the multivalent LNF-III conjugate, and not with free LNF-III oligosaccharides or other multivalent oligosaccharides. F9 embryonal carcinoma cells resemble cells undergoing early stages of embryonic development, as they also express surface Lewis x antigens at the undifferentiated stage and express less of the antigen upon differentiation. Aggregation of these cells was inhibited by LNF-III or EDTA, consistent with a calcium-dependent, homotypic Lewis x interaction (20). Furthermore, solid-supported F9 cells interacted preferentially with Lewis x-containing liposomes. Glycolipids bearing the Lewis x trisaccharide specifically interact with F9 cells, Lewis x-containing liposomes, Lewis x-functionalized columns, and other Lewis x-bearing lipids. Finally, latex beads or gold nanoparticles coated with Lewis x glycolipid aggregate in the presence of calcium ions (21, 22). In each case, the divalent cations Ca\(^{2+}\), Mg\(^{2+}\), or Mn\(^{2+}\) are required for the interaction. The Lewis x CCI was used to engineer the adhesion of non-Lewis x producing cells. Additions of Lewis x glycoporphin-gold nanoparticles to a culture of nonaggregating rat basophilic leukemia (RBL) cells resulted in the incorporation of these glycolipids into the plasma membrane and the subsequent clustering of the RBL cells in the presence of calcium (23).

#### Model Studies

Structural studies using NMR, MS, and X-ray crystallography provided molecular details of the Lewis x CCI. Electrospray ionization mass spectrometry of Lewis x in the presence of calcium ions gave rise to ion peaks consistent with two Lewis x moieties bound to a single calcium atom (24). Additionally, this complex was observed by cold spray ionization, which also detected Lewis x aggregates formed in the absence of calcium (25). NMR spectroscopy of Lewis x in the presence of divalent manganese cations identified tentative metal binding sites based on chemical shift perturbations induced by the paramagnetic ion (26). The binding of monomeric Lewis x to liposomes containing Lewis x has been detected using transfer NOESY experiments, although the strength of the association was exceedingly weak due to the monovalency of one of the binding partners (27). While calcium binding to the trisaccharide alone could not be detected by NMR, linking two molecules of Lewis x by a short tether provided a complex for which nuclear Overhauser enhancements (NOEs) were observed (28). The NMR experiments, along with computational analysis, suggested a model for the complex in which the calcium atom is sandwiched between the two linked trisaccharides with an association constant of 55 M\(^{-1}\). In the solid-state structure of Lewis x, rows of trisaccharides are arranged in a head-to-head arrangement with especially strong interactions between fucose and galactose residues (29, 30). Three intermolecular hydrogen bonds between each Lewis x pair are proposed to mediate their association.

---

Figure 2. (a) Discrimination between self and non-self, in a suspension of live glyconectin-bearing Microcosm problems (orange), Halichondria, purple (white), and Chone vitrifera (brown) sponge cell at 8°C. The microscopically observed colors are from the whole sponge. (b) Ca\(^{2+}\) dependent species-specific recognition and adhesion of pink, yellow, and white beads coated with glyconectins from each sponge. (c) Control experiment showing sorting of pink, yellow, and white beads coated with glyconectins from M. prolifera sorting of pink, yellow, and white beads, all coated with glyconectins from M. prolifera. Magnification =40. Adapted by permission from Macmillan Publishers Ltd. (13).
The Lewis x CCI was probed using a variety of quantitative biophysical approaches. Lewis x bearing a thiol-terminated lipid tail has been attached to gold surfaces and AFM tips. AFM measurements of the interaction between these tips and surfaces measured an adhesion force of 20 pN (31). The aggregation of Lewis x-coated gold nanoparticles in a calorimeter was found to be highly exothermic (31, 32). SPR measurements of the interaction between Lewis x-coated gold nanoparticles and a self-assembled monolayer on gold functionalized with Lewis x have been carried out, and association constants of 1.9 \times 10^7 M^{-1} have been reported (33). In an alternative approach, two vesicles containing Lewis x were held in tight contact using micropipette manipulation (34). An adhesion energy of approximately 0.6 kJ/mol was obtained from contact angle measurements of the vesicle interface.

CCI in Melanoma Cell Adhesion

Biological Context

Heterotopic CCI involving glycolipids were first identified from studies of the adhesion of murine B16 melanoma cells. These cells, which express high levels of the glycoprophospholipid GM3 (3), adhered to mouse AA12 T-cell lymphomas that express the glycolipid Gg3 (6), but not to the non-Gg3-expressing AA27 variant of the cell line (35). The adhesion of AA12 lymphoma cells to a GM3-coated surface was inhibited either by pretreatment of the AA12 cells with an anti-Gm3 monoclonal antibody or by blocking the GM3-coated surface with an anti-Gm3 antibody. Conversely, adhesion of B16 melanoma cells to Gg3-coated solid surfaces was correlated to levels of an anti-Gm3 antibody. Adhesion of B16 melanoma cells to GM3-coated solid surfaces was correlated to levels of a monoclonal antibody. Avidity of B16 melanoma cells to GM3-coated solid surfaces was correlated to levels of a monoclonal antibody. Avidity of B16 melanoma cells to GM3-coated solid surfaces was correlated to levels of a monoclonal antibody.

The GM3 in B16 cells is localized within glycolipid-enriched microdomains in the lipid membrane (37). The microdomains, often referred to as lipid rafts, also contain a variety of signal-transducing proteins and enzymes. Binding of B16 cells to immobilized Gg3 activates the signaling proteins Ras and Rho, suggesting that CCI does not simply mediate cell adhesion but may also be involved in signal transduction. The structure and function of the glycolipid signaling domains in B16 cells is perturbed by the addition of an unnatural, synthetically prepared glycosphingolipid (38). The synthetic glycolipid reduced GM3 clustering on the cell surface and diminished the activity of the kinase c-Src, identifying another link between CCI and signal transduction. The specific mechanism by which binding to GM3 activates signal transduction is not currently known.

Several studies suggest that interruption of the LacCer-GM3-mediated adhesion prevents metastasis, raising the possibility of targeting this CCI for cancer therapy. The metastatic potential of several variants of the B16 cell line is correlated with their LacCer binding ability (39). Injection of LacCer-containing liposomes to mice that have aggressive melanoma inhibited the development of metastatic lung tumors in the animals (39). Administration of multivalent glycoconjugates such as glycosylated peptides or glycosylated gold nanoparticles also inhibits metastasis in vivo (40, 41). Inhibition of GM3 biosynthesis has also been shown to reduce metastatic melanoma migration in an animal model (42). Further studies with these glycoconjugates are needed to clarify their mechanism of action in vivo.

Model Studies

Langmuir monolayers have been used to model the interactions between GM3 and LacCer or Gg3. Micelles of a lactosyl monolayer injected under a GM3-containing monolayer induces a change in the surface pressure (∆π) at the air-water interface in a concentration-dependent manner (43). These increases in ∆π are a consequence of glycolipid insertion into the membrane, but glycoconjugates that do not insert also increase ∆π. The role of the divalent cation in this interaction is unclear, as it is increased in some conditions but not others (43, 45, 46). In a separate study, the behavior of various polystyrene glycoconjugates with a GM3 monolayer was examined (47). Pressure-area compression isotherms of GM3 monolayers exhibited expanded areas when they were generated over a solution of the GM3 glycopolymer. The amount of Gg3 glycoconjugate adsorbed on the GM3 monolayer was quantified using a quartz crystal microbalance by transferring the glycoconjugate-bound monolayer to a quartz substrate (46, 47). An SPR study of this same system provided an association constant for Gg3 glycopolymer binding to a GM3 monolayer of 2.5 \times 10^7 M^{-1} (48, 49).

CCI in Myelin Compaction

Biological Context

The myelin sheath is a multilayered membrane that wraps around axons in the nervous system and facilitates the transport of action potentials by preventing dissipation of a signal as it travels along the axon. The lipid bilayers of the myelin sheath are extensions of the membranes of oligodendrocytes in the central nervous system and Schwann cells in the peripheral nervous system. The specific mechanism by which this CCI mediates adhesion is not currently known.
nervous system. Unlike many other cell membranes, myelin is highly enriched in glycolipids, primarily GalCer (2) and CBS (3), that are preferentially embedded in the outer bilayer (50, 51). In addition, the multilayers are tightly compacted, a process dependent on divalent cations (52). A CCI between GalCer and CBS is believed to play an important role in maintaining the structural integrity of the myelin sheath.

The cellular function of this CCI was studied by examining the interaction of GalCer/CBS containing liposomes with cultured oligodendrocytes (53). Anti-GalCer antibodies bind to glycolipid containing membranes on oligodendrocytes and this binding decreases when the cells are pre-incubated with the glycolipid liposomes. Antibody binding to myelin binding protein (MBP) was similarly inhibited by the liposomes, indicating a colocalization of the glycolipids and MBP in the oligodendrocyte membrane. Liposome incubation also resulted in loss of microtubule fine structures, suggesting that actin filaments are destroyed upon CCI. Similar results were obtained in subsequent studies using galactose (Gal) and sulfated galactose (sGal) conjugates of BSA (54). BSA conjugates containing both Gal and sGal elicited a stronger response than albumin modified with a single carbohydrate.

Model Studies

The interaction between these two glycolipids has also been examined in various chemical model systems. A ceramide sulfate (CBS) coated surface is adherent to liposomes containing GalCer but not to faciol or galactosyl ceramide (55). The aggregation of glycolipid containing liposomes in the presence of various divalent cations was examined (56). The largest amount of aggregation was observed between liposomes containing GalCer and CBS, and the magnitude of the interaction was inversely correlated with the ionic radius of the cations. The presence of longer or N-hydroxylated N-acyl chains on the glycolipids also increased the association between GalCer and CBS liposomes. Infrared spectroscopy of this interaction revealed that the presence of calcium changed the conformation of the sulfate group in CBS, thereby disrupting several intramolecular hydrogen bonds (57, 58). As membranes in myelin contain both GalCer and CBS, this CCI can occur in a cell membrane in myelin.

References

Carbohydrate–Carbohydrate Interactions


Further Reading


See Also

Lipid Domains, Chemistry of Neoglycoproteins, Chemistry of Sugar-Lectin Interactions in Cell Adhesion Glycoinformatics Surface Techniques and Surface Chemistry
Molecular Recognition, Computation and Modeling of

Anna Feldman-Salit and Rebecca C. Wade, *EMI Research*, Heidelberg, Germany
doi: 10.1002/9780470048672.wecb354

Molecular recognition is critical to many fundamental processes in biology, including enzymatic reactions, signal transduction, genetic information processing, as well as molecular and cellular transport. The question of how proteins selectively recognize and correctly associate with their partners is an active research problem. For many proteins, experimental information about the structures of their molecular complexes and the details of their interactions are lacking. Macromolecular, structure-based computational techniques provide a means to predict the interactions of proteins and to investigate their recognition mechanisms. Here, we first discuss the general mechanism of molecular recognition as a multistep process from diffusion of one molecule toward the second to tight complex formation. We then describe the main forces and the physical properties that govern biomolecular interactions and introduce the principles of modeling them. Finally, simulation methods and computational approaches for molecular docking are briefly presented.

An understanding of the mechanisms of molecular recognition provides the essential basis for rational structure-based drug design and bioengineering. However, biomolecular recognition is complex and is determined by a fine balance between different contributing properties such as shape, electrostatics, dynamics, and entropy. This complexity makes the modeling of biomolecular recognition and the prediction of the properties of biomolecular complexes challenging. Here, we first discuss the factors important for biomolecular recognition that should be considered in computational models. Then simulation methods, such as Brownian and Molecular Dynamics, as well as other computational techniques that are widely exploited for studying protein-protein and protein-ligand recognition are presented. Last, we describe the principles, approximations, and limitations of some existing computational molecular docking approaches, which provide a means to predict the structures of bound complexes.

Although molecular recognition is of importance to all biomolecules, the astounding molecular recognition properties of the immune system have been the subject of investigations since the nineteenth century. It is intriguing how a vast number of chemically similar antibodies are serologically indistinguishable but can react with different antigens can be produced in one species (1). Similarly, the major histocompatibility complex can recognize a huge variety of peptides. How the immune system achieves this spectrum of molecular recognition properties is still not fully understood. Nevertheless, studies of this molecular recognition process have been exploited practically, namely in the design of vaccines that exploit the molecular recognition between antibodies and cell surface antigens.

The recognition between proteins and nucleic acids has challenged scientists for a long time. RNAs, for example, are capable of enzymatic activity and make direct contributions to substrate recognition and catalysis in ribonucleoproteins such as the ribosome (2) and the spliceosome (3). Nearly all the functions of RNAs are associated with the binding of proteins. However, the question of how a protein recognizes a specific RNA site, what effects it has on the RNA structure and dynamics, and how it promotes a specific RNA function are still very demanding. It is, for example, of great interest to understand how aminoacyl-RNA synthetases achieve the specificity needed to ensure faithful translation of the genetic code.

**Biological Background**

Molecular recognition is one of the most fundamental and important processes in biology. It occurs between two or more molecules, and it is involved in biochemical processes such as enzymatic reactions, molecular transport in the cell, genetic information processing, and protein assembly.
Mechanisms of Molecular Recognition

The mechanism of molecular recognition can be described schematically as a multistep process. Two molecules first come into proximity from afar. If active transport does not occur, then this process will occur by diffusional association with formation of a diffusional encounter complex, which is a loose complex between the molecules. Depending on the degree of burial of the binding sites, diffusion may occur of a small molecule through its macromolecular receptor to the binding site and this may require conformational changes in the receptor. Finally, and usually after induced fit conformational changes, the molecules will form a fully bound complex (11). The extent to which induced fit exists as opposed to selection of a binding conformation among a conformational ensemble of a protein is the subject of debate (12); but presumably, it varies according to the system, as does the extent of conformational change.

The binding process may be mediated first by long-range electrostatic interactions between the two associating molecules and then complemented by short-range van der Waals interactions and hydrogen bonding. Water molecules can either be displaced on binding or retained to mediate binding of the molecules. The forces involved in the process of molecular recognition are discussed in more detail later.

Kinetics and thermodynamics of molecular recognition

The rate of binding of two molecules A and B and formation of their complex AB can be quantified by the bimolecular association rate constant (on-rate constant), $k_{on}$, whereas their unbinding is characterized by the dissociation rate constant (off-rate constant), $k_{off}$.

$$ A + B \rightarrow_{k_{on}} AB \rightarrow_{k_{off}} [A][B] $$

The binding affinity of two molecules can be quantified by the equilibrium dissociation constant, $K_d$.

$$ K_d = \frac{[A][B]}{[AB]} $$

With the relation between $K_d$ and the binding free energy (Gibb's energy), $\Delta G$, $K_d$ can be expressed as

$$ K_d = \frac{C_0 \exp(-\Delta G/RT)}{C_d} $$

where $R$ is the universal gas constant ($\sim 8.31 \text{ J} \cdot \text{K}^{-1} \cdot \text{mol}^{-1}$), $T$ is the temperature in K, and $C_0$ is the standard concentration. The dissociation constant is usually given in molar units (M).

The binding free energy $\Delta G$ is defined by the interplay between two thermodynamic entities: binding enthalpy, $\Delta H$, and binding entropy, $\Delta S$.

$$ \Delta G = \Delta H - T \Delta S $$

The phenomenon of "enthalpy-entropy compensation," in which one term favors binding and the other term disfavors binding, is observed for many molecular recognition processes. On one hand, the binding benefits from enthalpic contributions that include electrostatic, hydrogen bonding, and van der Waals interactions. On the other hand, the unfavorable loss of translational and rotational entropy occurs mainly due to the reduction in the number of degrees of freedom of the molecules on their complexation. Bimolecular binding is also disfavored by the loss of flexibility of the molecules on binding, although it can benefit from increased flexibility at a distance from the binding sites (allosteric effect). The displacement of water molecules on binding may also be entropically favorable. Different sources of entropic contributions to binding are discussed later (see the section entitled "Entropy").

Conformational changes on binding

The conformational flexibility of the molecules plays a very important part in their binding. When a protein–protein complex is modeled using the structures of the two proteins determined in their unbound states, some key side chains or structural motifs might be positioned to result in steric clashes or unfavorable electrostatic interactions (13, 14). Therefore, on binding, the intermolecular interactions should result in some degree of induced fit. Induced fit can be defined as the collection of conformational changes, which results in optimal interactions, when two molecules come into contact (14).

These conformational changes in biomolecules can vary from small-scale motions, which include bond stretching, bond angle bending, and dihedral angle variations, to larger motions of the main chain, which are the loops or entire domains of proteins. Such changes occur on different timescales, which range from picoseconds to hours, with amplitudes that extend to hundreds of Ångstroms (15). Small changes in molecular conformation can adversely affect molecular docking results when using rigid molecular models. For example, Betts and Sternberg (16) have found that RMS deviations of 0.6 Å for backbone and 1.7 Å for side chain atoms affect protein-protein docking.

The dynamics of protein side chains occurs on the time scale of picoseconds to milliseconds with RMS deviations up to about 3 Å. Such motions include the rotation of side chains between different rotamers. A "flexibility scale" for protein side chains has been determined (17). It is shown that protein-protein interactions are critically dependent on
“hot-spot” residues at the binding interface (18). Molecular dynamic simulations have revealed that the hot-spot residues are found frequently in the unbound proteins in the rotamers observed in the protein-protein complex (19, 20). In contrast, the side chains at the perimeter of the interface are mostly found in nonspecific rotamers in the unbound protein, which suggests that their final conformations are induced on binding.

It was proposed by Kimura et al. (19) that these side chains act more as “latches” that hold the molecules together rather than as “keys” fitting snuggly into the binding pocket. According to the mechanism of induced fit, these latches open before docking and “fasten” to their partners later during the binding process.

The conformational rearrangements of backbone atoms and of protein loops play a significant role in ligand recognition, protein-protein association, DNA-binding, and so on (21–24). For developing the docking methodology, Chen et al. (25) and Vajda et al. (26) classified 59 protein complexes based on their docking difficulty. The cases in which substantial conformational changes of the backbone or loops between the unbound and bound structures were observed were referred to as “difficult.” Such backbone motion may be up to 10 Å or more and can drastically modify the steric and the electrostatic properties of the protein face presented to the partner. In some cases, loop motion may be caused by large changes in as few as two torsion angles (27), but sometimes the deformations are distributed over more torsion angles (28, 29).

Large-scale domain or subunit motions are often classified into hinge and shear motions (30). Short loops or elements of secondary structure can serve as hinge regions. Interactions of ligands with binding domains may cause the protein to undergo certain conformational changes, which sometimes leads to closure of a binding site, see Fig. 1.

The extent and the variation of the conformational changes that can occur on molecular binding show the necessity for these changes to be taken into account when modeling the molecular recognition and the binding of two associating molecules.

Modeling Molecular Recognition

Modeling of the different steps of molecular recognition requires suitable models of the corresponding forces and the energetic contributions. These elements will be considered in the next section.

Shape complementarity and van der Waals interactions

In 1894, Emil Fischer postulated that biomolecular recognition is analogous to the matching of a “key” into its “lock,” in which the “lock” refers to an enzyme and the “key” refers to a substrate. Only the correctly shaped key (substrate) fits into the hole (active site) of the lock (enzyme). Describing molecular recognition in terms of structure and bonding in the mid-1940s, Pauling defined the specificity of biomolecular recognition, which occurs due to the mutually complementary configurations of molecular surfaces. In other words, the surface of one molecule conforms closely to the surface of the other molecule on biomolecular binding.

Molecular shape complementarity is critical to biomolecular recognition and specificity. Even if the molecules change conformation on binding and water molecules are trapped at the interface, bound complexes show high shape complementarity (31). This shape complementarity is dependent on van der Waals interactions between the binding molecules. Electron-electron repulsion prevents atomic overlap and intermolecular penetration. However, induced dipole effects as atoms approach lead to short-range attractive interactions.

The simplest approximation in modeling molecular recognition is to model just van der Waals repulsion by preventing steric overlap (e.g., with an infinite wall potential or a geometric shape description). The short-range attraction is usually accounted for in modeling biomacromolecules with an empirical energy function by means of a Lennard-Jones energy (32, 33):

$$E_{13}[r_{ij}] = -\frac{C_{ij}}{r_{ij}^{12}} + \frac{D_{ij}}{r_{ij}^{6}},$$

where the first term describes the attraction and the second term describes the repulsion between a pair of interacting atoms where $r_{ij}$ is the distance between them. The coefficients $C$ and $D$ depend on the type of the atoms and $D_{ij} = 0.5C_{ij}(r_{ij}^{6}/r_{ij}^{12})$, where...
\( \rho_{ij} \) is the sum of the van der Waals radii of atoms \( i \) and \( j \).\( C_{ij} \) is parameterized according to the Slater–Kirkwood equation based on atomic polarizability and the number of outer-shell electrons of interacting atoms (34, 35). For computational convenience, Lennard-Jones energies and forces are often computed only for pairs of atoms within a defined “cutoff” radius of about 10 Å (35).

**Electrostatic interactions**

Electrostatic interactions are long range, vary with distance as \( 1/r \), and play a significant role in molecular recognition. Therefore, they are particularly important to model when simulating the initial diffusional step of biomolecular recognition. Molecules are characterized not only by their net charges, but also by their charge distributions. For modeling electrostatic interactions, formal or partial charges are assigned to the atoms of the interacting molecules. The simplest model to describe the interactions between charges is Coulomb’s law, in which the electrostatic energy \( E_\text{el} \) of charges \( q_i \) in molecule A and \( q_j \) in molecule B is given by:

\[
E_\text{el} = \sum_{i,j} \frac{q_i q_j}{\epsilon_0 r_{ij}}
\]

where \( r_{ij} \) is the distance between atoms \( i \) and \( j \), and \( \epsilon \) is the relative dielectric constant. The potential of a charge \( q \) at the position of an atom \( i \) at distance \( r_i \) in homogeneous dielectric \( \epsilon \) is \( \Phi = q_i r_i / \epsilon \).

Weaker electrostatic interactions occur in a polar medium such as water rather than in vacuum because of dielectric screening. If the atoms in the system, including the water molecules and ions of the solvent, are modeled explicitly, then a relative dielectric constant \( \epsilon_m \) is usually assigned as 1 (\( \epsilon_m = \epsilon_0 \) or 1, where \( \epsilon_0 \) is the permittivity of free space). However, when the molecules and the medium are treated implicitly, the relative dielectric constant is used as a descriptor of the bulk polarizability of the medium. Proteins tend to have a lower relative dielectric constant than water because many dipolar groups can be considered frozen into a hydrogen-bonded lattice and therefore cannot reorient in an external electrostatic field. Typical assignments of the \( \epsilon_m \) for proteins range from 2 to 20. Water, however, has a high \( \epsilon_m \) around 80 because its dipoles can reorient freely. This dielectric heterogeneity should be accounted for in an electrostatic model of a system with protein molecules in water.

The Poisson equation (or Gauss’ Law) describes the electrostatic potential of a fixed charged density of the solute \( \rho_{\text{tot}}(r) \). The exterior charge density of the ions in the solution, \( \rho_{\text{sol}}(r) \), is modeled by assuming a Boltzmann distribution. The Poisson-Boltzmann (PB) equation is commonly applied to molecules in aqueous solution to compute the electrostatic potential of the system. The general form of the PB equation is:

\[
\nabla \cdot [\epsilon(r) \nabla \phi(r)] = -4\pi \rho_s(r),
\]

where \( \phi(r) \) is the electrostatic potential at position \( r \); \( \epsilon(r) \) is the position-dependent dielectric function; \( \rho_s(r) \) is given by the charge density of the solute, \( \rho_{\text{sol}}(r) \), and of the ions in the solvent, \( \rho_{\text{ion}}(r) \); \( \epsilon \) designates the vector differential operator.

The PB equation may be solved numerically for molecules (for reviews, see References 36–38). The finite difference, finite element, and multigrid methods are used most commonly to solve the PB equation. Usually, this technique is performed by mapping the molecules onto a three-dimensional cubic grid. To solve the PB equation, a suitable interior relative dielectric constant and definition of the dielectric boundary should be assigned (39, 40).

**Hydrogen bonding**

Hydrogen bonds are specific, short-range, nonbonded interactions. They occur between a proton donor of strongly polar groups such as FH, OH, NH, and SH, and a proton acceptor that is strongly electronnegative such as O, F, or N. Weak hydrogen bonds may also be formed by CH groups. The “attraction” of the proton has electrostatic nature. However, quantum-mechanical treatments reveal other contributions to hydrogen bonds. These treatments include the short range repulsion due to the smaller distance between acceptor and hydrogen donor than the sum of their van der Waals radii and the attraction that originates from charge transfer (for review, see Reference 41).

Several empirical analytic forms for the energy of a hydrogen bond have been proposed. A general expression for the energy of a hydrogen bond that describes their directional properties is:

\[
E_{\text{H2O}} = E_x + E_n + E_d = (-C_2 r^{-6} + D_2 r^{-10}) - E_x - E_d
\]

The distance-dependent term may have values of \( m \) and \( n \) of 10 and 12 or 6 and 8, respectively, \( \{E_x, E_n, E_d\} \), the latter are used in the GRAM program (42). The angular terms are functions of the angles made at the donor hydrogen and at the acceptor atom and depend on the chemical types and bonding of the hydrogen bonding atoms.

Hydrogen bonds play an essential role in the stabilization of secondary structures in proteins and in the interactions of macromolecules. In proteins, hydrogen bonds “lock” secondary structure elements such as \( \alpha \)-helices, beta-sheets, and beta-turns. Most hydrogen bonds occur between N-H and O=C groups in the polypeptide chain. Because of the strong hydrogen bonding abilities of water molecules to polar groups in proteins, it is believed that the structural water molecules tighten the protein matrix. However, evidence has been provided for the importance of structural water molecules in increasing protein flexibility (43).

Hydrogen bonds are of great importance for the specificity of molecular recognition at short range. Unspecific hydrogen bonding at the binding interface of two interacting macromolecules will disfavor binding. The charged or polar residues at the binding interface of one macromolecule can form hydrogen bonds to charged groups or to polar groups of another macromolecule. Interfacial water molecules may also mediate molecular recognition, due to their ability to donate and accept hydrogen bonds.
Hydrophobic effect

The hydrophobic effect refers to the favorable interactions between nonpolar surfaces immersed in water. These interactions are considered to provide the driving force for protein folding (44) and to make a major contribution to the stability of protein tertiary structures. The hydrophobic effect also plays an important role in protein interactions (45). The hydrophobicity of protein surfaces has been studied experimentally by affinity partitioning of proteins (44). Theoretical studies have shown that the presence of hydrophobic patches on the surfaces of proteins correlates with protein binding sites (47–49).

In simulations with explicit water molecules, the hydrophobic interactions must result from a complicated interplay of Lennard-Jones and electrostatic interactions between the atoms of the protein and the surrounding water molecules. In the case that the water is represented implicitly, the hydrophobic interactions are usually modeled by an empirical term that depends on the surface area buried on binding:

\[ E_{\text{SASA}} = \sum \alpha \Delta \text{SASA}_i \]

where \( \text{SASA} \) indicates the solvent accessible surface area, and \( \alpha \) is a coefficient that depends on the surface area definition and the atom type (50). The SASA was described first by Lee and Richards in 1971 (51). For computational calculations, the “rolling ball” algorithm of Shackle and Rupley is often used (52). The ball or sphere radius is typically 1.4 Å, which approximates the radius of a water molecule.

The buried surface area between two interacting molecules, A and B, is calculated as the difference between the summated SASA of each molecule separately and SASA of the complex:

\[ \Delta \text{SASA}_{\text{buried}} = \left( \sum_{A} \Delta \text{SASA}_A + \sum_{B} \Delta \text{SASA}_B \right) - \sum \Delta \text{SASA}_i \]

Entropy

Many biological processes such as protein folding, molecular recognition, or more specifically ligand-protein or protein-protein binding, are associated with changes in entropy. Entropy is defined as a measure of “freedom” of the system. Several types of entropy are involved in recognition processes.

On bimolecular binding, an entropic loss develops from the reduction of the translational \( S^{\text{trans}} \) and rotational \( S^{\text{rot}} \) freedom because of the transformation of two unbound molecules into one bound complex. \( S^{\text{trans}} \) and \( S^{\text{rot}} \) for the gas phase may be calculated by the “Sackur-Tetrode” approach (53). Modifications are required to estimate \( S^{\text{trans}} \) and \( S^{\text{rot}} \) in other environments (54).

In addition, the loss in \( S^{\text{trans}} \) and \( S^{\text{rot}} \) has been found to be compensated largely by intermolecular, mostly vibrational, motions \( S^{\text{vib}} \) in the complex. \( S^{\text{vib}} \) is expected to be smaller for high affinity complexes and larger for low affinity complexes (55). \( S^{\text{vib}} \) can be estimated using harmonic or quasiharmonic models of the system (for more details, see Reference 56).

Intuitively, binding is expected to restrict the flexibility of both partners and lead to a significant loss in conformational entropy. To overcome this unfavorable loss of conformational entropy, specific and favorable intramolecular or short-range intermolecular interactions such as hydrogen bonds, salt bridges, hydrophobic interactions and so on are formed. In some cases, the binding of the partners causes increased conformational flexibility of regions distant from the actual binding site, which results in allostERIC EFFECTS (57). The change in conformational entropy due to the protein backbone and side chains is usually estimated by considering the protein as having discrete, isoenergetic, conformational states (e.g., side chain rotamers). Then:

\[ \Delta S = k_B \ln(n_1/n_2) \]

where \( n \) indicates the number of states, and 1 and 2 refer to the initial and final situations. The conformational entropy \( S^{\text{vib}} \) is often subdivided into backbone \( S^{\text{back}} \) and side chain \( S^{\text{side}} \) entropies. \( S^{\text{back}} \) is determined by the probability distribution over the dihedral angle space. The contribution of backbone entropy is generally less significant for binding than for folding. \( S^{\text{side}} \) is usually estimated by rotamer counting, although more detailed probability calculations can be performed (also discussed in Reference 58).

During molecular binding, some water molecules are “released” from the interface of the contacting molecules, which causes a favorable entropic contribution, \( \Delta S^{\text{dielectric}} \). The solvation entropy is commonly computed using methods based on solvent accessible surface areas \( (\text{SASA}) \) (58, 59). However, the binding involves highly charged molecules such as nucleic acids, the SASA approach may not be adequate. Because of the presence of highly charged binding partners in the solvent, additional solvation entropy contributions \( (\Delta S^{\text{dielectric}} + \Delta S^{\text{ionic}}) \) should be taken into consideration. These contributions result from ordering of the water dipoles via long-range electrostatic interactions and the formation of an ionic atmosphere. Sharp et al. (60) calculated \( \Delta S^{\text{dielectric}} \) and \( \Delta S^{\text{ionic}} \) using a Poisson-Boltzmann model in conjunction with van't Hoff analysis and applied it to drug and protein binding to DNA. Water molecules may also be trapped at the interface with an entropic cost, which can be estimated by considering the loss of mobility of the water molecules compared to bulk solvent (61).
simulated process or step. Here, we first discuss the use of Brownian and molecular dynamics simulation methods and then discuss how the level of detail can be altered in constructing models for the simulation of molecular recognition.

**Brownian dynamics simulation**

The Brownian Dynamics (BD) simulation technique can be used to simulate the diffusion and the association of molecules in solution. BD simulations have been widely used to simulate protein–small molecule and protein–protein association (62). This method may be exploited to simulate the first step of molecular recognition when two molecules diffuse from a distance. From such simulations, it is possible to compute the structure and the diffusional encounter complex ensemble and to calculate the bimolecular association rate constant for two diffusing proteins or enzymes and their substrates or inhibitors. In these calculations, the effects of mutations and variations in ionic strength, pH, and viscosity can be investigated (63).

The general principle of BD is based on Brownian motion, which is the random movement of solute molecules in dilute solution that result from repeated collisions of the solute with solvent molecules. In BD, solute molecules diffuse under the influence of systematic intermolecular and intramolecular forces, which are subject to frictional damping by the solvent, and the stochastic effects of the solvent, which is modeled as a continuum. The BD technique allows the generation of trajectories on much longer temporal and spatial scales than is feasible with molecular dynamics simulations, which are currently limited to a time of about 10 ns for medium-sized proteins.

The basic BD algorithm developed by Ermake and McCammon (64) provides an approximate solution to the Langevin equation in the highly damped diffusive limit by using the positions of a solute particle at time t, together with the forces acting on them, to estimate the particle’s new position at time, \( t + \Delta t \). The translational behavior of the particle is described by:

\[
 r(t + \Delta t) = r(t) + D \Delta t + \frac{F}{k_B T} + R
\]

where \( F \) is the force that acts on the particle, \( D \) is the translational diffusion constant of the particle, and \( R \) is a random displacement that mimics the effects of collisions of the particle with solvent molecules. The rotational behavior of the diffusing particle is described by a similar equation to the translational behavior. The translational and rotational diffusion constants of proteins can be estimated by the Stokes-Einstein relationship, which depends on the solvent viscosity and the radius of the protein. Larger proteins diffuse slower.

For simulation of molecular recognition by BD to compute association rate constants and for the generation of the structures of diffusional encounter complexes, the following setup may be used. One solute, which usually is the larger one, is positioned at the center of simulation space. The second molecule diffuses from randomly chosen starting points at a specified distance, and many trajectories are generated. The probability of formation of diffusional encounter complexes is computed from these simulations. The method of Northrup et al. (65) allows computation of the association rate constant by combining an analytical solution to the diffusion equation with the numerical data from the BD simulations.

The programs most widely used for simulation of the diffusional encounter of two biomolecules are UHBD (66, 67), Macromod (68), and SDA (66, 67).

**Molecular dynamics simulation**

Molecular dynamics (MD) trajectories can be used to investigate the structure, dynamics, and thermodynamics of biological molecules and their complexes. The motions of the atoms in the system are simulated according to Newton’s equations of motion. The forces on the atoms are calculated using a molecular mechanics energy function. The force field for a protein is given by the sum of various components including bond stretching and bond angle bending, torsional potential, and nonbonded interactions (Lennard-Jones and Coulombic interactions). Several different algorithms may be used to generate snapshots of the system using a time step of about 1 fs (69, 70).

The motions of proteins are usually simulated in aqueous solvent. The water molecules can be represented either explicitly or implicitly. To include water molecules explicitly implies more time-consuming calculations, because the interactions of each protein atom with the water atoms and the water molecules with each other are computed at each integration time step. The most expensive part of the energy and force calculations is the nonbonded interactions because these scale as \( N^2 \) where \( N \) is the number of atoms in the system. Therefore, it is common to neglect nonbonded interactions between atoms separated by more than a defined cut-off (~10 Å). This cut-off is questionable for electrostatic interactions because of their UV dependence. Therefore, in molecular dynamics simulations, a Particle Mesh Ewald method is usually used to approximate the long-range electrostatic interactions (71, 72).

MD simulations are used mainly to investigate in detail the interactions of two associating macromolecules at relatively short distances, rather than to simulate their coming together from afar. Thus, an encounter complex or an approximate protein–ligand complex can be refined using MD simulations. Indeed, MD is an important tool in macromolecular complex structure determination in which it is used with simulated annealing and experimental constraints. Important applications of MD simulations include the computation of the binding free energies and intermolecular forces.

**Coarse graining of a simulation model**

Most biomolecular binding events occur on timescales orders of magnitude greater than can be attained in standard molecular dynamics simulations in which the motion of all atoms is
Molecular Docking: Prediction of Bound Complexes

Whereas dynamic simulations may be used to study the process of molecular binding and unbinding and to predict the structure of the bound complex, many computations are aimed solely at prediction of the structure of the bound complex. This is the so-called “docking problem.” Although this problem can be addressed by using simulations to mimic the physical binding process (77), many other approaches may be used to sample molecular arrangements and to predict the bound complex. Energy minimization from systematically or randomly distributed starting arrangements may be performed subject to a detailed or simplified force field (e.g., for protein-protein docking, (74)). Simulated annealing and Monte Carlo methods include temperature to permit high-energy configurations to be sampled and thereby energy barriers to be overcome (e.g., AutoDock, (78)). Genetic algorithms and particle swarm optimization methods are biologically inspired procedures for optimizing a function efficiently that have been applied to small molecule docking (e.g., GOLD (79), PLANETS (80)) and to protein-protein docking (81). They may be applied with driving “forces” for binding that range from simple geometric complementarity (82) to detailed molecular mechanics type force fields. Systematic searching is widely used. This can be facilitated by FFT methods, which were applied initially to protein-protein docking by optimization of geometric fit (83). Because the direct calculation of the correlation between two functions results in an order of N^6 computing steps, Katchalski-Katzir et al. (83) used Fourier Transforms to calculate the correlation function more rapidly. With this algorithm, the computing time is proportional to N^3 ln N^3, where N is the number of atoms in the complex. FFT methods have been used to optimize not only geometric fit but also hydrophobicity/complementarity (48) and electrostatic complementarity (DOT) (84).

Optimization may be performed in continuous or discrete space. For example, in optimizing the conformations of side chains in proteins, a discrete rotamer representation is often used. In 2003, Gray et al. (85) presented a docking program RosettaDock that allows the side chain conformational stiffness of the interface residues on both interacting partners to change. The side chain flexibility was modeled using a simulated annealing algorithm that searches through backbone-dependent rotamers from the Dunbrack rotamer library (86). The side chains of the partners are rebuilt from rotamers before docking (prepacking) and only the side chains of the interface are then refined later in docking. Additional improvements of this method, such as implementation of a torsion minimization step to sample off-rotamer space and information on side chain packing from the unbound structures, are described in Reference 87. However, in some cases, which may be binding hot spots, side chains adopt nonideal-rotamer conformations and these should be accounted for (88). In the ATTRACT approach of Zacharias (74), several conformations are considered for each side chain simultaneously. Side chain copy selection is coupled to energy minimization in translational and rotational degrees of freedom of the mobile protein partner in an iterative way. This technique is effective for complexes in which side chain conformational changes occur during complex formation. Similar to side chain mobility, this approach can also be extended to treat the flexibility of protein loops.

Conclusions and Future Directions

Biomolecular recognition can be described as a balanced interplay of different forces and energetic contributions. The long-range electrostatic forces together with the short-range van der Waals interactions, shape complementarity, hydrogen bonding, and hydrophobic effects determine the interactions of associating molecules. The flexibility of the partners on their binding is important for induced fit and makes an entropic contribution to the recognition process. Modeling of the structural, dynamic, and energetic determinants provides a tool to simulate the interactions of the macromolecules on their complexation. Different computational docking programs model the forces and the energetic contributions differently. Some docking programs mimic the diffusional association of the molecules; others base their approach mostly on the shape complementarity of associating partners. The varying degrees of the incorporation of electrostatic and hydrophobic interactions and the structural plasticity of the molecules on binding influence the accuracy.
of the prediction of the structure of the bond complex and the ability to compute kinetic and thermodynamic parameters. Because of the large number of degrees of freedom of biologically active molecules in solution, the problem of sampling and scoring of the docked solutions is very challenging. The computational methods must aim at simple, computationally feasible models that provide reasonable accuracy and help to answer the question of whether a given molecule can recognize and bind with their partners selectively and specifically. Current protein–protein docking algorithms can provide good predictions of the structures of complexes when the conformational changes on binding are very limited, but they usually fail when large conformational changes occur (89, 90); see also the CAciPRi experiment for the comparative evaluation of protein–protein docking for structure prediction (http://caspabi.ebi.ac.uk). Ongoing improvements of the computational models for biomolecular interactions include the treatment of large-scale conformational changes of the macromolecules on binding and the modeling of solvation and polarization effects, among others.

References


Molecular Recognition, Computation and Modeling of


Further Reading


See Also

Computational Chemistry in Biology
Non-covalent Interactions in Molecular Recognition
Protein-Nucleic Acid Interactions
Protein-protein Interactions
Receptor-ligand Interactions in Biological Systems
Receptor–Ligand Interactions in Biological Systems

Liwei Li, Department of Biochemistry and Molecular Biology and the Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis, Indiana
Samy O. Meroueh, Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, and the Department of Chemistry and Chemical Biology, Indiana University Purdue University, Indianapolis, Indiana
doi: 10.1002/9780470048672.wecb506

Despite the significant research that has been invested in understanding molecular recognition in biological systems, accurate prediction of macroscopic properties based on microscopic interactions remains elusive, which makes it difficult to identify systematically tight binding inhibitors in computational drug design. In the past, most ligand design efforts have centered on descriptors derived mainly from structure, neglecting entropic effects that develop from receptor flexibility. The lack of explicit incorporation of receptor motion has meant that compensatory effects between enthalpy and entropy, which are essential for the accurate estimation of free energy, have also been neglected. In addition, cooperative effects, which develop from alteration in the motion of the receptor because of ligand binding, are not captured. These effects pose a challenge for the design of small molecules that allosterically modulate protein-protein and protein-small-molecule interactions. Here, we describe macroscopic properties of receptor-ligand interactions, which are followed by a discussion of how these properties are predicted with microscopic interactions. A discussion then ensues on some fundamental aspects of the molecular recognition process that have attracted renewed attention, such as conformational selection versus induced-fit binding, entropy-enthalpy compensation, and allostery.

Introduction

Molecular recognition leads to the association of macromolecules, which enables them to perform functions that are vital to the survival of an organism. Evolutionary forces have exploited interactions between macromolecules for the purpose of performing one of a multitude of tasks, including chemical modification (1), transportation (2), signal transduction (3), or recruitment of other macromolecules (4), among other things. The recognition process is driven by a balance between potential energy that develops from physical interactions at the molecular level (5) and kinetic energy that originates from inherent dynamics contained in molecules (6). The physical interaction between macromolecules is driven by various interactions most prominent among them electrostatics, which develops from the positive and negative charge of nuclei and electrons (6) and the hydrophobic effect (7-11). These charges impart attractive and repulsive characteristics to molecules. But it is often overlooked that the dynamics of molecules also play an important role in the recognition process (6).

In solution, molecules constantly encounter each other through collisions. At close distances, repulsions from the nuclei dominate. The balance between repulsive and attractive forces often lead to an equilibrium state such that two molecules will remain associated in a complex. In this state, dispersion or London forces dominate. The motion experienced by the molecules also plays a crucial role in this process. A molecule contains translational, rotational, as well as internal vibrational motion. When two molecules associate, conservation of energy dictates...
that the energy that is initially contained in translation and rotation must be eventually absorbed internally by the molecules through vibrations (12). The failure to effectively absorb this energy could lead to a disruption of the balance between attractive and repulsive forces and the abrogation of the complex.

A thorough understanding of macromolecular and their ligands is essential to the survival of an organism. It is often the case that these same interactions will contribute to pathological processes. Examples include the binding of a growth factor to its cell surface receptor (3, 13), a matrix-metalloproteinase binding to its ligand (14), the binding of a kinase to its ATP substrate (15). Each of these interactions has been targeted by pharmaceutical companies for the development of drugs. Inhibition of the receptor tyrosine kinases has been particularly successful. Targeting the ab kinase, for example, has led to the discovery of Gleevec (Novartis, East Hanover, NJ), which is one of the most effective anticancer therapeutics to date (16). This drug was rationally designed such that structural information of its target was used during the drug-discovery process. It is likely that early version of the drug underwent a series of chemical modifications to optimize not only affinity to the target but also ADME (adsorption, distribution, metabolism, and excretion), toxicology, and pharmacokinetic properties (17). But high affinity to the drug target remains a key criterion that must be constantly maintained. In addition, with the multitude of targets that are now available, selective or broad-spectrum inhibitors are also desirable. Despite the significant progress that we have made over the past decade toward understanding molecular recognition, the systematic ability to design molecules with high affinity to a target based on its structure remains elusive (18).

The focus of this review article will be on the interaction between macromolecules and small-molecule ligands. The discussion will first center on the thermodynamic and kinetic characteristics that are used to measure the extent of binding. Subsequently, we discuss the interactions at the atomic level that drive complex formation. Then, a discussion follows of some tools available to predict macroscopic properties from microscopic properties. We then briefly discuss molecular motions as well as various aspects of receptor-ligand that have attracted renewed attention, such as conformational selection versus induced-fit, entropy-entropy compensation effect, and protein allosterism.

**Macroscopic Properties**

In solution, the mixing of two entities leads to the formation of a chemical equilibrium that is represented by the following equation:

\[ \text{M (aq)} + \text{L (aq)} \rightleftharpoons \text{ML (aq)} \]

where M is the macromolecular receptor, L is the small molecule ligand, and ML represents the complex between the receptor and ligand. Depending on the strengths of the interactions, the equilibrium can favor the formation of reactants or products. The direction of the equilibrium is measured by a ratio known as the equilibrium constant, which is also known as the dissociation constant \( K_D \). The association constant \( K_A \) defined as follows:

\[ K_A = \frac{[\text{ML}]}{[\text{M}][\text{L}]} \]

A measure of the spontaneity of the chemical event is encapsulated in the Gibbs free energy, which is related to the equilibrium constant following this relation:

\[ \Delta G = -RT \ln K_A \]

\( \Delta G \) is the free-energy change, \( \Delta H \) is standard free-energy change, \( R \) is the gas constant whose value is \( 8.31 \) J mol\(^{-1}\) K\(^{-1}\), and \( T \) is the temperature.

A chemical event that results in a reduction of the free energy is referred to as exergonic, whereas a process that results in an increase in the free energy is known as endergonic. Typical values of dissociation constants \( K_D \) for receptor-ligand interactions occur between \( 10^{-3} \) (millimolar) and \( 10^{-9} \) (nanomolar) (19–21), although reports suggest values that reach the pico- to a change in the free energy \( \Delta G^\circ \) of 1.4 kcal mol\(^{-1}\). Hence, \( \Delta G^\circ \) for a millimolar ligand is \(-4.2\) kcal mol\(^{-1}\), whereas that of a nanomolar ligand is \(-9.8\) kcal mol\(^{-1}\). Another well-known equation relates the free energy to the enthalpy \( \Delta H^\circ \) and entropy \( \Delta S^\circ \):

\[ \Delta G^\circ = \Delta H^\circ - T \Delta S^\circ \]

The enthalpy is defined by the following expression:

\[ \Delta H^\circ = \Delta E + \Delta P \Delta V \]

where \( E \) is the internal energy of the system, \( P \) is the pressure, and \( V \) is the volume. Most chemical events occur at constant volume and pressure, and hence the enthalpy can be conceptualized as a measure in the internal energy change. The formation of favorable interactions, such as hydrogen bonds or van der Waals contacts will result in a negative enthalpy change and hence an exothermic process. The loss of favorable interactions leads to a positive enthalpy or endothermic process.

Although the enthalpy is a measure of order in the system, the entropy is a measure of disorder. The second law of thermodynamics states that a spontaneous event results in an increase in entropy. Hence, a favorable process always results in an increase in disorder. The entropy is defined as:

\[ S = k \ln \Omega \]

where \( k \) is the entropy and \( \Omega \) is the phase space that corresponds to all the possible states that a system may adopt. In Newtonian mechanics, a state can be defined by position and velocity. In macromolecules, position can be translated into internal coordinates, such as bonds and torsions that define the conformation of the macromolecule. In the context of receptor-ligand binding, the change in the entropy is the log of the ratio of phase space of the receptor—ligand complex.
and the phase space of the product of the receptor and ligand. The number of states that are accessible for the free receptor and ligands will be significantly larger than that of the complex. Hence, the association of macromolecules and their ligands typically results in entropy loss (22). In some instances that involve the burial of hydrophobic residues, the rearrangement of water could lead to a favorable contribution from entropy (vide infra).

**Microscopic Interactions**

Molecules can have a net positive defined as negative charge depending on the hybridization state of atoms within them. In solution, the net charge often originates from ionizable groups that act as a Brønsted base or acid following this equation:

\[ \text{A}^+ + \text{H}_2\text{O} \rightleftharpoons \text{HA} + \text{H}_3\text{O}^+ \]

Carboxylate, thiols, or amino moieties are a few examples of ionizable groups at physiological pH. The propensity of an ionizable group to lose a proton is measured by the \( pK_a \), which is defined by the following:

\[ pK_a = \log \left( \frac{[\text{HA}]}{[\text{A}^-]} \right) \]

where \( K_a \) is the acid dissociation constant, \( [\text{A}^-] \) and \( [\text{HA}] \) are the concentrations of acid and its conjugate base. Based on this equation, the \( pK_a \) can also be viewed as the pH at which the concentration of base is equal to that of the acid. For example, at neutral pH, the \( pK_a \) of the amino group of a lysine residue is 10.4, whereas that of a glutamic acid residue is 4.5 (\( pK_a \) values are from Refs. 23 and 24). In the former, the amino group is protonated and positively charged at physiological pH, whereas the glutamic acid is unprotonated and negatively charged. Other examples of amino acids with positively charged side chains include the guanidine side chain of arginine (\( pK_a \) = 12.5). A another example of an amino acid with a negatively charged group include aspartic acid (\( pK_a \) = 4.0).

Within macromolecules, however, the \( pK_a \) of ionizable groups can deviate from their standard values, as the groups are not only surrounded by solvent but also by other residues. For example, a carboxylic acid moiety located in a region of low polarity or hydrophobicity will experience an increase in its \( pK_a \). A positive charge near the carboxyl group will reduce its \( pK_a \), whereas a negative charge will lead to a more basic side chain. Conversely, an amino group in a hydrophilic region will experience a reduction in its \( pK_a \), whereas a salt bridge will increase its \( pK_a \). A prominent example that illustrates these trends occurs in the TEM-1 \( \beta \)-lactamase, which is a resistance enzyme in bacteria that destroys \( \beta \)-lactam antibiotics (25). The active site contains two lysine residues and two serine residues in a cross-like arrangement (Fig. 1). This arrangement is universally conserved not only among the \( \beta \)-lactamases but also among their ancestors, the penicillin-binding proteins (PBPs) (26). A lysine residue that is essential for catalysis is found in a hydrophobic pocket that is shielded from direct contact with solvent (Fig. 1).

A nearby glutamate side chain (Glu-166) forms a salt-bridge interaction with Lys-73. Biochemical analyses had described a \( pK_a \) of 5 to Glu-166, nearly 0.5 units higher than that of standard \( pK_a \) of 4.5. An extensive set of biochemical, nuclear magnetic resonance (NMR), and molecular dynamics-based free-energy calculations led to a \( pK_a \) value of 8 for Lys73, which is nearly 2.8 units lower than its standard \( pK_a \). These \( pK_a \) shifts can be rationalized from the microenvironment around the residues. The reduced \( pK_a \) of Lys73 is promoted by the low dielectric medium that surrounds its amino group along with the positive charge of a nearby lysine (Lys234). The negative charge that emanates from Glu166, on the other hand, stabilizes the positive charge on Lys73 and thus raises its \( pK_a \). Interestingly, in the PBPs, which lack a glutamate in its active site (27), the \( pK_a \) of the catalytic lysine residue is reduced even more to 6, such that the amino group on the lysine side chain is neutral (25). The neutrality of the side chain is essential for this enzyme to carry out its transpeptidase function (27).

Electrostatic interactions, along with hydrophobic effects, form the basis of most intermolecular forces that drive receptor-ligand interactions in biological systems. The simplest form of electrostatic interaction is the salt bridge, which occurs between two oppositely charged functional groups (28). The formation of such an interaction within proteins is usually energetically costly (29). An example is the Lys73 and Glu166 salt bridge that was alluded to above (Fig. 1). Other examples of charge-charge interactions in biological molecules include the coordination of transition metals with negatively charged moieties such as carboxylates from glutamate and aspartate side chains or thiolates from cysteine residues (30, 31). Although opposite charges form strong interactions, their presence alters the charge distribution of surrounding residues as well. A fixed charge creates an electric field that emanates from it (for a positive charge), electrons on surrounding atoms will be polarized depending on the nature of the field. For a positive charge, the electron density on surrounding atoms will move toward the charge, whereas a negative charge will cause electrons to move away. This movement of electron density creates a dipole. The region where the electron density is accumulated will be negatively charged, whereas the region that lost electron density will develop a partial positive site. Some dipoles are permanent (32). A prominent example is that of water, in which the electrons accumulate on the electronegative oxygen atom, leaving a partial positive charge on the hydrogen atoms and a permanent dipole that can be visualized as a vector that is perpendicular to the base of the HOH triangle (33). It is worth mentioning that the interaction energy between a charge and a dipole is inversely proportional to the separation to the fourth power, whereas a dipole–dipole interaction is inversely proportional to the separation to the sixth power. Hence, these interactions dominate at short separations, whereas charge–charge interactions dominate at large separations.

A ubiquitous interaction that is electrostatic in nature and commonly occurs in biological systems is the hydrogen bond (see Fig. 1 for examples). It is defined as the interaction between an electronegative atom (acceptor) and a hydrogen atom that is covalently attached to another electronegative atom (donor) (34, 35). At the fundamental level, a hydrogen bond has been
described as a dipole–dipole interaction, but it is likely more complex mixture of physical effects that include covalent characteristics, polarization, as well as charge transfer (34-37). A common hydrogen bond in proteins occurs between an NH bond and an oxygen atom from a carbonyl carbon with an NO distance that falls between 2.8 and 3.4, and between 1.5 and 2 Å for the OH distance. A recent series of surveys conducted on protein–ligand complexes from the Repository Collaboratory of Structural Bioinformatics find that the strongest hydrogen bonds occur for the NH–O (ligand is donor and backbone is acceptor), NH–O (side chain is donor and ligand is acceptor), OH–O (ligand donor and backbone is acceptor), and OH–O (side chain donor and ligand acceptor) bonds with donor-hydrogen-acceptor angles near linearity (38). The textbook definition of the hydrogen bond mentions only N, O, S as donors, but recent work has suggested that carbon can also act as donor atom in the formation of weaker CH–X hydrogen bonds (39, 40), where X = C, N. A comprehensive database analysis suggested that CH–O hydrogen bonds occurred extensively in alpha helices (41). Solvent tends to weaken the CH–O hydrogen bond (42).

Other atoms, such as halogens, have been suggested to act as a bridge between a hydrogen bond donor and acceptor (43). In addition to promoting more favorable intermolecular interactions between receptor and ligand, electrostatic interactions play an important role in the interaction between water and solutes. Water molecules have the tendency to form ordered structures, although it is worth mentioning that the extent of ordering has been difficult to determine (44). Hydrogen bonds have a tendency to favor water ordering at 25 °C and are hence immiscible in water (45). This effect is also observed around nonpolar surfaces of macromolecules in which water molecules are more ordered that in bulk water, resulting in an entropic penalty (46). This effect is known as the hydrophobic effect and is caused by the special hydrogen bonding properties of water (7-11).

A key signature of the hydrophobic effect is a change in the heat capacity (47). It is widely accepted that this effect plays an essential role in protein folding, as hydrophobic residues tend to aggregate at the core of folded proteins whereas hydrophilic moieties are found on the exterior of proteins, maximizing contact with the water solvent (47). In the context of protein–ligand interactions, the hydrophobic effect is defined as the additional stability that is gained as a result of binding in water, specifically caused by the removal of nonpolar surfaces from exposure to water during complex formation. In the case of a small-molecule ligand, the burial of nonpolar functional groups into a protein cavity will be favorable, especially when the cavity is also hydrophobic in nature; here, the hydrophobic effect can exhibit a favorable enthalpic component (48). The magnitude of the hydrophobic effect is proportional to the amount of surface buried as a result of binding.

### Predicting Macroscopic Properties From Microscopic Interactions

The 1990s experienced a surge in the number of X-ray and NMR structures that were solved for macromolecules and their complexes. This surge coincided with 1) a rapid increase in computer power and 2) an increase in the availability of computational chemistry programs that could perform molecular dynamics simulations as well as quantum chemical calculations on systems of biological relevance (49, 50). These changes emboldened scientists to apply computational techniques to biological systems that were otherwise not previously amenable to accurate calculations. Early attempts were mostly limited at predicting the binding mode of compounds, which is a process known as docking. One of the first computer programs to perform docking was DOCK, written by Erwin Kuntz and coworkers at the University of California San Francisco (51). The process of docking consists of sampling many conformations of the ligand and bound to the receptor until the most energetically favorable complex is found (52). Today programs perform docking with both ligand and receptor flexibility (53), including GOLD (54), AutoDock (55), and FlexX (56). The main impediment with docking, however, has been that the accuracy of scoring function that is used to treat the interaction between ligand and receptor (57). Scoring functions were devised with the purpose of estimating the binding free energy of a molecule to its receptor. Several techniques have been used, which include empirical (58-60), knowledge-based (61-64), or physics-based methods (51, 65-67). But insufficient sampling of configurational space (docking typically is conducted on a single snapshot, with at most local flexibility), as well as the lack of accuracy in the force fields used to represent the interaction between compound and receptor, have hampered efforts at obtaining more accurate approximation of the free energy of binding (68). Scoring functions were devised as an alternative to highly intensive methods for computing the free-energy binding. The compute-intensive methods are typically based on molecular dynamics.
Receptor–Ligand Interactions in Biological Systems

![Figure 2](image)

Stereoview that illustrates the motion that is experienced by a protein–ligand complex during molecular dynamics simulation.

Dynamics simulations (MD), which consist of solving Newton’s equations of motion to determine the position and velocities of atoms with respect to time (69). With these simulations, one can gain detailed insight into the dynamics of the system, which can be used to estimate entropy of binding (Fig. 2) (70). Several packages have facilitated the process of carrying out MD simulations for biological systems; these methods include AMBER (71), CHARMM (72), and GROMACS (54, 73) among others. Two requirements exist for conducting an MD simulation, namely a three-dimensional structure, initial velocities for all atoms in the system, and the potential energy of the system (69). The structure is usually obtained from x-ray or NMR. The velocities can easily be obtained from the temperature. The potential energy is obtained through what is known as a force field, which is a mathematical representation of the energy of the system (68). A typical force field contains terms to represent bonds, bonds, bends, torsions, and intermolecular interactions in a system, and it takes the following form:

\[
E_{\text{AMBER}} = \sum_{\text{bonds}} k_b (r - r_{\text{eq}})^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_{\text{eq}})^2 \\
+ \sum_{\text{strain}} k_\gamma (\gamma - \gamma_{\text{eq}}) + \left(1 + \cos(n \theta - \gamma)\right) \\
+ \sum_{\text{dihedral}} k_\phi R_{\phi}^2 + \frac{\phi}{R_{\phi}^2} + \frac{A_{\phi}}{R_{\phi}^2} \\
\]

here, \(E_{\text{AMBER}}\) is the potential energy, \(r, r_{\text{eq}}\) and \(\theta, \theta_{\text{eq}}\) are equilibrium distance and angles for bonds and bends, respectively, and \(k_b, k_\theta, k_\gamma\) are force constants; \(n\) is multiplicity and \(\gamma\) is the phase angle for the torsional angle parameters. The accuracy of force fields remains the main impediment for the accurate computation of free energies based on molecular dynamics simulations (68).

Several methods based on molecular dynamics simulations have been devised to compute the free energy of binding (22, 74), including free-energy perturbation (FEP) and thermodynamic integration (TI). These methods are most useful for estimating the free-energy change as a result of the substitution of a functional group into another (e.g., the conversion of a hydrogen atom (initial state) to a methyl group (final state) (75)). In practice, these techniques consist of running multiple simulations at different values of a coupling parameter that ranges from 0 to 1. Depending on the substitution that is being made, the coupling parameter is multiplied to the relevant terms in the force field such that the initial state gradually converts to the final state. We have recently applied the thermodynamic integration approach toward estimating the pK\(_a\) of a residue in the class A \(\beta\)-lactamase (25) and the penicillin-binding proteins with remarkable success (76). The deprotonation reaction constitutes the full thermodynamic cycle for free-energy variations in aqueous condition, which forms the basis for Eqs. (1–3), and was used to calculate the pK\(_a\) of the lysine residue.

\[
\begin{align*}
\Delta\Delta G_{\text{aq}}(\text{A} - \text{AH}^+) & = \Delta\Delta G_{\text{aq}}(\text{B} - \text{BH}^+) \\
\Delta\Delta G_{\text{aq}}(\text{BH}^+ - \text{AH}^+) & = \Delta\Delta G_{\text{aq}}(\text{B} - \text{A}) \\
\Delta\Delta G_{\text{aq}}(\text{A} - \text{B}) & = \Delta\Delta G_{\text{aq}}(\text{AH}^+ - \text{BH}^+) + \Delta\Delta G_{\text{aq}}(\text{B} - \text{A}) \\
\end{align*}
\]

where \(\Delta\Delta G_{\text{aq}}(\text{X} - \text{XH}^+)\) is for the protonated Lys73 and \(\text{BH}^+\) is for the protonated Lys306 (see Fig. 3). The pK\(_a\) value of Lys306 is expected to be that of a typical lysine (because it is a surface residue). The following equation was used to compute the pK\(_a\) of \(\text{AH}^+\),

\[
\text{pK}_a(\text{AH}^+) = 0.729 \times \Delta\Delta G + \text{pK}_a(\text{BH}^+) = 10.8
\]

where \(\Delta\Delta G = \Delta\Delta G_{\text{aq}}(\text{A} - \text{AH}^+) - \Delta\Delta G_{\text{aq}}(\text{B} - \text{BH}^+)\) (in kcal mol\(^{-1}\)). The constant 0.729 is the value of 1/(2,303 RT) with gas constant \(R = 1.984 \times 10^{-3}\) kcal K\(^{-1}\) mol\(^{-1}\) and temperature \(T = 300\) K, and \(\text{pK}_a(\text{BH}^+) = 10.8\) for the unperturbed surface lysine in the aqueous milieu. \(\Delta\Delta G_{\text{aq}}(\text{X} - \text{XH}^+)\) was computed using the thermodynamic integration approach, which is the free-energy difference between the unprotonated state \(\text{X}\) and protonated state \(\text{XH}^+\),

\[
\Delta\Delta G_{\text{aq}}(\text{X} - \text{XH}^+) = \Delta\Delta G_{\text{aq}}(\text{X}) - \Delta\Delta G_{\text{aq}}(\text{XH}^+)
\]

With the thermodynamic integration method the difference is evaluated with a 12-point Gaussian quadrature.

\[
\Delta\Delta G(\lambda) = \int_{0}^{1} \left[ f_{\text{Vf}}(\lambda) - f_{\text{Vb}}(\lambda) \right] \ d\lambda = \sum_{i=1}^{12} w_i \left( \frac{f_{\text{Vf}}(\lambda_i) - f_{\text{Vb}}(\lambda_i)}{R(\lambda_i)} \right)
\]

where \(w_i\) and \(\lambda_i\) are known (77).
Receptor–Ligand Interactions in Biological Systems

However, when it comes to computing the free-energy difference between two very different molecules, FEP or TI are not suitable (78). Other techniques have been devised for that purpose, such as the MM-PBSA (79, 80) and LIE (81) methods. The MM-PBSA approach is based on the following thermodynamic cycle:

\[
\Delta G_{\text{binding}} = \Delta G_{\text{gas}} + \Delta G_{\text{sol}} + \Delta G_{\text{bind}}^{\text{th}}
\]

The free-energy change \( \Delta G_{\text{binding}} \) is thus expressed by the following equation:

\[
\Delta G_{\text{binding}} = \Delta G_{\text{gas}} + \Delta G_{\text{sol}} - \Delta G_{\text{bind}}^{\text{th}}
\]

Induced-Fit Versus Conformational Selection

More than 100 years ago, Emil Fischer had proposed the lock-and-key hypothesis to explain how an enzyme associates with its substrate (97). He proposed that the substrate (key) adopts a shape that is highly complementary to the receptor (lock), which suggests that an enzyme was predisposed to bind to its substrate. But crystal structures of macromolecules in their apo and liganded states have shown that significant conformational changes can often occur during complex formation (98–100). An update of the lock-and-key model, which is known as the induced-fit hypothesis, was later put forward by Koshland (101). It assumes that the enzyme or macromolecule does not sample supplementary structures to the ligand, and hence the necessary conformational changes required for complex formation are caused directly by the ligand on binding.

An alternative to the induced-fit model has been proposed in light of a growing body of data suggesting that proteins are inherently highly dynamic and hence sample many conformational states (90, 102–104). This model is often referred to as the pre-existing equilibrium or conformational selection model as an alternative to the induced-fit model. It is closely related to the Monod-Wyman-Changeux (MWC) model, which was introduced more than four decades ago to describe allosteric effects in hemoglobin (105). They proposed that a protein exists in an ensemble of conformational states that co-exist in equilibrium; the ligand will bind to one of these conformations and shift the equilibrium toward complex formation. Several authors have discussed these matters extensively, including Hecht et al. (106) and Miller and Dill (107). A prominent example in support of conformational selection was highlighted more than a decade ago by Foote and Milstein (108). Using stopped-flow fluorescence experiments, they found that antibodies existed at an equilibrium between different conformers, with ligands binding preferentially to one form. More recently, James et al. (109) employed X-ray crystallography and pre-steady-state kinetics to propose that antibodies adopt multiple conformers and thereby bind unrelated antigens. But they also suggest the possibility of induced fit on association of receptor and ligand. The conformational selection model was recently supported with computation using normal mode analyses (110).

We recently studied conformational change caused by ligand binding in the bacterial ribosome acyltransfer site (A site), where antibiotics are known to lead to mistranslation and the eventual demise of the organism (111). The conformational change was discovered after the elucidation of the three-dimensional crystal structure of the 30S ribosome (112). It entails the flipping of two bases A1492 and A1493 from an intrahelical stacked conformation to an extrahelical conformation that is completely solvent exposed. We have recently

The free-energy change \( \Delta G_{\text{binding}} \) is thus expressed by the following equation:

\[
\Delta G_{\text{binding}} = \Delta G_{\text{gas}} + \Delta G_{\text{sol}} - \Delta G_{\text{bind}}^{\text{th}}
\]

In practice, the method consists of running a molecular dynamics simulation of the receptor ligand complex, then collecting a series of snapshots, which is followed by energy calculation, solvation energy calculation, and entropy calculation for complex, receptor, and ligand. We have applied this method to protein–ligand as well as nuclei-acid interactions in the past. The method seems to perform best in calculations of differences in free energy (82), but sometimes it performs well for absolute value calculations of the free energy (83, 84).
find that the binding of vancomycin to the acyl-D-Ala-D-Ala tidyl stem, which terminates in acyl-D-Ala-D-Ala. The study (NAM). The NAM residues are incorporated with a pentapep-

Enthalpy-Entropy Compensation (EEC)

It has been known for several decades that the formation of complexes between molecules through weak interactions results in a change of the enthalpy that is accompanied by an opposite change in the entropy (114-116). This effect, which is known as enthalpy-entropy compensation (EEC) (117), can be a hindrance in drug-design efforts. For example, a recent effort at optimizing a picomolar HIV-1 inhibitor based on the crystal structure of the inhibitor bound to the active site resulted in a compound with similar binding affinity (118). The addition consisted of replacing a thioester group by a sulfonyl group. The original compound had $\Delta H = -8.2 \text{ kcal/mol}$ and $\Delta S = -6.7 \text{ kcal/mol}$. The optimized compound showed a gain in enthalpy by 3.9 kcal/mol, which signifies that more favorable interactions were created. However, an equal and opposite compensation of the entropy resulted, leading to no gain of free energy.

Our recent work on the study of binding of the antibiotic van-
comycin to bacterial cell wall with isothermal titration calorime-
try also illustrated the effect of entropy enthalpy compensation (119). The backbone of the cell wall is made up of alter-
ating N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM). The NAM residues are incorporated with a pentapep-
tidyl stem, which terminates in acyl-D-Ala-D-Ala. The study found that the binding of vancomycin to the acyl-D-Ala-D-Ala fragment resulted in the same free-energy change as the second vancomycin binding to cell wall fragment with two NAG-NAM units. Despite the similarity in the binding affinity, the entropy and enthalpy contributions for binding to the cell wall fragment were different, and the 5 kJ/mol gain in enthalpy was compensated by a loss of 5 kJ/mol in entropy (119).

It has been argued that EEC is a mere artifact (120). In an attempt to establish the existence of the EEC effect, a collection of thermodynamic data collected for the binding of 136 drugs to their receptors revealed strong correlation between enthalpy and entropy (116). Some have attributed these results to inaccuracies of the van’t Hoff analysis that is used for the measurement of thermodynamic parameters (121). But isothermal titration calorimetry data (ITC) does not suffer from the same inaccuracies that plague the van’t Hoff analysis. We have recently created a dataset that contains more than 400 entries containing structural and isothermal titration data (122).

A plot of the enthalpy and entropy results in the following linear relationship:

$$\Delta H (\text{kcal mol}^{-1}) = -8.2154 + 282.89 \Delta S (\text{kJ mol}^{-1})$$

Aalysis of the data from our dataset has firmly established the existence of the EEC effect (122).

Nearly a decade ago, Dunitz et al. (114) attempted to ratio-
nalize the EEC effect with the use of statistical thermodynamic concepts. Using a model for two covalently attached atoms whose bond is described with a Morse potential, he showed that the vibrational entropy decreased exponentially with the frequency of vibration of the bond. In the context of macro-
molecules, this signifies that the high frequency motions such as bond stretching and bending that occur with frequency greater than 1000 cm$^{-1}$ contribute little to the entropy, whereas lower frequency motions, such as those that involve entire domains, make the largest contribution. Using a hypothetical model of a water molecule bound to a macromolecule with a dissocia-
tion enthalpy of 5 kcal/mol, the author estimated an entropic contribution of 4.5 kcal/mol, nearly completely balancing the enthalpy. A plot of the vibrational entropy versus the disso-
ciation enthalpy revealed that for weak interactions (less than 5 kcal/mol), the contributions of enthalpy and entropy compen-
sated each other. These results support earlier plots by Searle et al. (115) that revealed a similar effect from a curve de-
rived based on qualitative thermodynamic considerations. It is of interest to note that the plot of entropy versus enthalpy monoton-
ically increased with respect to enthalpy, which is likely caused by the neglect of the role of solvation in the model.
water molecules generate favorable entropy that should dampen the steep enthalpy increase. These early theoretical treatments of EEC have been followed by more recent work by Ford (123) who argues that in the gas phase, EEC is not a general feature of association. He bases this argument on the observation that both compensation and anticompensation (negative correlation between $\Delta H$ and $\Delta S$) develop from this model. Other investigators have derived a theoretical treatment (117) that supports the EEC effect, but only within a 20% range of the experimental temperature $T_c = - \Delta H / \Delta S$. Other notable investigators that have extensively dealt with the issue of enthalpy–entropy compensation, including Sharp (117), Cooper (124), and Jen-Jacobsen et al. (125).

### Allostery

In 1967, it was discovered that the binding of oxygen to one of the monomers of the hemoglobin tetramer led to a change in the affinity of oxygen on the other subunits of hemoglobin, despite the large distances that separate the oxygen binding sites (126). This effect, which is now known as allostery, has been since observed in a multitude of systems (127, 128). In general, a triggering event at one site of a macromolecule or macromolecular complex leads to an effect at a distal site. The triggering event could be caused by binding of a ligand, or a macromolecule. But it could also be caused by a chemical reaction. The resulting effect that occurs at a distal site could range from modulation of the binding affinity of a ligand (103, 129, 130), to alteration of catalysis (131-133), to causing significant conformational change that is usually essential for the function of the protein (134-136). An extensive list of structures of allostery in their initial and final states has been recently compiled (137).

Just as in ligand binding, allostery can be discussed in terms of the conformational selection and induced-fit mechanisms. The difference is that the triggering event would be causing an effect at a distal site. In the context of conformational selection, the ligand will select not only those conformers that facilitate its binding but also those that lead to an effect at the distal site. If the effect is an enhancement of catalysis, then the ligand would be selecting those conformers with residues at the distal site that are positioned such that the transition state free-energy barrier for reaction would be reduced. If, on the other hand, the effect is enhancement of binding, then the ligand would be selecting conformers that are both suitable for its binding as well as suitable for the binding of another ligand at a distal site.

Induced fit in allostery would entail a propagation of conformational changes from the triggering site to the distal site. This idea was initially proposed to explain cooperativity in hemoglobin by Perutz et al. (138, 139) who suggested a set of consecutive conformational changes that lead hemoglobin from its T-state (tense) to its R-state (relaxed). More recently, Pettigrew and co-workers (140, 141) have proposed the existence of a coupling pathway between two distal sites to explain how glyceral kinase from Escherichia coli but not Hemophilus influenzae is inhibited allosterically by phosphohexo-kinase system protein (142). Through specific mutations, they could interconvert these systems, which confirms the existence of allostery and suggests that allostery was an evolutionary effect. Another example where specific pathways are believed to connect distal sites is the RNA synthesis system. Cognate recognition and translation of aa-tRNA and the codon on mRNA signals RNA synthase to perform its function at a catalytic site that is located more than 70 Å away (143). A thorough recent study has invoked structural change as the main conduit between distal sites, evidence has been mounting that in addition to structural changes, molecular motion, both in the short (103) and longer (102, 103, 132, 143) timescales, plays an important role in allostery.

A prominent allostery system in which ligand binding at one site causes a change in the affinity of another ligand at another site is the catalytic activator protein or CAP (Fig. 4). In solution, the 47-kDa dimer exists as a homodimer and exhibits homotropic negative cooperativity, whereby the binding of the first cAMP molecule reduces the affinity of the second cAMP by nearly two orders of magnitude (144-146). The distance between the cAMP binding pockets on the two monomers is 24 Å and precludes any contribution from electrostatic interactions. A recent NMR study has suggested that the anti-cooperative effect could be caused by changes in the dynamics of the system, rather than by structural changes (103). The NMR data show that the binding of the first ligand causes significant increase in the slower motions but a slight tightening of the faster motions, which is evidenced by a slight increase of the order parameter $S_2$. Hydrogen/deuterium (H/D) exchange experiments show no overall change in the conformational change experienced by the system. The binding of the second ligand, however, causes significant tightening of the structure, both in the ps–ns and the μs–ms regime (103). The negative cooperativity is attributed to this loss of motion. In addition, significant stability of the structure is detected from lower H/D exchange rates.

We have conducted a thorough computational study of allostery in CAP to unravel the structural and dynamic changes that occur as a result of cAMP binding (132). Our extensive simulations for the apo, singly liganded, and doubly liganded CAP have confirmed results obtained from NMR (103), namely that cAMP binding to one monomer causes changes both in the shorter and longer timescale motions. The shorter timescale motions were measured by computing the order parameter $S_2$, whereas the longer timescale motions were obtained through a principal component analysis of the dynamics trajectory. Our simulations, however, included the DNA-binding domain that was absent in the NMR studies. We observed a remarkable conformational change that entailed the complete separation of the ligand binding and DNA-binding domains (Fig. 4). The unique aspect of this conformational change was that it only occurred in unoccupied monomers (Fig. 4). Hence it seems that ligand binding causes significant stabilization of the ligand-binding domain. When our calculations were repeated in the presence of DNA, however, we found that a switch in the cooperativity occurred, whereby the binding of the second cAMP was significantly favored by several kcal/mol (132). This finding explained
Receptor–Ligand Interactions in Biological Systems

Figure 4 A stereoview of CAP dimer in (a) doubly liganded, (b) singly liganded, and (c) unliganded forms. The protein is rendered in cartoon representation, colored yellow for the DNA binding domain, and green for the ligand-binding domain. The cAMP ligands are depicted as red spheres.

the seemingly contradictory fact that CAP has evolved to disfavor the binding of another cAMP, but that doubly liganded CAP was necessary for transcription activation (147).

The transglutaminase type 2 (TG2) enzyme is one example where a triggering event at a GTP-binding site causes alteration of catalysis at a distal site (133). TG2 is a multifunctional protein with roles in receptor signaling, extracellular matrix assembly, and apoptosis (Fig. 5) (133). The catalytic function of the enzyme, which consists of posttranslational amine modification of proteins or cross-linking of interchain glutamine and lysine residues (148), is allosterically controlled by calcium and GTP (149). But the three-dimensional structure of TG2 (Fig. 5) reveals that the binding site of GTP is located more than 20 Å away from the catalytic center, which precludes any direct electrostatic or bonding interaction. It has been shown that transglutaminases exist in an active and inactive conformation, and that GTP binding promotes the formation of the active conformer (150). Recently, a study from Begg et al. (133) has suggested that an arginine residue (Arg-579) is critical for TG2 transition to a compact and inactive form. They suggest that the GTP-mediated conformational transition into inactive form is caused by the formation of a hydrogen bond with the catalytic residues of the active site Cys-277 and a tyrosine residue. They argue that GTP masks the charge of a nearby Arg-579, which promotes the Cys-277 and Tyr-516 hydrogen bond.

In other systems, the triggering event causes significant conformational change. The elongation factor Tu (EF-Tu) is one example in which the hydrolysis of a phosphoester bond from GTP causes interdomain separation within the protein (Fig. 6). EF-Tu is a member of the GTPase superfamily that guides aa-tRNA to the acyltransfer site (A site) of the ribosome, which mediates the accommodation step and initiation of the protein biosynthetic process (see Fig. 6a). Cognate recognition between the anticodon of the nascent tRNA and the codon of the mRNA in the ribosome causes GTP hydrolysis more than 50 Å away in EF-Tu’s active site (Fig. 6b). The allosterically induced conformational change was rationalized as a series of perturbations that start from the GTP binding site that propagate into the nucleotide binding sites (domains II and III, see Fig. 6). It is not unconceivable that alteration in the shorter and longer timescale motions occur as a result of the hydrolysis of GTP. It is likely that the loss of a phosphate moiety will have the effect of increasing local fluctuations in the shorter timescale, which eventually permeates into the longer timescale motion, resulting in the separation of the domains, an entropically highly favorable process.

In many cases, dynamical events as a result of ligand binding involve multiple steps that combine conformational and dynamical changes. An example is a recent study that combined computational and experimental techniques to study the activation of gelatinase-2 (MMP-9), which is a zinc-based endopeptidase whose aberrant control has been implicated in various pathological processes, including cancer (151). The activation process consists of the removal of a ~100-residue peptide through proteolytic cleavage mediated by human kalikrein and involves the dissociation of a bond between a cysteine thiol and the catalytic zinc ion, which is a process known as the “cysteine switch.” A stopped-flow X-ray absorption analysis provided, for the first time, a detailed account of the coordination states of the zinc ion during the activation. The process was found to be triggered by binding of the kalikrein protease. It is interesting to note, however, that the binding site of kalikrein occurs at a distal site from the catalytic zinc ion. This finding suggests that alterations in dynamics and structure mediated by this interaction occur through an allosteric mechanism.

Conclusion

Despite the extensive amount of work that has been done toward understanding the molecular recognition process, the design of small molecules that exhibit potency to biological receptors remains a challenge. Most efforts to date have been driven by
Receptor–Ligand Interactions in Biological Systems

Figure 5  A stereoview of TG2 crystal structure with bound GDP. The TG2 structure is rendered in cartoon representation. The β-sandwich domain, catalytic core domain, first and second β-barrel domain is colored in green, red, blue, and yellow, respectively. The active and GDP molecule is rendered in cyan and orange spheres, respectively.

Figure 6  (a) Stereoview of the three-dimensional structure of EF-Tu/GTP complex. The protein is shown in ribbon representation, and domain I is colored in blue, whereas domain II, and III are shown in yellow. GTP is rendered in capped-sticks representation (P, O, C are shown in yellow, red, and white, respectively. Mg²⁺ is shown in orange spacefill representation. (b) Stereoviews of the three-dimensional structure of the EF-Tu/GDP complex, which illustrates the remarkable conformational change that the protein undergoes as a result of GTP hydrolysis.

structure, often without the consideration of motion that is experienced by the receptor. Structure-based design often entails the addition of various moieties to improve affinity. The result, however, is that the affinity does not improve because of an effect known as entropy–enthalpy compensation. Incorporating entropic effects is essential for a complete description of the recognition process. The motion that develops because of entropy often leads to effects such as cooperativity and allostery, which add another level of complexity to the ligand design process.

References


Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for

Vivian Hook, Shin-Rong Hwang, Jill Wegrzyn and Steven Bark, Skaggs School of Pharmacy, Pharmaceutical Sciences, University of California, San Diego, California
doi: 10.1002/9780470048672.wecb381

Peptide neurotransmitters and peptide hormones, collectively known as neuropeptides, are required for cell–cell communication in neurotransmission and for regulation of endocrine functions. Neuropeptides are synthesized from protein precursors (termed proneuropeptides or prohormones) that require proteolytic processing within secretory vesicles that store and secrete active neuropeptides. This article describes the application of chemical biological approaches advantageously used to define protease pathways involved in neuropeptide biosynthesis. Activity profiling of proteases, combined with mass spectrometry, has allowed identification of the novel cathepsin L cysteine protease pathway for neuropeptide biosynthesis, which contributes to neuropeptide production with the subtilisin-like prohormone convertase pathway. Furthermore, proteomic approaches for identifying proteases and protein systems present in secretory vesicles define the protease pathways and the functional protein systems that jointly operate in the secretory vesicle for production and secretion of active neuropeptides. Neuropeptidomic approaches allow defined primary structural analyses of neuropeptides. Future studies that gain understanding of protease mechanisms for generating active neuropeptides will be instrumental for translational research to develop therapeutic strategies for health and disease.

Neuropeptides for Cell–Cell Communication in Nervous and Endocrine Systems

Neuropeptides mediate neurotransmission as peptide neurotransmitters and mediate cell–cell communication as peptide hormones for endocrine regulation of target cellular systems. The term “neuropeptides” refers to this large, diverse class of peptide neurotransmitters and peptide hormones that typically consist of 3–40 residues. More than 100 different neuropeptides exist, and new neuropeptides are yet to be discovered.

The unique primary sequence of each neuropeptide defines its selective and potent biological actions. The same neuropeptides often serve important functions in both the nervous system as neurotransmitters and as peptide hormones in peripheral endocrine systems. For example, enkephalins function as neurotransmitters in the brain and are involved in peripheral actions, including regulation of intestinal motility and immune cell functions (1, 2). ACTH (adrenocorticotropic hormone) is present in the brain where it functions as a neuromodulator; furthermore, ACTH is a prominent peptide hormone released from the pituitary gland for control of glucocorticoid production in the adrenal cortex (3). Neuropeptides such as β-endorphin, NPY (neuropeptide Y), galanin, CRF (corticotropin releasing factor), vasopressin, insulin, and numerous others (Table 1) mediate diverse physiological functions that include analgesia, feeding behavior and blood pressure regulation, cognition, stress, water balance, and glucose metabolism, respectively (4–8).

Proteolytic processing for neuropeptide biosynthesis

Proneuropeptide (prohormone) precursors

Neuropeptides are derived from larger protein precursors known as proneuropeptides or prohormones. Proneuropeptides refer to protein precursors of neurotransmitters as well as peptide hormones, whereas prohormones refers primarily to
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for

Peptide neurotransmitters in the brain. Neuropeptides in the brain function as peptide neurotransmitters to mediate chemical communications among neurons. Neuropeptides are synthesized within secretory vesicles that are transported from the neuronal cell body via the axon to nerve terminals. The proneuropeptide (or prohormone) is packaged with the newly formed secretory vesicle in the cell body, and proteolytic processing of the precursor protein occurs during axonal transport and maturation of the secretory vesicle. Mature processed neuropeptides are contained within secretory vesicles at the synapse, where activity-dependent, regulated secretion of neuropeptides occurs to mediate neurotransmission via neuropeptide activation of peptidergic receptors.

Table 1 Neuropeptides in the Nervous and Endocrine Systems

<table>
<thead>
<tr>
<th>Neuropeptide</th>
<th>Regulatory Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTH</td>
<td>steroid production</td>
</tr>
<tr>
<td>α-MSH</td>
<td>skin pigmentation, appetite</td>
</tr>
<tr>
<td>beta-endorphin</td>
<td>analgesia, pain relief</td>
</tr>
<tr>
<td>calcitonin</td>
<td>calcium regulation</td>
</tr>
<tr>
<td>cholecystokinin</td>
<td>learning, memory, and appetite</td>
</tr>
<tr>
<td>CRF (corticotropin releasing factor)</td>
<td>ACTH secretion</td>
</tr>
<tr>
<td>enkephalin</td>
<td>analgesia, pain relief</td>
</tr>
<tr>
<td>galanin</td>
<td>cognition</td>
</tr>
<tr>
<td>glucagon</td>
<td>glucose metabolism</td>
</tr>
<tr>
<td>insulin</td>
<td>glucose metabolism</td>
</tr>
<tr>
<td>NPY</td>
<td>obesity, blood pressure</td>
</tr>
<tr>
<td>PACAP</td>
<td>water balance</td>
</tr>
<tr>
<td>somatostatin</td>
<td>neuronal differentiation</td>
</tr>
<tr>
<td>vasopressin</td>
<td>growth regulation</td>
</tr>
</tbody>
</table>

Peptide neurotransmitters and hormones are collectively termed neuropeptides. Neuropeptides typically consist of small peptides of approximately 3–40 residues. Several neuropeptides and several of their regulatory functions are listed; these neuropeptides and others function in multiple roles as physiological regulators (too numerous to list in this short table). Abbreviations are adrenocorticotropin hormone (ACTH), α-melanocyte stimulating hormone (α-MSH), neuropeptide Y (NPY), and pituitary adenylate cyclase-activating peptide (PACAP).

Although each proneuropeptide precursor possesses a distinct primary sequence, proteolytic processing occurs at dibasic endocrine peptide hormone precursors. To encompass peptide functions in both the nervous and endocrine systems, the terminology of “neuropeptide” and the respective “proneuropeptides” will be used in this article to refer to “neuroendocrine” functions of neuropeptides.

Proneuropeptide precursors share distinct and common features. Notably, the small active form of each neuropeptide is a segment present within its full-length precursor protein. A proneuropeptide may contain one copy of the active neuropeptide, as represented by proNPY, proagalanin, and provasopressin (Fig. 2). Alternatively, a precursor may contain multiple related copies of the active neuropeptide. For example, proenkephalin contains four copies of [Met]enkephalin, one copy of the related [Leu]enkephalin, and one copy each of the ME-Ag-Gly-Leu and ME-Ag-Phe (Fig. 2). Proteolysis of these precursors, especially tissue-specific proteolytic mechanisms, is required for biologically active neuropeptides to be generated.
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for... 

Figure 2. Proneuropeptides: structural features for proteolytic processing. Neuropeptides are synthesized as proneuropeptide precursors, also known as prohormones, that require proteolytic processing to liberate the active neuropeptide. Proteolytic processing occurs at dibasic and monobasic sites, as well as at multibasic sites. The precursor proteins may contain one copy of the active neuropeptide, such as the proneuropeptides for NPY, galanin, CRF, and vasopressin. Some proneuropeptides such as proenkephalin contain multiple copies of the active neuropeptide; proenkephalin contains four copies of (Met)enkephalin (ME), one copy of (Leu)enkephalin (LE), and the related opioid peptides ME-Arg-Phe (H) and ME-Arg-Gly-Leu (O).

Proteolytic processing of precursors for neuropeptide biosynthesis

Biosynthesis of neuropeptides begins with the translation of the respective mRNAs to generate the preproneuropeptide or preprohormone precursors. Proteolytic processing begins co-translationally at the rough endoplasmic reticulum (RER) where the NH₂-terminal signal peptide of the preproneuropeptide is cleaved by signal peptidase. The resulting proneuropeptide or prohormone is routed through the Golgi apparatus and is packed into newly formed secretory vesicles together with processing proteases. As the secretory vesicle matures, proteolytic processing occurs so that the mature secretory vesicle contains processed, biologically active neuropeptide that awaits cellular stimuli for regulated secretion.

Proteolytic processing at the dibasic or monobasic sites of proneuropeptides occurs primarily within regulated secretory vesicles (12-18). Cleavage at the COOH-terminal side of the paired basic residues results in peptide intermediates with basic residue extensions on their COOH-termini, which must then be removed by Lys/Arg carboxypeptidase to generate the mature neuropeptide (Fig. 3). Alternatively, cleavage of the precursor at the NH₂-terminal side of dibasic residue sites will generate peptide intermediates with basic residue extensions on their NH₂-termini, which then will be removed by aminopeptidase B to generate the active neuropeptide. Processing may also occur between the dibasic residues, which then will require both carboxypeptidase and aminopeptidase exopeptidase activities to generate the final neuropeptides.

Neuropeptides may also undergo post translational modification that modifies the biological activities of peptides. Activities of the neuropeptides may be altered by disulfide bond formation, glycosylation, COOH-terminal α-amidation, phosphorylation, sulfation, and acetylation (6, 7). This article, however, will focus on protease mechanisms for neuropeptide biosynthesis.

Biochemical properties expected of proteases for neuropeptide production

Elucidation of proteases in brain and neuroendocrine tissues is complicated because of the many different cell types and the presence of proteases in many subcellular compartments of these cells. To ensure that authentic proteases are identified for producing an active peptide, the neuropeptide field has used key criteria for successful elucidation of proteases that generate peptide neurotransmitters and hormones. These criteria...
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for Proneuropeptide Processing

Distinct cysteine protease and subtilisin-like protease pathways have been demonstrated for pro-neuropeptide processing. Recent studies have identified secretory vesicle cathepsin L as an important processing enzyme for the production of the endogenous enkephalin opioid peptide. Preference of cathepsin L to cleave at the NH2-terminal side of dibasic residue processing sites yields peptide intermediates with NH2-terminal residues, which are removed by Arg/Lys aminopeptidase. The well-established subtilisin-like protease pathway involves several prohormone convertases (PC). PC1/3 and PC2 have been characterized as neuroendocrine processing proteases. The PC enzymes preferentially cleave at the COOH-terminal side of dibasic processing sites, which results in peptide intermediates with basic residue extensions at their COOH-termini that are removed by carboxypeptidase E/H.

Application of these criteria has led to elucidation of the recently identified cysteine protease pathway and serine protease pathways, mediated by cathepsin L and proprotein convertases, respectively, for neuropeptide production (Fig. 3).

Chromaffin granules: model neurosecretory vesicles for proneuropeptide processing proteases

Elucidation of protease pathways for neuropeptide biosynthesis has been facilitated in the field with the use of isolated chromaffin secretory vesicles, a well-established model neurosecretory vesicle system used for investigation of proteases that synthesize neuropeptides and small-molecule neurotransmitters (19). Chromaffin secretory vesicles contain proneuropeptide precursors that undergo proteolytic processing to generate enkephalin, NPY, galanin, somatostatin, VIP, and other neuropeptides. These vesicles were used to identify the cathepsin L as a novel proneuropeptide processing enzyme using chemical biological approaches and have identified prohormone convertase processing enzymes (20–23).

Chemical biology for activity-based profiling and identification of proteases

Recent achievements in the development of active-site directed affinity probes for proteases and other enzyme classes provide direct chemical labeling of proteases of interest in the biological system (24–27). These specific activity probes allow joint evaluation of selective protease inhibition concomitant with labeling of relevant protease enzymes for more analyses. Moreover, activity-based probes that selectively label the main protease subclasses—cysteine, serine, metallo, aspartic, and threonine—can provide advantageous chemical approaches for functional protease identification. Activity probe labeling of proteases allows direct identification of enzyme proteins by tandem mass spectrometry. Such chemical probes directed to cysteine proteases have been instrumental for identification of the new cathepsin L cysteine protease pathway for neuropeptide biosynthesis, as summarized in this article.

Activity-based chemical profiling identifies the cathepsin L cysteine protease pathway in secretory vesicles that contributes to neuropeptide biosynthesis

A cysteine protease identified active site-directed affinity labeling of the 27 kDa enzyme of the "prohormone thiol protease" (PTP) complex (18, 22), which represents...
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for the major proenkephalin (PE) processing activity in chromaffin secretory vesicles (20, 21). The high molecular weight of the PTP complex of approximately 180–200 kDa (21) suggested the presence of several protein subunits because proteases typically possess lower molecular masses than that of native PTP. Studies then were targeted to identify the catalytic subunit of PTP responsible for PE-cleaving activity.

The activity probe DCG-04, combined with differential labeling in the presence of CA074, allowed identification of the 27 kDa protein as the active protease subunit of the PTP complex (18, 22). Two-dimensional gels resolved DCG-04-labeled proteins of 27–29 kDa (Fig. 4), which was identified as cathepsin L by mass spectrometry of tryptic peptides. Confirmation of the localization of cathepsin L in secretory vesicles was demonstrated by immunofluorescence confocal microscopy and immunoelectron microscopy, which illustrated the presence of cathepsin L in enkephalin and neuropeptide-containing secretory vesicles. The secretory vesicle function of cathepsin L contrasts with its well-known lysosomal function for degradation of proteins. These findings suggested a new biological function for cathepsin L in secretory vesicles for producing the enkephalin and related neuropeptides.

Gene analyses of Cathepsin L in neuropeptide biosynthesis by protease gene knockout and gene expression approaches

Cathepsin L knockout mice
Cathepsin L-deficient mice show decreased levels of enkephalin in the brain, with reduction by approximately one half (22). In addition, enkephalin brain levels are also reduced by about one half in PC2-deficient mice (28). These results support dual roles for both cathepsin L and PC2 in enkephalin production. Ongoing studies indicate multiple neuropeptides that are substantially decreased by more than 50% in the brain and endocrine tissues of cathepsin L knockout mice (Funkelstein et al., submitted for publication). With the observed alterations in brain neuropeptides, it will be of interest in future studies to assess the behavioral effects of the loss of neuropeptides in cathepsin L knockout mice. Cathepsin L knockout mice are viable and show phenotypes of hair loss and cardiac myopathy (29, 30). The mechanism for these functional effects of cathepsin L deficiency could possibly involve neuropeptides. New and continued investigations of neuropeptides in cathepsin L knockout mice will provide knowledge of the relative roles of cathepsin L in the production of particular neuropeptides.

Cellular gene expression of cathepsin L for enkephalin neuropeptide production in the regulated secretory pathway
Cellular routing and trafficking of cathepsin L to secretory vesicles for proneuropeptide processing was demonstrated by the coinjection of cathepsin L with proenkephalin in neuroendocrine PC12 cells (derived from rat adrenal medulla) (31). Expression of cathepsin L resulted in its trafficking to secretory vesicles with subsequent processing to enkephalin and related neuropeptides.

Figure 4 Activity-based profiling for identification of proenkephalin cleaving activity as cathepsin L. Activity-based profiling (APB) uses the strategy of labeling the active site of active proteases, often with an inhibitor-related probe, to identify proteolytic activity. Inhibition of proenkephalin cleaving activity by the cysteine protease inhibitor E64c in isolated chromaffin secretory vesicles (also known as chromaffin granules) allowed affinity labeling of the 27 kDa active protease enzyme proteins by a biotinylated form of E64 known as DCG-04 (Panel a). The inhibitor-labeled proteins were separated by two-dimensional gels (Panel b) and subjected to peptide sequencing by mass spectrometry, which revealed the identity of the proneuropeptide processing activity as cathepsin L.
vesicles that contain enkephalin and chromogranin A. Furthermore, cathepsin L expression resulted in cellular processing of proenkephalin into (Met)enkephalin that undergoes regulated secretion from PC12 cells. Cathepsin L generated high-molar-weight PE-derived intermediates (of about 23, 18–19, 8–9, and 4.5 kDa) that represented PE-derived products in vivo (20). Such results demonstrated a cellular role for cathepsin L in the production of (Met)enkephalin in secretory vesicles for its regulated secretion.

**Cathepsin L cleavage specificity indicates the subsequent aminopeptidase B exopeptidase step for neuropeptide production**

Studies of the cleavage specificity of cathepsin L demonstrated that it prefers to cleave on the NH₂-terminal side of dibasic residue processing sites of enkephalin-containing peptide substrates B.A.M-22P and Peptide F (22) and to cleave at the N-terminal sides of dibasic residues within peptide-MCA substrates (32). The cleavage specificity of cathepsin L results in enkephalin intermediate peptides with NH₂-terminal basic residue extensions, which are then removed by Arg/Lys aminopeptidase. Secretory vesicles from adrenal medullary chromaffin cells (33) and from pituitary (34) contain Arg/Lys aminopeptidase activity for neuropeptide production.

Recent molecular cloning studies have identified aminopeptidase B as an appropriate Arg/Lys aminopeptidase (35). Molecular cloning of the bovine aminopeptidase B (AP-B) cDNA defined its primary sequence that provided production of specific antisera to demonstrate localization of AP-B in secretory vesicles that contain cathepsin L with the neuropeptides enkephalin and NPY. AP-B was also found in several neuroendocrine tissues by western blots. Recombinant bovine AP-B (35) and rat AP-B were compared. Recombinant bovine AP-B showed preference for Arg-MCA substrate compared with Lys-MCA. AP-B was inhibited by arphamenine, an inhibitor of aminopeptidases. Bovine AP-B showed similar activities for Arg-(Met)enkephalin and Lys-(Met)enkephalin neuropeptide substrates to generate (Met)enkephalin, whereas rat AP-B preferred Arg-(Met)enkephalin. Furthermore, AP-B possesses an acidic pH optimum of 5.5–6.5 that is similar to the internal pH of secretory vesicles. The significant finding of the secretory vesicle localization of AP-B with neuropeptides and cathepsin L suggests a role for this exopeptidase in the biosynthesis of neuropeptides.

These findings indicate differences in the cleavage specificity of cathepsin L for the N-terminal side of dibasic residues within propeptides, compared with cleavage at the C-terminal side of dibasic residue processing sites by the prohormone convertases 1 and 2 (PC1/3 and PC2) (15–17, 23). These differences result in the requirement for different exopeptidases following endopeptidolytic processing by cathepsin L compared with PC1/3 or PC2. Although cathepsin L cleavage of neuropeptide precursors results in peptide intermediates extended with basic residues at their N-terminal that can be removed by aminopeptidase B, PC1/3 and PC2 cleavage at the C-terminal side of dibasic residues of propeptides results in peptide products containing C-terminal basic residue extensions that are removed by carboxypeptidase E. These dual protease pathways provide alternative routes for cellular processing of propeptides into active peptide neurotransmitters and hormones.

**Chemical biology defines the novel cathepsin L cysteine protease pathway combined with the prohormone convertase subtilisin-like pathway for neuropeptide production**

Significantly, the approach of activity profiling for cysteine propeptides has established cathepsin L as a new protease pathway for neuropeptide biosynthesis. Together with current knowledge in the field, these data demonstrate the existence of two distinct protease pathways for converting propeptides into active peptide neurotransmitters and hormones. These dual pathways consist of the newly discovered cysteine protease pathway for propeptide processing, which consists of cathepsin L followed by Arg/Lys aminopeptidase (aminopeptidase B), and the previously known proprotein convertase (PC) family of subtilisin-like proteases (15–17) that process propeptides with carboxypeptidase E (Fig. 3). Elucidation of these two protease pathways resulted from the application of the biochemical criteria required for processing propeptides.

**Cathepsin L: member of clan CA and the C1A papain subfamily**

Properties of cathepsin L may be compared among papain-like cysteine cathepsins (36–39) for understanding its role in producing neuropeptides. Cathepsin L belongs to the C1A subfamily of Clan CA (36). Clan CA was formed based on recognition of the first cysteine protease papain. The crystal structure of papain shows two structural domains separated by an active-site cleft. The N-terminal domain is comprised of α-helices and the C-terminal domain contains a β-barrel.

Clan CA is composed of twenty families. Family C1 within clan CA is divided into two subfamilies that consist of C1A (papain subfamily) and C1B (blysozyme hydrolase subfamily) groups. The larger subfamily C1A consists of secreted and lysosomal proteases that include the animal cysteine cathepsin proteases cathepsins B, H, and L as well as the plant proteases papain, chymopapain, and actinidain.

A division within the C1A subfamily exists between the papain-like proteases and the cathepsin B-like proteases. Included among the papain-like proteases are cathepsins O, H, L, X, and S. Cathepsin O is divergent, whereas the others are more closely related. Among the cathepsin B-like proteases are dipeptidyl-peptidase I and the endopeptidases from *Giardia*.

The distinction between exopeptidases and endopeptidases is merged for some members of the subfamily C1A. Dipeptidyl-peptidase I acts principally as an exopeptidase, removing N-terminal dipeptides, but may have some endopeptidase activity. Cathepsin B and H both possess endopeptidase activity but also possess exopeptidase activities. Cathepsin B acts as a dipeptidyl-peptidase, releasing C-terminal dipeptides. Cathepsin X is a carboxypeptidase.
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for

The proteases that enter the secretory or lysosomal pathways are synthesized as precursors, with N-terminal propeptide and signal peptides (36-39). Most members of family C1 have propeptides similar to that of papain. The propeptides act by blocking the active site. Papain-like propeptides are indicated by the ERFNNH motif. Propeptide inserts relative to papain occur within the catalytic domain in other family members. Cathespin B contains the "occluding loop" that carries the histidine residues important for peptidyl-dipeptidase activity. Most members of subfamily C1A are monomeric. In the case of cathespin L, it exists as a single-chain form (≈28 kDa on SDS-PAGE) and as a heavy-chain and light-chain form (≈24 and ≈4 kDa, respectively, on SDS-PAGE).

The specificity subsite that is dominant in most proteases of subfamily C1A is S2, which commonly displays a preference for occupation by a bulky hydrophobic side chain. Cathespin L possesses such S2 subsite preference for aromatic residues in the P2 position.

These properties contribute to the function of secretory vesicle cathespin L as a processing enzyme that produces neuropeptides.

Neuropeptides in mice deficient in PC2, PC1/3, and CPE: evidence for other processing enzymes

PC2 and PC1/3 prohormone convertases

A review of neuropeptide data from PC2 and PC1/3 knockout mice shows that the majority of neuropeptides studied are partially reduced in the knockout compared with wild-type controls. These results indicate roles for PC2 and PC1/3 in neuropeptide production. Importantly, the partial reduction of neuropeptides examined in PC2- and PC1/3-deficient mice also indicates possible roles for other proteases for processing proneuropeptides, such as cathepsin L, in secretory vesicles for neuropeptide production that has been discussed in this article. Thus, the chemical biology approach has identified the novel cathespin L cysteine protease pathway as a candidate route for neuropeptide production.

More specifically, to integrate data from PC2- and PC1/3-deficient mouse studies with recent data that demonstrate a candidate role for cathepsin L in neuropeptide production, results of neuropeptides studied in PC2- and PC1/3-deficient mice are summarized here. In PC2 knockout mice, many neuropeptides were partially reduced, with the exception of α-MSH that was nearly obliterated. PC2-deficient mice show increases in the POMC-derived peptide hormones ACTH and β-endorphin (1-31), which identifies them as substrates for PC2 (40). A mou

Carboxypeptidase E

After the actions of PC1/3 and PC2 for cleavage at the C-terminal side of paired basic residues (Fig. 3), the resulting peptide products contain basic residues at the C-termini that are then removed by carboxypeptidase E (CPE). Studies of CPE<sup>−/−</sup> mice that contain mutant, inactive CPE, showed that these animals show altered neuropeptide production (49-55). Studies of CPE peptide substrates in the fat/fast mice have been facilitated with specific isolation of peptides with C-terminal basic residues by the anhydrotrypsin affinity column for enrichment of CPE substrates (49-53). Analyses of such CPE substrates have demonstrated that numerous neuropeptides use CPE for their biosynthesis (49-55).

Analogously, it will be important to evaluate cathespin L-generated peptide products that serve as substrates for the subsequent anhydrotrypsin B (A-P-B) step. Significantly, the A-P-B substrates contain basic residues at their N-termini; these substrates do not contain basic residues at their C-terminal as CPE substrates do. Therefore, although the anhydrotrypsin affinity column can be used to isolate CPE substrates, selective analyses of A-P-B peptide substrates will require other approaches for isolation and analyses. It will be of interest in future studies to characterize A-P-B neuropeptide substrates generated by secretory vesicle cathepsin L.

Neuropeptidomics: LC-MS/MS tandem mass spectrometry analyses of neuropeptides

Primary structure analyses of neuropeptides are essential for understanding the structure-function features of neuropeptides. The development of mass spectrometry (MS) tools and approaches for identifying neurological peptides has become essential for defining neuropeptide structures that correspond to...
Neuropeptides, Chemical Activity Profiling and Proteomic Approaches for Neuropeptide Biosynthesis

Proteomics of Secretory Vesicles for Defining Proteases and Related Systems for Neuropeptide Biosynthesis

Direct proteomic approaches can be used to identify protease enzyme proteins, as well as protein categories in the biological system, that are present in secretory vesicles for neuropeptide production and secretion. Knowledge of the secretory vesicle proteome can advance our understanding of neuropeptide biosynthetic mechanisms that operate within this organelle. Recent examination of proteins in model chromaffin secretory vesicles revealed several functional protein categories that together support secretory vesicle production of neuropeptides and bioactive catecholamines for cell–cell communication (Fig. 5) (71). Protein systems involved in vesicular neuropeptide biosynthesis were examined in proteomic studies of soluble and membrane fractions of dense core secretory vesicles purified from neuroendocrine chromaffin cells. Proteins were separated by SDS-PAGE, and proteins from systematically sectioned gel lanes were identified by micropreparative LC-MS/MS (uLC-MS/MS) of tryptic peptides (71). Proteomic results revealed functional categories of prohormones, proteases, catecholamine neurotransmitter metabolism, protein folding, redox regulation, ATPases, calcium regulation, signaling components, exocytotic mechanisms, and related functions.

Several proteases of different mechanistic classes were identified by proteomics of secretory vesicles. These proteases included the subtilisin-like prohormone convertases 1 and 2 along with the metalloprotease carboxypeptidase E (CPE), which participates in prohormone processing (15–17). CPE has also been proposed to function as a prohormone sorting receptor (72). Regulators of PC1 and PC2 were found to consist of proSAAS and 7B2 (73, 74), respectively. Interestingly, cathepsin B (cysteine protease) (75) and cathepsin D (aspartic protease) (76) were identified, which indicates a novel location for these previously only known lysosomal proteases. The localization of cathepsin B in these secretory vesicles has been confirmed by immunoelectron microscopy (77). Cystatin C was identified in the membrane and soluble components, which may participate in the regulation of cathepsins L and S; cystatin C is a member of the cystatin superfamily of protease inhibitors (78, 79). Ubiquitin, a highly conserved 76 amino acid protein that is covalently linked to proteins targeted for degradation by the ubiquitin-proteasome system (80, 81), was identified; furthermore, ubiquitin-binding protein 52A was also identified. These findings may possibly be interpreted to suggest that ubiquitin-targeted protein degradation by proteosomes may occur in secretory vesicles. TIMP, tissue inhibitor of metalloproteinase, was also present (82). These identified protease system components were mostly present in both soluble and membrane fractions.

Furthermore, membrane-selective functions were implicated by proteomic data of these secretory vesicles. The membrane fraction exclusively contained an extensive number of GTP nucleotide-binding proteins related to Rab, Rho, and Ras signaling molecules (83, 84), together with SNARE-related proteins and annexins that are involved in trafficking and exocytosis of secretory vesicle components (85, 86). Membranes also preferentially contained ATPases that regulate proton translocation.
vesicles (also known as chromaffin granules) were isolated and subjected to proteomic analyses of proteins in the soluble and membrane components of the vesicles. Protein systems in secretory vesicle function consisted of those for 1) production of hormones, neurotransmitters, and neuromodulatory factors, 2) generating selected internal vesicular conditions for reducing condition, acidic pH conditions maintained by ATPases, and chaperones for the vesicles. Protein systems in secretory vesicle function consisted of those for 1) production of hormones, neurotransmitters, and neuromodulatory nucleotide-binding, calcium regulation, and vesicle exocytosis. These protein systems are coordinated to allow the secretory vesicle to synthesize and protein folding, and 3) vesicular trafficking mechanisms to allow the mobilization of secretory vesicles for exocytosis, which uses proteins for release neuropeptides for cell–cell communication in the control of neuroendocrine functions.

These results implicate membrane-specific functions for signaling and exocytosis that allow secretory vesicles to produce, store, and secrete active neuropeptides for the control of physiological functions. The protein systems used in these chromaffin vesicles, which represent dense core secretory vesicles (73), resemble those of brain synaptic vesicles (88) and secretory vesicles in the liver (89). Proteomic studies provide inference for secretory vesicle protein systems used for functions of these vesicles, including their biogenesis, that are required for production of enkephalin and related neuropeptides in brain and endocrine tissues. Secretory vesicles at synaptic nerve terminals in the brain are essential for chemical neurotransmission among neurons. Proteomic studies of synaptic proteins have revealed their regulation by brain injury (90), brain-derived neurotrophic factor (BDNF) (91), and drug regulation by morphine (92). The protein systems that support secretory vesicle exocytosis of peptide neurotransmitters and receptor activation at synaptic junctions of neurons function in concert to achieve neuropeptide-mediated communication in neural circuits.

Future Perspectives—Chemical Approaches for Elucidating Neuropeptide Mechanisms for Translation into Therapeutic Applications

It is extremely important to apply knowledge of protease mechanisms for neuropeptide biosynthesis to small-molecule strategies for the development of therapeutic agents that can modulate the production of specific peptide neurotransmitters or hormones. Current and future research using new approaches and tools, as discussed in this article, can provide insight into selective pharmacological approaches for exogenous therapeutic regulation of neuropeptide actions. Numerous health and disease conditions are regulated by neuropeptides. Proteases are essential for the conversion of inactive proprotein precursors into the active neuropeptides. Two main protease pathways have been elucidated for processing proenkephalins and hormones: the recently discovered cysteine protease cathepsin L with aminopeptidase B and the well-established subtilisin-like serine proteases that consist of prohormone convertases 1 and 2 followed by carboxypeptidase E/H. Endogenous regulators modulate these two protease pathways as endogenous peptide inhibitors, activators, and in vivo secretory vesicle proteins. Neuropeptides in CSF (cerebrospinal fluid) in neurological diseases can monitor brain nervous activity because neuropeptides represent active neurotransmission (93, 94).

Knowledge of specific regulators for particular neuropeptides can lead to future translational research for small-molecule regulation of prohormone convertases and cathepsin L pathways in the control of physiological functions. For example, regulation of opioid peptide production—enkephalin, β-endorphin, and dynorphin—may lead to new drugs for analgesia and pain relief. Specific small-molecule control of hypothalamic NPY in the control of feeding behavior may lead to improvement in obese conditions. Regulation of hypothalamic CRF and pituitary ACTH production is important for the control of steroid biosynthesis in the adrenal cortex for metabolic regulation. PC-related proteases have been implicated in steroid and lipid metabolism, tumor progression, atherosclerosis, and other physiological and disease conditions (5–8).

Application of chemical biology and proteomic approaches for understanding protease mechanisms in the biosynthesis of neuropeptides, Chemical Activity Profiling and Proteomic Approaches for NEUROPEPTIDES, SECRETION

Figure 5 Proteomics reveals functional secretory vesicle protein systems for neuropeptide biosynthesis, storage, and secretion. Chromaffin secretory vesicles (also known as chromaffin granules) were isolated and subjected to proteomic analysis of proteins in the soluble and membrane components of the vesicles. Protein systems in secretory vesicle function consisted of those for 1) production of hormones, neurotransmitters, and neuromodulatory factors, 2) generating selected internal vesicular conditions for reducing condition, acidic pH conditions maintained by ATPases, and chaperones for protein folding, and 3) vesicular trafficking mechanisms to allow the mobilization of secretory vesicles for exocytosis, which uses proteins for release neuropeptides for cell–cell communication in the control of neuroendocrine functions.
neuropeptides in health and disease is an exciting area of research for neuropeptide regulation of neuroendocrine systems.

Acknowledgments

Support from the National Institutes of Health for this research is appreciated.

References

45. Villeneuve P, Feliciangeli S, Croissandeau G, Seidah NG, Mbikay M.

Further Reading

See Also
Neurotransmitter: Production and Storage
Protease Pathways, Small Molecules to Elucidate Neuropeptidomics
Systems Biology
Neurotransmission, Measuring Chemical Events in

Christopher B. Jacobs and Trisha L. Vickrey  Department of Chemistry, University of Virginia, Charlottesville, Virginia
B. Jill Venton Department of Chemistry, Neuroscience Graduate Program, University of Virginia, Charlottesville, Virginia
doi: 10.1002/9780470048672.wecb383

Measuring chemical events in neurotransmission is challenging because fast measurements of small amounts of neurotransmitters and neuromodulators are needed to track the dynamics of chemical changes accurately. In this article, we outline the basics of three popular methods for measuring neurochemical changes: electrophysiology, microdialysis, and electrochemistry. Electrophysiological techniques measure changes in membrane potentials associated with neurotransmission. These methods are often used to measure receptor-gated ion channel currents and are popular for neuropharmacology studies. Microdialysis is a sampling technique that can be used to monitor basal levels of neurotransmitters directly. When coupled to a separation technique, microdialysis is advantageous because it can be used to detect virtually any compound in the brain. Electrochemical techniques are popular because microelectrodes allow rapid, direct detection of neurotransmitters with minimal tissue disturbance. Although the analytes must be electroactive, electrochemistry has been used successfully to monitor neurochemical changes in various preparations, from single cells to behaving animals. Future research in monitoring neurochemical events will include improving the temporal resolution, spatial resolution, and selectivity of measurements.

Neurotransmission is the transfer of an informational signal, a chemical messenger, between two neurons. The traditional picture of synaptic transmission is shown in Fig. 1. A terminal from one neuron forms a synapse with a dendrite of another neuron. After the neurotransmitter is synthesized in the perikaryon of the terminal, it is packaged into specialized synaptic vesicles. The vesicles range in size from about 50 nm to 100 nm and can store between 3000 and 30,000 neurotransmitter molecules (1). An action potential propagating to the terminal changes the cell membrane potential and can induce the fusion of vesicles to the synaptic membrane. This fusion results in exocytosis, the coordinated release of transmitter molecules into the synapse (Fig. 1). Transmitters can then bind to postsynaptic receptors and activate signaling pathways through GTP-binding proteins or open gated ion channels, which leads to localized changes in membrane potential (2). Membrane potentials, ionic currents, and action potentials can be measured using electrophysiology techniques.

Released neurotransmitters can have fates other than binding postsynaptic receptors. Transmitters can bind to presynaptic receptors that act as a feedback loop to regulate additional release of the chemical messengers. Transporters take up neurotransmitters back into the neuron and clear them from the extracellular space. Neurotransmission can also be ended by enzymatic degradation of transmitters, although this process is kinetically slower than uptake. Although the traditional picture of neurotransmission is of short-range signaling, neurotransmitters can also diffuse out of the synapse and act at distal targets, which allows longer distance signaling (3). This process is called volume transmission, and extracellular detection methods such as microdialysis and microelectrodes detect these extrasynaptic concentrations.
Neurotransmission, Measuring Chemical Events in Vesicles

Empty Vesicles

Presynaptic Receptors

Neurotransmitters

Transporter

Postsynaptic Receptors

Figure 1 General concept representation of synaptic transmission. Vesicles that contain neurotransmitter molecules dock to the cell membranes and release their contents into the synapse by exocytosis. Neurotransmitters can diffuse across the synapse and bind to postsynaptic receptors or can diffuse out of the synapse. This extrasynaptic neurotransmitter can bind to presynaptic receptors, diffuse away and activate receptors on distal neurons, or be cleared from the extracellular space by transporters.

Electrophysiology

Electrophysiological techniques measure changes in potentials associated with neurotransmission. The potential of the neuronal membrane is controlled by ionic concentrations, which are regulated by the active transport of ions. As postsynaptic dendrites receive neurotransmitter signals, receptor-gated Na\(^{+}\) and Ca\(^{2+}\) channels are activated, which allows influx down the ionic gradient. Membrane depolarization occurs as cations enter the cell. Once the depolarization reaches a certain threshold, successive Na\(^{+}\) and Ca\(^{2+}\) ion channels open along the axon in a domino fashion to propagate the electrical signal. This propagation of voltage is called an action potential, and it will not occur unless a threshold depolarization is reached. After maximal depolarization, Na\(^{+}\) and Ca\(^{2+}\) ion channels begin to close, and simultaneously, K\(^{+}\) ion channels open and K\(^{+}\) effluxes out of the nerve cell. This process reverses the polarization, and because of the length of time the K\(^{+}\) channels stay open, the abundance of cations outside the cell causes a hyperpolarization of the membrane potential.

Transmembrane movement of ions is controlled by gated ion channels and can be represented by a simple RC circuit, in which the ion channels act as resistors and the cell membrane acts as a capacitor. Hodgkin and Huxley, who won the 1963 Nobel Prize for Physiology or Medicine for their work, quantified these transmembrane voltage changes in the axon of a giant squid and determined the time-dependent behavior of Na\(^{+}\) and K\(^{+}\) channels. Current electrophysiological studies measure neurogenic changes through various intracellular and extracellular techniques. Although these techniques are diverse, common intracellular techniques include the two-electrode voltage clamp, patch clamp, and whole cell methods that measure ion channel currents and common extracellular techniques that measure voltage changes outside cells.

Two-electrode voltage clamp

Voltage clamp methods typically involve two micropipette electrodes that impale the cell; one electrode measures voltage changes, whereas the other applies a stimulating voltage. Transmembrane voltage is clamped at a constant voltage eliminating capacitive current, and allowing conductance to be measured, which is proportional to voltage. The clamped voltage then is varied in a stepwise fashion with each step resulting in a change in ion current, which is measured. For this reason, voltage clamp techniques are useful in studying voltage-gated ion channels. In a typical experiment, voltage steps are applied to an oocyte or insect muscle cell, and the effects of the addition of pharmacological agents on ion channel currents are tested. For example, Fig. 2a shows voltage clamp recordings of inward calcium channel currents in parasitic nematode muscle cells after a series of depolarizing steps (4). The current versus voltage graph shows the maximal current was induced by a step to 0V. The study showed that bathing a voltage-clamped cell in the nematode neuropeptide, AF2, increased the inward calcium currents, which suggests that the neuromuscular function of the parasitic nematode was modulated by AF2. Such studies can be useful for identifying possible target sites for drugs that act at voltage-gated channels.

Patch clamp

Some of the most popular electrophysiology methods currently used are patch clamp techniques, in which individual ion channels can be studied. Unlike the two-electrode methods described above, these methods use a single electrode to measure voltage changes. A glass pipette electrode with a flat, open tip is placed onto a cell, and piece of membrane that contains an ion channel is “sucked” into the pipette tip. This action allows the direct current of ions moving through channels to be measured. Although the pipette tip has a high resistance, negative pressure...
Neurotransmission, Measuring Chemical Events in

Figure 2  a) Two-electrode voltage clamp experiment. The cell potential was held at −35 mV, depolarizing steps from −35 to +20 mV were made (top left), and resultant inward currents from Ca\(^{2+}\) channels were measured for each step (bottom left). The right graph shows an I-V curve for nematode muscle Ca\(^{2+}\) ion channels. (Data reprinted from Ref. (4) by permission of Macmillan Publishers Ltd.) b) Whole cell experiment. Representative trace of an individual EPSC before (left) and after (right) application of an adenosine A2A receptor agonist CGS 21680. The agonist causes the ESPC amplitude to decrease. (Data reprinted from Ref. (5) by permission of Elsevier.) c) Inside-out experiment. A continuous single-channel current is shown from in an inside-out membrane patch of a rat cultured hippocampal neuron before and after addition of diazepam. The top trace shows currents induced by 5 µM GABA. The bottom four traces show currents induced after a 1 µM solution of diazepam is introduced. A gradual increase in single-channel current amplitude occurs. (Data reprinted from Ref. (8) by permission of Macmillan Publishers Ltd.)

is applied to form a seal with a gigahm resistance, which is commonly referred to as a gigaseal. The pipette tip is usually filled with a solution that approximates intracellular fluid. Two types of experiments can be performed once the pipette electrode is in place: voltage clamp or current clamp. Similar to two-electrode voltage clamp methods described above, in voltage clamp mode the membrane potential is held at a constant voltage, and current through the ion channels is measured. In current clamp mode, the current is kept constant, and voltage changes are measured. To identify which types of channels contribute to the signal, electrophysiology techniques are commonly used in conjunction with pharmacology experiments, in which drugs are applied to block specific channels. Several configurations of this method correspond to whole cell, outside-out and inside-out models (Fig. 3).

Whole cell model

In the whole cell model, a gigaseal is formed as the pipette is attached to the cell, and then a more dynamic suction is applied, which causes the interior of the cell to be sucked into the pipette tip (Fig. 3a). This action allows current and conductance of the entire cell to be measured. Therefore, the whole cell model measures changes caused by many ion channels on the entire cell membrane. Additionally, the liquid content of the cell will mix and equilibrate with the solution in the pipette, which allows pharmacological agents to be administered into the cell. Of the patch clamp techniques, the whole cell method is the most common and can be used to determine how pharmacological agents affect the total conductance of neurons.

An example of a whole cell study is shown in Fig. 2b. Excitatory postsynaptic currents (EPSCs), which are caused
Neurotransmission, Measuring Chemical Events in Whole Cell Method

- Outside-out Method
  - In the outside-out model, the pipette is attached to the entire cell as in the whole cell model, followed by a sharp pull that causes the cell membrane to break and reseal with the pipette tip (Fig. 3b). With the extracellular region exposed, channel activity as a response to different external stimuli can be probed. This configuration is less common than the inside-out method. Using an outside-out method, single-channel opening activity has been recorded while various neurotransmitters were released. For example, this patch clamp method was used as a detector for capillary electrophoresis separations of GABA, glutamate, and NMDA (7).

- Inside-out Model
  - In the inside-out model, gigaseal formation is followed by a sharp pull, which detaches the cell membrane and exposes the inner membrane of the cell to the bathing solution. The portion of the membrane that was on the inside of the cell is then on the outside of the pipette, whereas the outer portion is in equilibrium with the fluid inside the pipette (Fig. 3c). This configuration is useful for studying the effects of intracellular molecules or drugs on individual ion channels and has been used in pharmacology experiments to study ligand-gated ion channel activity. For example, benzodiazepines, which are used to treat anxiety, are thought to augment synaptic inhibition in the central nervous system. Inside-out patch clamp was used to study the effect of the benzodiazepine, diazepam, on single-channel conductance (8). Diazepam caused a seven-fold increase in the conductance of GABA\(_A\) chloride channels from rat cultured hippocampal neurons (Fig. 2c). This study demonstrates that the inside-out method can provide information about the effect of drugs on the current of a particular ion channel. Inside-out patch clamp has also been used to study pain sensory transduction in rat ganglion neurons through capsaicin-activated ion channels (9). Capsaicin is known to activate certain cation channels and cause severe pain. One study used patch clamp to determine that 12-(S)-hydroxyeicosatetraenoic acid (HPETE), a lipoxygenase product, was an endogenous activator of capsaicin channels. The inside-out technique is widely used to study various channels, both inhibitory and excitatory, and various preparations from retinal neurons to central neurons. For example, inside-out patch clamp has been used to study the photosensitivity of rat intrinsically photosensitive retinal ganglion cells (ipRGCs), which are photoreceptors that control pupil response (10). In this study, inside-out recordings showed that ipRGCs are photosensitive. This study gives insight into the complex cascade of events that lead to vision.

Extracellular Methods

With intracellular methods, the change in voltage caused by the movement of positive ions into the postsynaptic cell through glutamate-mediated channels, were monitored. This study examined whether adenosine receptor activation could modulate EPSCs. Whole cell patch clamp was performed on neurons in striatal rat brain slices. Figure 2b shows an EPSC evoked by electrical stimulation before and after application of an A\(_2\)A adenosine receptor agonist. The decreased amplitude of the EPSC after the adenosine receptor agonist is applied suggests that adenosine receptor activation inhibits glutamate channel-mediated EPSCs (5). Many researchers also study inhibitory postsynaptic currents (IPSCs) mediated by gamma-aminobutyric acid (GABA\(_A\))-gated chloride channels. For example, whole cell patch clamp was used to differentiate two different types of IPSCs in substantia gelatinosa neurons of the mouse spinal cord (6). IPSCs with a fast decay had a different pharmacological profile than those with a slow decay.

Whole cell methods continue to be popular for neuropharmacology studies because they allow researchers to understand the complex actions and regulation of receptors gating ion channels.
an electric field in the extracellular space, and these changes in electrical activity can be measured by extracellular recordings.

Both invasive and noninvasive techniques are routinely used. Extracellular techniques can be used in many preparations, including awake, behaving animals.

In a typical invasive setup, a wire electrode or an array of wires is implanted into the brain, and changes in potential are measured. Electrodes that are directly next to a neuron will record voltage changes when that neuron fires. This setup has been used to measure neuronal firing patterns in animals performing complex behaviors, such as drug self-administration (11). These studies have shown that dopamine neurons fire as predictive signals of reward and will fire in response to cues that have been previously associated with rewards (12).

Multielectrode arrays have also been used to map the response of different neurons to external stimulations. For example, Nicolelis et al. have used multichannel electrodes to map tactile responses of an anesthetized rat (13). Figure 4 demonstrates the neuronal response to stimulation of the infraorbital nerve, which carries information from the whisker’s mechanoreceptors to the brain. Typical data is shown, in which firing patterns before and after the stimulus (dashed line, time 0) are given for multiple trials (top of figure) and then binned to get average firing rates (histograms). The results show that the hippocampal neurons have a longer latency to fire but are activated for longer than the neurons in the somatosensory cortex and thalamic ventral posteromedial nucleus. Such studies help identify and interpret the neuronal pathways that process tactile stimulation.

In a noninvasive setup, surface electrodes are placed on the surface of the skin and can be used to study individual or multiple neurons. Additionally, many studies look at motor units, which are a motor neuron and all muscle fibers that the neuron innervates. For example, stable recordings of single motor unit potentials can be made using electromyography (14). Surface electrodes are of interest in electroencephalography (EEG) measurements, in which electrodes are placed on the scalp. Noninvasive EEG measurements are particularly useful in human studies. For example, EEG recordings have been used to study adenosinergic neurotransmission in homeostatic sleep-wake regulation in humans (15). These studies found that adenosinergic neurotransmission plays a role in non-rapid eye-movement sleep homeostasis.

Summary and future directions

Intracellular patch clamp methods provide a sensitive, high-resolution approach to studying ion channels. Patch clamp studies are particularly useful in pharmacological studies, such as identifying receptor agonists and antagonists. Studies can be performed in brain slices or in vivo. However, several pervasive problems with patch clamping exist. Patch clamping is inherently an invasive technique, and morphological changes in the cell can occur during the clamping process that interfere with the overall function of the ion channel investigated (16).

Although patch clamp methods are robust and have provided information about the function of ion channels, they are labor-intensive, require a high degree of technical skill, and are low-throughput. Therefore, a research topic of current interest is developing automated, high-throughput patch clamp methods. For example, a planar patch clamp electrode, which uses a planar substrate rather than a pipette to determine current for whole cells, has been developed. Planar patch clamp requires less skill and allows higher throughput measurements to be made (17). Interest in developing methods to screen ion channels as drug targets has led to the commercial availability of automated planar patch clamp systems (18). Additional development of these high-throughput methods is expected to lead to widespread adoption of these techniques.

With extracellular measurements, action potential firing adjacent to an electrode gives information about neuron output and synaptic input. These methods are good for measuring neuronal activity of awake, behaving animals. The use of implantable electrodes provides high resolution of single neurons, whereas nonimplantable electrodes sample a larger surface area, which causes them to have a lower resolution. However, noninvasive measurements are the most useful for higher order animals and humans. Signals from extracellular recordings are much smaller and harder to detect than with intracellular methods. For example, the number or type of neurons that are firing at a given time can be difficult to determine (19). Additionally, action potentials are often obscured or unsynchronized, and extracellular recordings can be a summation of multiple neuronal signals. However, with the development of better computer programs and bioinformatics capabilities for data analysis, these problems can be overcome (20). Additional integration of both intracellular and extracellular methods will provide a more holistic picture of neurotransmission.

Figure 4 Electrically evoked single-unit tactile responses recorded with microdrive multielectrode arrays in rat hippocampus CA1, primary somatosensory cortex, and ventral posteromedial nucleus brain regions. (Upper) raster plot of single-unit spikes before and after electrical current stimulation to whiskers. Each row is a separate trial. (Lower) Summed activity for all trials in 1-ms bins that demonstrate a response to electrical stimulation. The graphs show different latencies in firing for the three different regions. [Data reprinted from Ref. (13). Copyright (2002) National Academy of Sciences, U.S.]

In a noninvasive setup, surface electrodes are placed on the scalp. Noninvasive EEG measurements are particularly useful for higher order animals and humans. Signals from extracellular recordings can be a summation of multiple neuronal signals. However, with the development of better computer programs and bioinformatics capabilities for data analysis, these problems can be overcome (20). Additional integration of both intracellular and extracellular methods will provide a more holistic picture of neurotransmission.
Microdialysis

Microdialysis has served as a powerful tool in the direct measurement of cerebral neurotransmitters. It can be used as a qualitative technique to monitor changes in neurotransmitters or as a quantitative technique to determine actual neurotransmitter concentrations such as basal levels. Measurements are made using a probe sheathed with a semipermeable membrane, which allows extracellular molecules of a low molecular weight, such as neurotransmitters, to pass but excludes larger molecules such as proteins. The probe is perfused with artificial cerebral spinal fluid, and molecules diffuse across the membrane according to their concentration gradients. Fluid fractions are collected at the probe outlet. Fractions are typically analyzed using separation techniques, such as high-performance liquid chromatography (HPLC) or capillary electrophoresis (CE). Microdialysis can provide both highly selective and sensitive analysis of neurochemical changes, but the temporal resolution has been limited traditionally to 5–20 minutes (21, 22). Continuous sampling of chemical events allows processes such as drug metabolism and neurotransmitter release to be studied. Microdialysis also is useful for studying neurotransmission in various preparations, including in behaving animals.

Microdialysis Theory

In microdialysis, a solution isotonic to that being sampled is perfused through an inlet tube of the probe (Fig. 5). A concentration gradient is established on each side of the membrane between the perfusate and the sampling region. This gradient causes neurochemicals from the sampling region to diffuse through the membrane, and the compounds are collected for analysis at the outlet of the probe. Analyte removal from the sampling region keeps the concentration gradient intact and allows continual sampling of the extracellular space. This process is highly dependent on flow rate of perfusion.

At typical flow rates, the concentration in the dialysate, Cext, is less than the actual concentration in the extracellular fluid, Cin (23). The ratio of Cin/Cext is defined as relative recovery, R, and must be considered for probe calibration and sampling optimization. In vitro, R is easily calculated because the dialysate and the extracellular fluid are homogeneous; therefore, probe calibration is easily obtained. However, in in vivo studies, calculation of R is difficult because of the active removal of neurotransmitters by uptake and transporters. Movement of analytes is impeded by tissue that surrounds the probe, and this movement cannot be easily accounted for with in vitro calibrations. Therefore, the most common method to determine concentrations in vivo is the zero-net flux method, in which known analytic concentrations are added to the perfusate (Cin), and then the analytic concentration is measured at the probe outlet (Cout). The difference between analytic concentration at the inlet and outlet is used to establish the actual analytic concentration in the tissue, and the relative recovery rate can be calculated. This calibration method can be used to estimate basal levels of neurotransmitters. For example, the zero-net flux method has been used to determine that basal concentrations of dopamine are approximately 1–3.5 nM (24, 25). Although basal level concentrations of neurotransmitters can be determined with this method, the contribution of tissue damage to the microdialysis signal cannot be neglected. Because microdialysis is an invasive technique, a traumatized layer of tissue will be next to the probe, which influences the rate of neurotransmitter release and uptake. Bungay et al. have suggested in the case of dopamine that trauma leads to underestimation of extracellular dopamine concentration and have proposed a quantitative model to correct for that (23).

Microdialysis coupled to high-performance liquid chromatography

Detection of neurotransmitters in microdialysate samples has been achieved traditionally by coupling microdialysis sampling to highly sensitive analysis methods such as high-performance liquid chromatography (HPLC). HPLC is a type of column chromatography, which separates the collected mixture of analytes and resolves them into individual components. Therefore, multiple chemical species can be identified in a given sample. Monitoring changes in concentrations over time gives insight into a wide range of processes such as transporter function, effects of pharmacological agents, and behaviorally evoked neurotransmitter changes.

Glutamate is the most abundant excitatory neurotransmitter in the nervous system. Under normal conditions, glutamate transporters remove extracellular glutamate into glial cells. However, during disease or ischemia, a lack of oxygen delivery to cells, excess glutamate can accumulate and result in excitotoxicity.
Microdialysis studies have played an important role in understanding the relationship between glutamate levels and ischemic events. Glutamate has been shown to increase during a stroke, which contributes directly to neuron damage and additional vulnerability to ischemia (26, 27). Subsequently, methods to reduce glutamate release during an ischemic event have been of interest. Recent microdialysis studies indicate that spinal cord necrosis can be reduced by administering magnesium sulfate to the area experiencing ischemia (28).

Figure 6  Effects of intrathecal magnesium sulfate on ischemia-induced release of glutamate in the dialysate collected from the spinal cord of rabbits. Data are expressed as percentage of the mean baseline levels in each group. Aortic artery occlusion to cause ischemia results in an increase in glutamate in the placebo group (open circles), but administration of MgSO4 before ischemia is induced prevents the increase (triangles). A sham group in which ischemia is not induced is also shown (squares). (Data reprinted from Ref. (28) by permission of Lippincott, Williams, and Wilkins.)

Microdialysis coupled to capillary electrophoresis

Recently, several groups have used capillary electrophoresis (CE) instead of HPLC as an analysis method for microdialysis samples. The chief advantage of CE is that high-speed separations of very small sample volumes can be achieved. The improved temporal resolution allows samples to be analyzed every 10 seconds, which facilitates better understanding of the dynamics of neurotransmitter changes during behavior. This method has enabled the online analysis of microdialysis samples using capillary electrophoresis and eliminated the need to store fractions. The disadvantage is that samples must be derivatized to make them fluorescent so that highly sensitive laser-induced fluorescence can be used for detection.

Microdialysis coupled to CE was used to detect changes in amino acids, such as glutamate, during exposure of rats to a predator fox odor (31). Figure 7 shows that large changes in glutamate were observed after presenting the fox odor, particularly in a subset of rats that were high responders. These changes lasted only a few minutes. The data are compared with a similar experiment that measured glutamate changes in response to the presentation of the fox odor with typical microdialysis–HPLC with only 10-minute temporal resolution (32). The magnitude and duration of the glutamate change are more accurately tracked with the higher temporal resolution CE method. Microdialysis–CE has been used to study amino acid neurotransmitters during learning (33) and catecholamines such as dopamine and norepinephrine during the sleep–wake cycle (34). Future research in this field includes miniaturizing the CE detection on a chip (35) to make the technique more amenable to wide-scale implementation.
Microdialysis may also see more use in the future as a diagnostic tool to improve temporal resolution. Microdialysis coupled to CE provided information not previously found with HPLC because of increased time resolution. (Data adapted from Ref. (31) by permission from Blackwell Publishing.)

Summary and future directions

Although microdialysis is sensitive and selective, it is an invasive technique that causes trauma and tissue damage, such as edema and alteration of the blood-brain barrier and glucose metabolism in areas near probe insertion. Consequently, interpretation of results garnered from microdialysis has been controversial. Several studies indicate cerebral blood flow and glucose metabolism recovery within 24 hours (27), and many studies delay measurements after probe implantation to allow tissue recovery. However, recent studies that compare dopamine levels at the insertion point and 1 mm adjacent to this area indicate permanent changes in tissue because of trauma (36). Furthermore, neuronal loss and tissue disruption have been observed through light and electron microscopic analysis (37).

Nevertheless, microdialysis measurements agree well with other methods used to measure neurochemical changes such as microelectrode techniques (38). For example, pulse voltammetric methods have determined basal concentrations of dopamine to be approximately 1.5 ± 0.5 nM (39), compared with 1–3.5 nM estimated with microdialysis measurements (24, 25). Undoubtedly, efforts to minimize tissue damage should be made to ensure accurate results. However, microdialysis has seen widespread use because the data on chemical changes has consistently provided insight into brain functions and pathologies. The efforts to improve temporal resolution for microdialysis will allow it to be used to track chemical changes specific to certain behaviors. Microdialysis may also see more use in the future as a diagnostic tool in human neurosurgery for monitoring neurotrauma (40).

Electrochemical Techniques

Electrochemical detection is one of the most common methods for neurotransmitter monitoring. Many neurotransmitters are electroactive, including dopamine, norepinephrine, epinephrine, and serotonin, and can be detected directly using microelectrodes. The small size of typical microelectrodes, from 7 to 30 μm in diameter, results in less tissue damage than the implantation of larger probes, such as microdialysis probes or conventional electrodes. Three of the most common methods of voltammetric detection for neurotransmitters are amperometry, high-speed chronoamperometry, and fast-scan cyclic voltammetry. In each of these methods, a voltage waveform is applied to the electrode. When the potential is sufficient, electroactive neurochemicals can be oxidized or reduced, and a current is measured that is proportional to the concentration of analyte at the electrode surface. The range of voltages that can be used is finite because of the possible oxidation or reduction of physiologic solutions and oxygen, which limit detection to the range of about −1 V to 1.5 V. The shape of the potential waveform applied in each technique is unique, and the resulting chemical information is different for each method (41).

Constant potential amperometry

Amperometry applies a constant potential to the microelectrode that will oxidize the analyte at the electrode surface (Fig. 8a). The current is limited solely by the mass transport rate to the electrode. Measurements can be made with extremely high temporal resolution, typically 500 Hz, because time resolution is not limited by the electrochemistry at the electrode. However, little analyte selectivity occurs with amperometry because a change in the concentration of any moleule electroactive at the applied potential will alter the measured signal.

One major use of amperometry is to measure the efflux of neurotransmitters from individual cells (42). Because the cell type and the molecule being released is known, selectivity is not an issue for these experiments. This technique, developed by R. Mark Wightman, is now widely adopted and involves placing a microelectrode directly above a cell (42). A secretagogue known to stimulate release is applied, and vesicular release is monitored with high temporal resolution. Recently, amperometry has been used to show that ionic secretagogues can influence the exocytotic release of specific neurotransmitters. An example experiment is shown in Fig. 8d, in which dopamine release is monitored after application of K+ to isolated dopaminergic retinal neurons (43). Typical data consists of multiple peaks, each of which is representative of neurotransmitter release from a single vesicle (Fig. 8d). The area under the peak is proportional to the total number of molecules in that vesicle and can be calculated. The amount of dopamine released from retinal cells was calculated to be about 32,000 molecules per vesicle (43). Recent studies have shown that small features detected with amperometry are caused by incomplete vesicular release associated with vesicle fusion (1, 44). Amperometry is a good method for studying basic mechanisms of neurotransmission or the effects of drugs on neurotransmitter release from single cells.

Chronoamperometry

For high-speed chronoamperometry, a recurrent square waveform is applied to the electrode, with potentials sufficient...
enough to oxidize the analyte of interest and subsequently reduce the analyte to the initial form. Rapidly changing the potential causes an inherent charging current at the electrode surface, but this current is dissipated quickly because of the micron size of the electrode. Therefore, the current measured at the end of the step, after the charging current has decayed, is proportional to the analyte concentration (Fig. 8b). The ratio of the peak reduction current to the peak oxidation current is a measure of the reversibility of the electrochemical reaction and can be used to differentiate between some molecules. However, this selectivity is limited, and positively identifying the analyte solely by this technique is difficult. The chronocoulometry waveform is typically repeated at 5-25 Hz.

Greg Gerhardt pioneered a chronocoulometry method that has been used to observe diffusion and uptake parameters in the rat brain (45). In this technique an exogenous neurotransmitter, such as dopamine, is introduced into the brain by pressure ejection. Gerhardt’s group could compare dopamine uptake rates of wild-type rats and rats that had been treated with nomifensine, a dopamine uptake inhibitor (46, 47). They found that nomifensine slowed dopamine uptake rates, as seen by the slower return to baseline after injection. (Data from Ref. (47) used by permission from Elsevier.) Cyclic voltammetry data of dopamine concentrations from uptake inhibitor nomifensine was administered. Uptake rate is reduced by nomifensine as seen by the increased concentration and the slower return to baseline after injection. (Data taken from Ref. (43) by permission from the American Chemical Society.)

Figure 8 Electrochemical techniques. Applied potential versus time and resulting current versus time graphs for a) amperometry, b) high-speed chronocoulometry, and c) fast-scan cyclic voltammetry. The dotted circle in b) shows where currents are measured after the charging current has decayed. Panels d-f show real data from experiments using each of the techniques. Amperometry data (d) show a current increase that corresponds to detection of electroactive species. Each peak is representative of dopamine release from a single vesicle in retinal neurons, and the area under each peak was used to quantitate the number of molecules released. (Data taken from Ref. (43) by permission from the American Chemical Society.) d) Dopamine concentrations measured by chronocoulometry. Upward-facing arrows indicate time of dopamine injection; downward-facing arrows indicate time of nomifensine injection. Curve A is dopamine detected after injection of 30 pmol of DA in normal tissue, whereas traces B and C are after the dopamine uptake inhibitor nomifensine was administrated. Uptake rate is reduced by nomifensine as seen by the increased concentration and the slower return to baseline after injection. (Data from Ref. (47) used by permission from Elsevier.)
the charging current is relatively stable at carbon electrodes, the current before analyte introduction can be subtracted from the current after analyte introduction. This method is known as background-subtracted FSCV, and it is best used for measuring concentration changes. The data collected is a plot of current versus voltage, called a cyclic voltammogram, which gives a chemical signature of the molecules detected. The peak locations and shapes help differentiate between different molecules, which gives this method the most selectivity of any of the electrochemical techniques discussed here. Some molecules, such as dopamine and norepinephrine, however, have similar CVs and are difficult to resolve.

R. Mark Wightman’s lab used carbon-fiber microelectrodes to pioneer the use of FSCV in measuring dopamine concentration changes with high temporal resolution. Early experiments focused on measuring dopamine concentration changes after electrical stimulation of the dopamine cell bodies (50). In more recent studies, behaviorally evoked dopamine concentrations have been measured. One example is the detection of dopamine in rats when cocaine was self-administered (Fig. 8f) (51). The subsecond temporal resolution allowed the observation that extracellular dopamine increased continuously for about 4 seconds before the lever press for cocaine administration and after the lever press. This observation shows that one role of dopamine is as an anticipatory signal (51). FSCV has also been used to measure other neurotransmitters, such as serotonin (52) and norepinephrine (53), in vivo and in brain slices. Our lab has recently extended the use of FSCV to detect the neuromodulator adenosine successfully. In vivo detection of adenosine and dopamine simultaneously was achieved, and the temporal resolution allowed us to demonstrate that adenosine accumulation was slower than dopamine (34). Fast-scan cyclic voltammetry is the best choice for detecting neurotransmitters that undergo volume transmission in vivo because the cyclic voltammogram provides a means to identify the species detected and the high temporal resolution allows correlation with behavior.

Summary and future directions

The main advantages of electrochemical methods for neurotransmitter detection are the high temporal resolution, the high sensitivity, and the small size of the microelectrode. The different electrochemical techniques have varying levels of selectivity; however, these methods are not as selective as separation methods coupled to microdialysis. The majority of studies have used a combination of anatomical knowledge, pharmacology, and electrophysiology to identify the analyte being detected. Most electrochemical methods are best for measuring fast changes and do not give information about basal levels. A major disadvantage of the electrochemical techniques is that the method is limited to the direct detection of electroactive molecules. New innovations are constantly being made to minimize these limitations. For example, enzyme electrodes are being developed for the indirect detection of non electroactive species. Enzymes are immobilized in a polymer coating, and the microelectrode detects an electroactive by-product, such as peroxide, from enzymatic activity (55, 56). Research continues into reducing the size of electrodes and increasing the number of molecules that can be detected simultaneously.

Conclusions and future of the field

Measuring neurochemicals is challenging because the brain is a complex matrix. Most studies have concentrated thus far on determining the effects of a single neurotransmitter. However, future research in this field inevitably will begin to examine interactions between various neurotransmitter systems, and techniques will be needed to measure multiple compounds. Separations-based methods, such as microdialysis, are particularly suited for this challenge. In addition, many researchers are starting to use multiple techniques for a single experiment. An example of this use is the combination of electrochemistry and electrophysiology techniques that can be now employed simultaneously at the same electrode (57). Another challenge is to reduce the amount of damage done to tissue from invasive techniques. Methods to microfabricate microdialysis probes or reduce the size of microelectrodes would be useful for maintaining tissue integrity. Noninvasive techniques, such as imaging, are becoming more prominent, especially for human studies. Many of the most popular methods, such as functional magnetic resonance imaging, measure changes in oxygen levels and blood flow but do not directly measure neurotransmitters (58). Positron emission tomography measures changes in receptor occupancy, a more direct measure of neurotransmission (59), but advances are needed to synthesize more positron-emitting ligands for different receptors. Imaging methods for cellular activity are also becoming more popular. For example, fluorescence resonance energy transfer (FRET) can be used to detect binding at receptors (60). However, these studies remain fairly complicated and often require genetic engineering of FRET-based pairs into a system. Optical probes, such as quantum dots or fluorescence-based dyes, may also soon see widespread use in detecting chemical changes in cell culture or brain slices.

The biggest advances for in vivo experiments will be in the area of designing better techniques for behaving animals. For electrochemistry, several groups have explored a wireless system of data collection in which the controls for electrochemistry and data are collected in instrumentation that can be placed in a backpack on the rat (61). The data then is sent wirelessly to a computer, so that the animal is not tethered to the computer. Similarly, telemetric electrophysiological systems are being developed that will enable wireless collection of unit recordings (62). These systems can be used to record spikes of neuronal activity in small behaving animals, such as birds. Integrated systems that measure both behavioral activity and neurochemical changes will also continue to be developed. Making measurements of neurotransmitters will continue to be challenging, but new techniques will lead to better insight into the basic neurobiology and regulation of neurotransmission.

References


Neurotransmission, Measuring Chemical Events in


45. Zahniser NR, Dickson SD, Gerhardt GA. High-speed chronoums


Further Reading


The nervous system is composed of networks of cells that engage in coordinated circuits to permit neural function. Within these precise neural circuits, communication between individual cells is primarily chemical in nature. Neurotransmitter release via exocytosis of neurotransmitter-filled synaptic vesicles is a fundamental step in this process. Here we overview the biochemical processes that regulate exocytic neurotransmitter release by focusing on three key stages: 1) loading of neurotransmitter into synaptic vesicles, 2) synaptic vesicle docking and priming reactions, and 3) calcium triggering of the vesicle fusion reaction. We also introduce the controversial topic of fusion-pore modulation as it pertains to the release of neurotransmitter. Lastly, we discuss current methods for detecting and quantifying neurotransmitter release.

Biological Relevance of Neurotransmitter Release

The process of information flow between neurons is termed synaptic transmission, and in its most basic form it is characterized by unidirectional communication from the presynaptic to postsynaptic neuron. The process begins with the initiation of an electrical impulse in the axon of the presynaptic neuron. This electrical signal—the action potential—propagates to the axon terminal, which thereby stimulates the fusion of a transmitter-filled synaptic vesicle with the presynaptic terminal membrane. The process of synaptic vesicle fusion is highly regulated and involves numerous biochemical reactions; it culminates in the release of chemical neurotransmitter into the synaptic cleft. The released neurotransmitter diffuses across the cleft and binds to and activates receptors on the postsynaptic site, which thereby completes the process of synaptic transmission.

Neurotransmitter release is not assured in response to synaptic stimulation. Rather, the process of vesicle fusion for individual release-competent vesicles is probabilistic. This process confers a discrete probability (between 0 and 1) that a given synapse will release neurotransmitter after an action potential (the synaptic release probability). For the majority of synapses in the central nervous system, the release probability at a defined synaptic contact is below 0.3, which leads to the often-quoted statement that the release process is “reliably unreliable” (1). Despite this fact, it has been demonstrated that some central nervous system synapses (in a variety of brain regions) do exhibit release probabilities as high as 0.9 (2–4). This higher synaptic release probability could be explained by more release-competent vesicles and/or because the vesicular release probability of the individual release-competent vesicles is higher. Moreover, release probability is highly dynamic; it incorporates several forms of short-term plasticity (5).

Although this article focuses almost exclusively on the essential aspects of release of classic small-molecule neurotransmitters from neurons of the central nervous system, it is appropriate to mention that the neurotransmitter release process encompasses several additional, salient facets. Release of neurotransmitters from neurons can occur at various locations on the neuron (e.g., axo-dendritic, axo-somatic, axo-axonal, dendro-dendritic, and, in the case of the neuromuscular junction, from axon to muscle). In addition, neurotransmitter release can occur from various different cell types (e.g., neuroendocrine cells and glial cells), in which case it can be debated whether the use of the term neurotransmission is truly appropriate. Release of neuropeptides is typically from dense-core vesicles rather than small synaptic vesicles, but many mechanisms parallel those for classic neurotransmitters. A group of diffusible messengers that includes nitric oxides, endocannabinoids, and hydrogen peroxide is often classified as retrograde neurotransmitters. However, these messengers are synthesized de novo rather than stored in vesicles and released, and thus they will not be considered here. Finally, even in the case of release of classic neurotransmitters from neurons of the central nervous system, highly specialized synapses have been described and investigated in great detail (e.g., ribbon-type synapses of the retina and Calyx-type synapses in the auditory pathway). These collective special considerations, although not adequately discussed herein, serve as excellent examples of the wide diversity...
of signaling mechanisms employed by the nervous system to achieve information transfer.

**Biological Chemistry of Neurotransmitter Release**

Within the presynaptic terminal, neurotransmitter-filled vesicles are clustered tightly in high numbers. The first electron micrograph images of synapses in which clusters of synaptic vesicles could be seen clearly were obtained in the mid-1950s (6, 7). This work coincided with the classic experiments of Bernard Katz and colleagues on the quantal nature of neurotransmission at the frog neuromuscular junction and thus greatly strengthened the foundation for the quantal hypothesis of neurotransmitter release (8-10). Together, these findings led to the vesicle hypothesis, for example, that a single synaptic vesicle is the morphological correlate of a quantum of neurotransmitter (10).

**Synaptic vesicle morphology and organization into functional pools**

As judged by electron micrograph images, small synaptic vesicles have a clear core and are approximately 35–50 nm in diameter (11–13). By contrast, dense core vesicles, which are found in neuroendocrine cells and for the storage of neuropeptide transmitters in the nervous system, can be as large as 300–400 nm in diameter. Within the presynaptic terminal, synaptic vesicles seem to be morphologically identical at near nanometer resolution. However, three distinct functional pools can be identified based on the differential ability of synaptic vesicles to be recruited for fusion. The vesicles that are docked at the membrane surface in a region called the active zone and that have undergone a series of priming reactions to achieve fusion competence represent the readily releasable pool (RRP) (14). It is believed that release of neurotransmitter will occur predominantly (if not exclusively) from this vesicle pool. Because the synaptic vesicle priming reactions may be reversible, a small subset of the morphologically docked vesicles may exist in the unprimed state and therefore would be unavailable for release (15).

After the fusion of a synaptic vesicle, the RRP is refilled from the recycling pool of synaptic vesicles. For central nervous system synapses (e.g., synapses of hippocampal neurons), the recycling pool of synaptic vesicles consists of approximately 30 vesicles, approximately three to five times the number of RRP vesicles (15, 16). During repetitive synaptic stimulation, the rapid refilling of the RRP from the recycling pool sustains continuous neurotransmitter release. An overview of the synaptic vesicle cycle is shown in **Fig. 1**.

The third and largest synaptic vesicle pool is termed the reserve pool and does not contribute to neurotransmitter release under normal physiological conditions. It is proposed that reserve pool vesicles are only recruited with extremely intense extended bouts of synaptic stimulation, conditions under which the recycling pool of vesicles is depleted (17). When vesicle pool sizes are expressed as percentages of the total synaptic vesicle cluster, these percentages hold well across many synaptic types and species. The RRP typically represents 1–2%, the recycling pool 10–20%, and the reserve pool 80–90% of the total vesicle cluster (18).

The synaptic vesicle as an organelle for neurotransmitter storage and release

Glutamate, gamma amino butyric acid (GABA) and glycine, acetylcholine (Ach), and monoamines are examples of major small-molecule neurotransmitters in the nervous system. Although each neurotransmitter has a distinct structure and biological activity, all share the common feature of being concentrated into synaptic vesicles before release from the presynaptic bouton of the neuron. In this respect, the synaptic vesicle is an organelle specialized for storage and release of neurotransmitter. Furthermore, the synaptic vesicle contains numerous proteins (both transmembrane and vesicle-associated) that are key players in the biochemical reactions that lead to release of neurotransmitter. The very recent detailed molecular characterization of the synaptic vesicle as a model trafficking organelle, including some of the first available estimates of key synaptic vesicle protein copy numbers (13), reemphasizes the fact that the synaptic vesicle is center stage in the consideration of the neurotransmitter release process (**Table 1**).

Although the synaptic vesicle landscape is dominated by an array of proteins, it is of note that synaptic vesicles that contain different neurotransmitters are thought to have largely similar overall protein composition. For example, all synaptic vesicles
require proteins that are essential for membrane trafficking and fusion. The best-studied synaptic vesicle proteins include the transmembrane proteins synaptobrevin, synaptophysin, synaptotubulin (also referred to as vesicle associated membrane protein or VAMP), and synaptic vesicle protein 2 (SV2), as well as the peripherally associated synapsins and the Ras that are attached through lipid modifications. The proton pump is also a key synaptic vesicle component and is critical for establishing the electrochemical gradient across the synaptic vesicle membrane (low pH in the lumen) to power neurotransmitter uptake into the synaptic vesicle lumen. The synaptic vesicle proton pump is unique in that only one copy per vesicle of this multi-protein complex exists, and this proton pump complex is by far the largest component of the synaptic vesicle [13, 18].

The neurotransmitter phenotype, i.e., what type of neurotransmitter is stored and ultimately will be released from the synaptic bouton, is determined by the identity of the neurotransmitter transporter that resides on the synaptic vesicle membrane. Although some exceptions to the rule may exist; all synaptic vesicles of a given neuron normally will express only one transport phenotype (this concept is enveloped in what is known as Dale's principle; see also Reference 19). To date, four major vesicular transporter systems have been characterized that support synaptic vesicle uptake of glutamate (VGLUT 1-3), GABA and glycine (VGA T), acetylcholine (VAChT), and monoamines such as dopamine, norepinephrine, and serotonin (VMA T 1 and 2). Vesicles that store and release neuropeptides do not have specific transporters to load and concentrate the peptides but, instead, are formed with the peptides already contained within.

**Synaptic vesicle docking and priming reactions**

Synaptic vesicle attachment or "docking" occurs in the active-zone region of the presynaptic terminal. This specificity for synaptic vesicle attachment implies a recognition mechanism between the synaptic vesicle and the active zone. In this context, some or perhaps all molecules responsible for synaptic vesicle docking might be expected to exhibit preferential enrichment or exclusive localization to the active-zone region. Despite intense efforts and numerous candidate molecules, the precise molecular mechanisms of synaptic vesicle docking remain poorly characterized at this time. The cytosolic protein Munc18 (which represents the major mammalian version of the sec1/munc18-like or SM family of proteins) is gaining acceptance as an important docking molecule [20], although no clear model for Munc18 function has emerged. Munc18 is proposed to provide the molecular link between the active-zone region and synaptic vesicles. It almost certainly links to the presynaptic membrane via a direct interaction with the N-terminal region of syntaxin [21]; however, the interaction by which Munc18 can also link with synaptic vesicles has proven highly elusive. No additional putative interacting partners have been validated as essential for synaptic vesicle docking. This fact may indicate redundancy in synaptic vesicle docking mechanisms. Alternatively, very recent evidence that demonstrates that syntaxin is required for synaptic vesicle docking [22], despite prior evidence to the contrary [23, 24], may largely explain the notable lack of progress in unraveling synaptic vesicle docking mechanisms to date, although this controversial finding demands additional validation. Notably, neither syntaxin nor Munc18 are preferentially enriched in the active zone; therefore, it is likely that some other protein participates to impart the regional specificity. Such a role has been ascribed to the Munc13 protein, which is reported to be concentrated at active zones [25] and to function upstream of syntaxin in the docking of synaptic vesicles [22].

**Prime** can be defined as the transition of synaptic vesicles from the docked state into the state of release competence. The available evidence supports a critical role for soluble N-ethylmaleimide-sensitive factor attachment protein receptor (SNARE) proteins in the biochemical interactions that underlie synaptic vesicle priming. Our current understanding of intracellular membrane trafficking—including synaptic vesicle fusion—has developed from a general model known as the SNARE hypothesis [26]. In this model, the formation of parallel-aligned α-helices between vesicle SNAREs and target-membrane SNAREs results in the formation of a remarkably stable SNARE complex [27]. In the case of synaptic vesicle fusion, the synaptic vesicle protein synaptobrevin serves as the vesicle SNARE, whereas SNAP-25 (synaptosomal-associated protein of 25 kDa) and syntaxin on the presynaptic plasma membrane serve as the target-membrane SNAREs. These three

**Table 1**  
Estimated copy number per vesicle and proposed functions for selected major synaptic vesicle proteins

<table>
<thead>
<tr>
<th>Synaptic vesicle protein</th>
<th>Copies per vesicle</th>
<th>Proposed function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synaptobrevin/VAMP2</td>
<td>70</td>
<td>Priming</td>
</tr>
<tr>
<td>Synaptophysin</td>
<td>32</td>
<td>Vesicle recycling?</td>
</tr>
<tr>
<td>Synaptotagmin 1</td>
<td>15</td>
<td>Calcium sensor</td>
</tr>
<tr>
<td>Neurotransmitter transporter</td>
<td>9-14</td>
<td>Neurotransmitter loading</td>
</tr>
<tr>
<td>Rab3B</td>
<td>10</td>
<td>Docking, priming</td>
</tr>
<tr>
<td>Synapsins</td>
<td>8</td>
<td>Regulation of vesicle mobility?</td>
</tr>
<tr>
<td>SV2</td>
<td>2</td>
<td>Priming, transporter?</td>
</tr>
<tr>
<td>Proton pump (V-ATPase)</td>
<td>1</td>
<td>Neurotransmitter loading</td>
</tr>
</tbody>
</table>

* Estimates of protein copy number per vesicle originally reported by Takamori et al. [13].

**Figure 2**  
**Neurotransmitter: Release**  
**Table 2**  
Estimated copy number per vesicle and proposed functions for selected major synaptic vesicle proteins

<table>
<thead>
<tr>
<th>Synaptic vesicle protein</th>
<th>Copies per vesicle</th>
<th>Proposed function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synaptobrevin/VAMP2</td>
<td>70</td>
<td>Priming</td>
</tr>
<tr>
<td>Synaptophysin</td>
<td>32</td>
<td>Vesicle recycling?</td>
</tr>
<tr>
<td>Synaptotagmin 1</td>
<td>15</td>
<td>Calcium sensor</td>
</tr>
<tr>
<td>Neurotransmitter transporter</td>
<td>9-14</td>
<td>Neurotransmitter loading</td>
</tr>
<tr>
<td>Rab3B</td>
<td>10</td>
<td>Docking, priming</td>
</tr>
<tr>
<td>Synapsins</td>
<td>8</td>
<td>Regulation of vesicle mobility?</td>
</tr>
<tr>
<td>SV2</td>
<td>2</td>
<td>Priming, transporter?</td>
</tr>
<tr>
<td>Proton pump (V-ATPase)</td>
<td>1</td>
<td>Neurotransmitter loading</td>
</tr>
</tbody>
</table>

* Estimates of protein copy number per vesicle originally reported by Takamori et al. [13].
SNARE proteins assemble into a heterotrimeric SNARE complex (or core complex). The crystal structure of the core complex revealed a bundle of four $\alpha$-helices, one each contributed by synaptobrevin and syntaxin and two contributed by SNAP-25 (38). The process of SNARE complex assembly proceeds from N-terminal to C-terminal direction in what is commonly referred to as a “zippering” action, which brings the C-terminal membrane anchors of the transSNAREs together (29). This action is proposed to force the closely apposed membranes together to initiate fusion. The transition from unstructured monomeric SNAREs to the assembled SNARE complex likely proceeds by multiple sequential steps rather than at once, and zippering can only occur when all SNAREs are aligned in parallel. One model proposes that initially an acceptor complex forms between SNAP-25 and syntaxin on the plasma membrane (30–32). Only then would the acceptor complex interact with the vesicular SNARE synaptobrevin, aligning the N-terminal ends in parallel, first in a loose conformation, and then undergoing N-terminal to C-terminal zippering. The partially zipped SNARE complex is arrested in a partial fusion or prefusion state and awaits the influx of calcium to activate the neuronal calcium sensor and stimulate the completion of the fusion reaction. These processes are summarized in Fig 2. One important open question is how SNARE zippering might become arrested before completion. Recent evidence supports a role for complexins at this step. Complexins are enriched in the presynaptic compartment, are crucial for highly synchronous evoked neurotransmitter release, and exhibit direct binding to the SNARE complex (33, 34). Although much evidence on complexin function is conflicting or controversial, most available evidence is consistent with a model in which complexins help to stabilize the partially zipped SNARE complex by direct binding but prevent full SNARE assembly and/or completion of fusion before the arrival of the calcium trigger (34–37). However, complexins do not merely function as inhibitory fusion clamps; it has been demonstrated that complexins are capable of both inhibitory and facilitating functions and thus are well suited to impart exquisite control to the final stages of calcium-dependent neurotransmitter release (34).

Although little doubt remains that the formation of the SNARE complex is the critical step in synaptic vesicle priming, several other proteins have been implicated as important regulators of the priming reactions. Two prominent examples are Munc18 and Munc13; these crucial proteins apparently function in both synaptic vesicle docking and priming. The available evidence is largely compatible with an essential role for both Munc18 and Munc13 in facilitating priming by regulating the proper assembly of the SNARE complex, which may be accomplished by the stabilization of the putative SNAP-25-syntaxin acceptor complex mentioned previously. It is unclear whether there is convergence exists between Munc18 and Munc13 in their respective modes of action in this context. To add another layer of complexity, Munc13 also has been shown to antagonize the action of a soluble protein called Tomosyn (38). Tomosyn contains a C-terminal SNARE motif through which it can form a tight complex with SNAP-25 and syntaxin and effectively deter the formation of the prototypical neuronal SNARE complex (39). Thus, Tomosyn serves as a negative regulator of vesicle fusion. Munc13 either prevents the formation of the Tomosyn-containing SNARE complexes or disrupts Tomosyn-containing SNARE complexes to liberate SNAP-25/syntaxin heterodimers, which then would become available to bind to synaptotubulin on the synaptic vesicle. Either model is again consistent with a positive regulatory role...
for Munc 13 in synaptic vesicle priming. Other putative regulators of synaptic vesicle priming include RIM, Rab3, and SV2, although the exact mechanisms for how these proteins regulate priming are much less defined at this time.

**Triggering of the synaptic vesicle fusion reaction**

Once synaptic vesicle docking and priming is completed, the final triggering of vesicle fusion with the presynaptic terminal membrane occurs rapidly in response to action potential invasion of the terminal. This step is highly dependent on calcium ions, which enter the terminal through voltage-gated calcium channels (40–42). The idea that calcium entry into the terminal is a key step in the neurotransmitter release process formed the basis of what is known as the “calcium hypothesis” (43). Although the identity of the molecular calcium sensor for neurotransmission was not known at the introduction of the calcium hypothesis 40 years ago, it is now widely accepted that the calcium-dependent triggering of synaptic vesicle fusion is imparted by the synaptic vesicle protein synaptotagmin (44–46).

The work of Dodge and Rahamimoff (47) initially detailed the quantitative dependence of neurotransmitter release on external calcium concentration at the neuromuscular junction. The relationship was reported to be highly nonlinear with an approximately fourth-order cooperativity. Remarkably, these findings are in near perfect agreement with a more modern study of the relationship between neurotransmitter release and intracellular calcium concentration at a central nervous system synapse (48). Although the basis for the cooperativity of release is not yet clear, multiple calcium coordination sites exist with the tandem C2 domains of synaptotagmin, which suggests that this cooperativity of release may originate with the intrinsic calcium-binding ability of the synaptotagmin protein itself. This suggestion is supported by studies that show that mutations in the synaptotagmin C2 domains alter the apparent degree of cooperativity (45, 49).

How does synaptotagmin trigger synaptic vesicle fusion in response to calcium influx and binding? Synaptotagmin engages in calcium-stimulated binding of both SNARE complexes and membrane phospholipids (50–52). Although still debated, available evidence supports a model of synaptotagmin action in which calcium binding to the C2 domains causes membrane penetration and induction of positive membrane curvature. This buckling of the active zone membrane locally reduces the energy barrier for fusion, which allows the fusion process to proceed to completion (53). A direct interaction of synaptotagmin to the SNARE complex is important in this model (calcium-dependent and/or-independent) to ensure that the local membrane buckling is targeted appropriately to the membrane beneath the vesicle; thus, the SNARE binding and phospholipid penetration activities of synaptotagmin likely work in concert to liberate neurotransmitter (52–55).

Vesicle fusion and neurotransmitter release in response to action potentials is termed evoked release (the main focus of this article); however, it is important to note that action potentials are not absolutely required for liberation of neurotransmitter. Spontaneous release events (also referred to as spontaneous miniature events or “minis” for short) occur in the absence of an overt stimulus (56–58). Evoked and spontaneous synaptic vesicle fusion likely share many commonalities, including the involvement of the same neuronal SNARE complex. However, perturbations to SNAREs can differentially affect evoked and spontaneous events, which suggests that the fusion reactions mediating evoked and spontaneous release events might differ to some extent (56–58).

**Regulation of neurotransmitter release via modulation of the fusion pore**

Although the molecular components of the neuronal fusion pore remain largely unknown, modulation of the fusion pore has been proposed as a mechanism for regulating release of neurotransmitter. Specifically, two modes of release can be summarized: 1) a classic full-fusion model in which the fusion pore completely dilates and the vesicle fully collapses into the plasma membrane of the active zone (59), which thereby releases a vesicle’s full complement of neurotransmitter, and 2) a kiss-and-run model in which a narrow fusion pore is stabilized transiently but then closed rapidly (60). In the kiss-and-run model, the presence of a narrow fusion pore is hypothesized to restrict neurotransmitter efflux, and it thereby might reduce quantal size (62). The prevalence of kiss-and-run-type fusion events has been debated intensely; particularly, the prevalence of these events in central nervous system neurons (which we focus on here) has been debated, with some studies of hippocampal neurons estimating greater than 80% of all fusion events being of this type at low firing frequencies (62) and others reporting negligible kiss-and-run in the same preparation (63). The most recent and direct study to date reported only a 3% rate of occurrence for kiss-and-run at the specialized Calyx of Held synapse (64). Thus, although its frequency is widely debated, it is generally accepted that a nonclassic mode of fusion that is reasonably well described by the kiss-and-run model does in fact exist.

Does kiss-and-run-type fusion influence neurotransmitter release at central nervous system synapses? This topic is a current topic of research that is yet to be resolved; even the existence of kiss-and-run exocytosis in the central nervous system is still controversial. It is worth noting that most current methods for detecting alterations to fusion pore size and/or conductance at central nervous system synapses have been unable to provide simultaneous direct measurements of neurotransmitter release. Thus, a marked absence of evidence exists to support the idea that fusion pore modulation can modulate the amount of neurotransmitter released during vesicle fusion. Of those that have achieved the technical feat of measuring neurotransmitter release during kiss-and-run-type fusion events, one study using a nonphysiological manipulation found no evidence for modulation of quantal size at glutamatergic central nervous system synapses (65), and another study detected, in a subset of release events, a predominant flickering fusion pore mechanism in which each flicker released only 25–30% of the total neurotransmitter contained within a vesicle at dopaminergic central nervous system synapses (66). It is likely that much of the controversy surrounding kiss-and-run can be resolved through...
advancements in technical approaches for measuring neurotransmitter release.

Methods for Quantification of Neurotransmitter Release

To quantitate neurotransmitter release, we ideally would use approaches that can make noninvasive, selective, analytical measurements in situ with millisecond temporal resolution and micron or better spatial resolution. Each of these criteria has been fulfilled by one or more currently available methodologies. However, to date, no single technique is universally suited to all contexts of neurotransmitter-release quantification. Thus, compromises are made in the criteria that are less important to maximize the criteria that are most important to the biological question at hand. In choosing a technique, perhaps the first question is, what type of biological preparation is most appropriate for the study? Advantages of in vitro preparations make these preparations better suited for some lines of work, but advantages of in vivo preparations are better for others. Importantly, some techniques are better suited for some biological preparations than others. Table 2 lists the preparations in which each of the techniques discussed below have been most commonly employed.

Table 2 Tools for the quantification of neurotransmitter release and the biological preparations in which they have been used

<table>
<thead>
<tr>
<th>In vitro</th>
<th>In vivo</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell culture</td>
<td>Tissue slice</td>
<td>Anesthetized</td>
</tr>
</tbody>
</table>

- **Radiolabeled neurotransmitters**
  One approach that has been used quite widely to quantitate neurotransmitter release employs radiolabeled (tritiated) neurotransmitter analogs (e.g., Reference 67). First, tissue is incubated in a buffer solution that contains tritiated neurotransmitter. During this time, the radiolabeled transmitter is taken up into cells by endogenous plasma-membrane transporters and packaged into vesicles by vesicular transporters. The tissue preparation then is rinsed in buffer to remove extracellular radiolabeled transmitter leaving only that which was taken up into cells. This stored transmitter is then released over time by exocytosis. To quantitate its release, the tissue is continuously perfused with buffer, and time-dependent aliquots are collected. Radioactivity is measured in the aliquots with a scintillation counter and is used as an index of endogenous neurotransmitter release. Rather than estimate absolute neurotransmitter release, this method is typically used to compare the relative release between two or more conditions.

- **Electrophysiological methods**
  Electrophysiological techniques have high temporal (millisecond) and spatial (micron) resolution. They are carried out at discrete electrodes and so obtain information from a single spatial location or from a finite number of locations if a multiple-electrode system is used. These recording techniques...
use fairly standard electrophysiology recording resources and capabilities to measure neurotransmitter release, and so they are widely used.

Postsynaptic electrophysiological recordings detect the neurophysiological response of a target cell following release of neurotransmitter and represent the least deviation from a standard electrophysiological recording experiment. These recordings are usually carried out at a patch-clamp electrode using a voltage clamp to measure postsynaptic currents (68), but the current clamp can be used if measurement of postsynaptic potentials is preferred. Selectivity of the responses for the neurotransmitter of interest can be achieved with appropriate pharmacological isolation by using a cocktail of antagonists for other neurotransmitters and/or with electrophysiological isolation (by manipulating the holding potential). Postsynaptic signals can be detected for both spontaneous and evoked neurotransmitter release events. The frequency of spontaneous postsynaptic events is often used as an index of neurotransmitter release. Information can also be obtained from the amplitude of postsynaptic responses; however, this metric encompasses both changes in neurotransmitter release (presynaptic effects) and changes in synaptic strength (postsynaptic effects). Presynaptic effects are usually inferred from experiments where postsynaptic responses are evoked by paired pulses of extracellular electrical stimulation separated by tens of milliseconds. The paired-pulse ratio (PPR), that is the amplitude of the response to the second pulse divided by that of the first, is believed to be a reflection of release probability, with low PPR signifying high release probability (argued to be because of depletion of readily releasable vesicles on the first pulse; see Reference 5). Thus, changes in PPR are taken to indicate changes in the amount of neurotransmitter release that can be evoked by an electrical impulse.

During neurotransmitter release events from a cell, the surface area of its plasma membrane changes; it increases during fusion and decreases during subsequent exocytosis. Because the plasma membrane acts as an electrical capacitor, these dynamics can be detected as changes in the total membrane capacitance. Thus, membrane capacitance measurements with whole-cell or cell-attached patch clamp can be used to quantitate exocytosis (69). Using voltage clamp, a sine wave command signal is applied, and the current is measured with a lock-in amplifier to derive the instantaneous capacitance. With these recordings, individual exocytotic events can be detected. In cell-attached preparations, conductance of fusion pores that form within the patch can be measured simultaneously.

The electrophysiological biosensor or “sniffer patch” (70) uses an outside-out patch excised from a donor cell, which has high-density expression of ionotropic receptors for the neurotransmitter of interest. Ideal donor cells are those in which the receptor has been stably overexpressed and low expression of other potential interfering receptors is detected. The electrode, incorporating the patch, is placed close to a putative release site, and channel conductance within the patch is measured in voltage-clamp mode. The neurotransmitter detection range of this technique is quite narrow, around the EC50 of the receptor, but it can be broadened somewhat by increasing the density of receptor expression in the patch. Nonetheless, the detection limit is, by definition, in the physiological range.

The sensor response is nonlinear with analyte concentration, but it can give quantitative information with appropriate calibration.

Electrochemical methods

Electrochemical detection involves the induction of a change in redox state (electrolysis) by application of an electrical potential to an electrode (71). Compounds that can be readily detected by this means are termed electroactive. Under physiological conditions, these compounds tend to be in their reduced state in the nervous system because of the rich level of antioxidants (e.g., ascorbic acid) and, thus, can be oxidized by application of a positive potential to the electrode. The evolved electrons are detected at the electrode in the form of electrical current. This current is proportional to the number of electroactive molecules at the surface of the electrode, and therefore it is proportional to their concentration in the bulk solution. By implanting an electrode in the extracellular space close to the release site and detecting changes in the local (extracellular) concentration of the neurotransmitter, neurotransmitter release can be monitored. The key advantage of this approach is the high temporal resolution that can be in the millisecond domain. Neurotransmitters that can be detected this way include dopamine, norepinephrine, epinephrine, serotonin, and melatonin.

Several variants of this technology differ by the voltage command waveform applied to the electrode to induce electrolysis. The simplest waveform is a constant direct-current potential. This form is known as constant-potential amperometry. This technique yields a constant readout of neurotransmitter fluctuations in real time and provides the fastest available chemical monitoring. For example, using constant-potential amperometry, data has been acquired in the high microsecond range that is sufficient to resolve release during multiple partial fusion events (flickering) (66). However, the disadvantage of this type of electrochemical detection is that it lacks chemical selectivity. Thus, its use is limited to environments where the identity of the analyte is predictable (see Reference 71).

Another commonly used waveform is a square wave where the electrode is held at a nonoxidizing potential and transiently pulsed to an oxidizing potential. An example of this form is high-speed chronopotentiometry, which uses oxidizing pulses that typically last for around 100 ms and are repeated each second. This method provides information (current) both on the oxidation of compounds at the surface of the electrode and on the reduction of the oxidized material. This additional information that is obtained during each measurement aids in the identification of the analyte.

To improve chemical selectivity, a triangular input waveform can be used that separates compounds into resolvable peaks. This form—cyclic voltammetry—can be carried out with high temporal resolution using high scan rates to allow the waveform to be completed in a short time. In fast-scan cyclic voltammetry (also known as fast cyclic voltammetry), waveforms last around 10 ms, and measurements are typically made every 10–200 ms. The newest generation of electrochemical-based biological detection devices is the biosensors. These devices combine the high spatial and temporal resolution of an electrochemical microsensor with a biologically selective recognition element.
current devices, this element is most often an enzyme that is selective for the neurotransmitter of interest. The enzyme exists in a polymer layer that is embedded on an electrochemical probe. When the analyte binds to the enzyme, a chemical coreaction is initiated that generates an electroactive reporter molecule. The reporter molecule is detected by the electrochemical probe using constant-potential amperometry and thus transduces the biological signal to an electrochemical signal. This approach has three main advantages over other electrochemical sensors. First, the range of possible analytes is not limited to those that are electroactive. In fact, it is advantageous for the analyte not to be electroactive so that its direct detection at the electrochemical sensor does not interfere with detection of the reporter molecule. Second, the selectivity of the probe is conferred by a recognition element that can be highly specific. Third, this technology lends itself to upgradeability; new recognition elements can be incorporated into probes as they are developed.

Optical methods

Optical methods offer the benefit of extremely high spatial resolution, in the nanometer range. Moreover, unlike electrode-based techniques (e.g., electrophysiological and electrochemical methods), these measurements can be made over a wide area; that is, rather than obtaining a single time-dependent measurement, multiple parallel processes can be monitored simultaneously at discrete spatial loci (e.g., Reference 72). Currently, optical methods for quantification of neurotransmitter release are not widely used in vivo because of insufficient deep-brain optical access. However, a rapidly growing movement is working to develop these tools for use in systems neuroscience (e.g., using techniques that can penetrate deep enough to measure from superficial cortical layers or fiber optics to access deeper regions). Fluorescent styryl dyes such as FM 1-43 have been used to approximate neurotransmitter release by measuring rates of exocytosis (16, 72, 73). These dyes reversibly label endosomal membranes and can be taken up into intracellular synaptic vesicles during endocytosis in systems in which vesicle recycling is not used, if desired.

Exocytosis. The primary advantage is that this methodology is not constrained to a single analytical tool, and detection schemes can be tailored to the needs of the particular experiment and to measure multiple analytes simultaneously. In fact, any tools of analytical chemistry can be used, including those that cannot easily be miniaturized into an in situ device, for example, gas chromatography or mass spectroscopy. Sampled material is often analyzed offline, but online analysis can be achieved by coupling a detector to the sampling-device outlet if desired.
Microdialysis sampling (78) is conducted using a probe that has an inlet and outlet tube (joined via a semipermeable membrane often a concentric design with the inlet inside the outlet tube). The probe has a molecular weight cutoff high enough that it is permeable to small-molecule and peptide neurotransmitters. Buffer (usually artificial cerebrospinal fluid) is perfused through the inlet tube past the membrane. When the probe is implanted in the brain, neurotransmitters in the extracellular space diffuse across the membrane into the probe down their concentration gradient. An equilibrium between the extracellular and intraprobe neurotransmitter concentration is accomplished, which is dependent on the flow rate, membrane size, and other factors. The ratio between the analytic concentration inside the probe and that on the outside at equilibrium is the relative recovery. Under most experimental conditions, the relative recovery is below 100%, and thus the neurotransmitter concentration in the dialysate is an underestimate of the absolute extracellular concentration. To better measure the absolute basal transmitter level, the no net flux method can be used in which the analytic of interest is perfused through the probe at different concentrations. When the analytic concentration is higher in the inlet than outside the probe, a net flux moves into the probe, and the concentration will be lower in the outlet than the inlet. Conversely, when the analytic concentration is lower in the inlet than outside the probe, a net flux moves into the probe (and the concentration will be higher in the outlet than the inlet). Therefore, the point of no net flux (which can be interpolated if not measured directly) represents the condition in which the inlet concentration matches the concentration of the analytic outside the probe.

Another means of sampling is push-pull perfusion (79) using a probe with an inlet and outlet tube. With this method, a small amount of cerebrospinal fluid is pulled directly from the brain through the outlet tube and replaced with artificial cerebrospinal fluid. This approach has greater spatial resolution than microdialysis; and because cerebrospinal fluid is collected directly, no concern develops that account for spin–lattice relaxation (T1) and spin–spin relaxation (T2) and depend on tissue molecular interactions. Spin–lattice relaxation is the relaxation of the proton spin axis in the longitudinal axis of the magnetic field (z-axis), and spin–spin relaxation is in the transverse (xy) plane. With the appropriate pulse sequence, a blood–oxygen-level-dependent (BOLD) signal, the hemodynamic response, can be extracted from the T2 signal and is used to infer regional brain activity. These types of measurements are typically made during sensory stimulation, in behavioral or cognitive tasks (functional MRI; fMRI) or after administration of pharmacological agents (pharmacologic MRI; phMRI).

One of the largest drawbacks of these techniques is the time it takes to collect sufficient material for an analytic sample. Typically, this time has been on the order of tens of minutes, and thus these techniques are best suited for measuring steady-state levels of neurotransmitter and slow signal changes. However, in recent years, the temporal resolution has been improved significantly to the level of seconds by using capillary and microfluidic devices for sample collection combined with online separation and detection (79, 80).

**Neuroimaging**

Neuroimaging is the least invasive means to measure neurotransmitter release and can be used in living animals or humans without surgical procedures. These techniques permit monitoring over a large area (e.g., the whole brain) in three dimensions with millimeter spatial resolution and temporal resolution in seconds to minutes (81).

Positron emission tomography (PET) detects and spatially localizes radioactive sources, and it can be used to quantitate neurotransmitter release by measuring displacement of radioiodinated ligands (usually antagonists) from neurotransmitter receptors. The radioligand is administered systemically and is sequestered in tissue by selective binding to its receptor. On radioactive decay, a positron is emitted that collides with an electron (annihilation) and produces two photons (gamma rays) that travel in opposite directions. The scanner has a ring of gamma detectors that senses the arrival time of the photons. From the position in the ring and the difference in time of flight between the photons, the location of the annihilation within the plane of the ring can be calculated. The third dimension can be reconstructed by conducting serial measurements in adjacent “slices.” Thus, PET provides a spatial map of the radioligand throughout the brain, which can be updated every few minutes. When an increase in endogenous neurotransmitter release occurs, the radioligand is displaced from receptors, and the tissue radioactivity level is reduced. Note that because this method relies on competitive displacement of a receptor-bound radioligand, increases but not decreases in endogenous neurotransmitter levels (i.e., increases but not decreases in the rate of ongoing neurotransmitter release) can be detected.

Magnetic resonance imaging (MRI) is included here because it is the least invasive method for gathering information about neurotransmission in living humans. However, it should be noted that the inferences about neurotransmitter release are extremely indirect. This technique is used to measure, on a timescale of seconds, local hemodynamic changes that are assumed to correlate with neurotransmission. The experiment is carried out in a high-field magnetic field (≥1 tesla). Excitation by radio-frequency pulses alters the spin axis of hydrogen nuclei (protons; including those in water), which can be measured as changes in the field strength. The dynamics of relaxation of the proton spin axis back to equilibrium is that account for spin–lattice relaxation (T1) and spin–spin relaxation (T2) and depend on tissue molecular interactions. Spin–lattice relaxation is the relaxation of the proton spin axis in the longitudinal axis of the magnetic field (z-axis), and spin–spin relaxation is in the transverse (xy) plane. With the appropriate pulse sequence, a blood–oxygen-level-dependent (BOLD) signal, the hemodynamic response, can be extracted from the T2 signal and is used to infer regional brain activity. These types of measurements are typically made during sensory stimulation, in behavioral or cognitive tasks (functional MRI; fMRI) or after administration of pharmacological agents (pharmacologic MRI; phMRI).

**References**

Neurotransmitter: Release


12. Harris KM, Sultan P. Variation in the number, location and size of synaptic vesicles provides an anatomical basis for the nonuniform probability of release at hippocampal CA1 synapses. Neuropharmacology 1995;34:1387-1395.


Release


See Also

Membrane Trafficking
Membrane Fusion, Mechanisms of
Neurotransmitter: Production and Storage
Neurotransmitter: Uptake and Degradation
Neurotransmission, Chemical Events in
Synaptic Chemistry
Iron-Sulfur World, Origin of life in an
Günter Wächtershäuser, Chapel Hill, North Carolina
doi: 10.1002/9780470048672.wecb264

The theory of a chemo-autotrophic origin of life in an Iron-Sulfur World postulates quenched, aqueous flows of volcanic exhalations as sites for the emergence of the pioneer organism of life, which are characterized by a composite structure with an inorganic substructure and an organic superstructure. Iron, nickel, and other transition metal centers in the substructure are catalytic for the reductive formation of low-molecular-weight organic compounds by carbon fixation from volcanically derived carbonyl, cyanide, sulfido, aminido, and other ligands. Some organic products become transition metal ligands in statu nascendi with the effect of ligand-accelerated autocatalysis as a chemical basis of pioneer reproduction, inheritance, and evolution. This unitary structure-function relationship of the pioneer organism constitutes the "Anlage" for two major parallel strands of evolution, genetization, and cellularization, and the "mystery" of an increase of complexity throughout evolution is simply caused by the primal synthetic drive of the autotrophic pioneer organism that may well be persisting to this day.

Conventional theories on the origin of life are characterized by a conjunction of several deep-seated misconceptions. Since the time of Nägeli (1), the oldest and most persistent misconception is based on the fact that the bulk material of all extant organisms consists of polymers. It is, therefore, assumed widely that the first organisms of life were engaged already in the polycodensation of monomers to autocatalytic nucleic acids ("RNA World") or to autocatalytic sets of polypeptides ("Protein World"). The origin of the monomers is relegated typically to an obscure chemical evolution in a "prebiotic broth" over thousands or millions of years. This view has been criticized severely, notably by Shapiro (see Further Reading). It suffers from a stark paradox. Liquid water is postulated as the medium of polycodensation; yet it is precisely this medium that tends to counteract polycodensation or polycodensation agents because of the hydrolytic mass effect of liquid water, notably under hot, non-neutral conditions. In the last 20 years, a new approach has been developed (2–5), which aims to avoid these pitfalls. The new theory is based not on a polymer metabolism with catalytic polymers, polycodensation, and replication, but rather on a small-molecule metabolism with catalytic transition metal precipitates and autocatalytic carbon fixation, which gives rise to small organic ligands. It is not restricted to main group chemistry, but it is critically dependent on transition metal coordination structures, organo-metal chemistry, and redox reactions, for which liquid water is a benefit rather than a detriment. In view of the dominance of iron among transition metals and the ubiquity of sulphur, the term "Iron-Sulfur World" has been coined for the setting of the origin of life (3), which is viewed not as a protracted affair of prebiotic chemistry but as a nearly instantaneous induction of an autocatalytic "pioneer organism." The new theory is recognized in the field as one of the "two main theories on the origin of life" and as having generated a conflict over "hot, volcanic origin" versus "cold, oceanic origin" (6).

General Organization and Operation of the Pioneer Organism

The new theory places the pioneer organism in a locally and temporally coherent volcanic flow setting. Therefore, all its assumptions must be compatible with each other. Such a comprehensive theory cannot be conceived in one stroke. It has to evolve. Indeed, the theory presented here has been evolving over the past 20 years, its progress guided by chemical experiments and by the increase of its relative explanatory power (Popper). The partial aspects of the theory as it now stands shall be addressed one at a time in a manner that makes their coherence within the theory apparent.
Structural organization of the pioneer organism

An origin of life by autocatalytic carbon fixation requires a sufficient degree of localization and physical/compositional coherence for preventing decay of the chemical potentials by diffusion and dilution, while permitting access of the volcanic nutrients. These requirements are satisfied by a minimal structural organization that consists of an inorganic substructure of crustal origin and an organic superstructure of volcanic gas origin. Catalytic transition metal centers (e.g., Fe, Co, Ni, Cu, Mn, Mo, W, and V) in the inorganic substructure are exposed to volcanic gas nutrients (CO, HCN, CH₃SH), which are quenched in the liquid water phase and bonded as ligands to the transition metal centers. The chemical potential of these inorganic ligands in conjunction with the catalytic properties of the transition metal centers leads to synthetic reaction cascades that generate low-molecular-weight organic compounds, which are functionalized sufficiently for bonding in situ as ligands to the transition metal centers (2). Thereby they become constituents of the organic superstructure attached to the transition metal centers in outer or inner surfaces of the inorganic substructure. The organic superstructure and the proximal regions of the inorganic substructure define the “pioneer organism.”

The organic superstructure constitutes a dynamic ligand sphere. Its constituents may undergo lateral transfer along the surfaces. At any given time the composition of the superstructure is a steady state between the rate of formation of organic constituents by carbon fixation and the rate of loss of organic constituents by detachment and diffusion into the vast expanses of the ocean (2).

The inorganic substructure should also be viewed as being dynamic. According to Ostwald’s rule of steps, it will change its constitution mainly by conversions of the structure-forming bridging ligands. It may begin as an extended polynuclear, poly-modal, heteroleptic, and hydrated complex or as amorphous, bridging ligands. It may begin as an extended polynuclear, poly-modal, heteroleptic, and hydrated complex or as amorphous, hydrated hydroxide minerals or hydrogels (e.g., [Fe₃N][OH]₄), which subsequently stabilize by crystallization. The additional development depends on the relative rates of introduction of polymerizing bidentate ligands (cyanidation, dehydrogenation, or sulfidation) versus depolymerizing ligands (carboxylation or ammoniation). Under conditions of high-sulfide activity, the substructure will convert to sulfides, which are at first amorphous or poorly crystalline, and finally to pyrites.

Inorganic starting conditions of the pioneer organism

The discovery that liquid water may well have existed on Earth 4.4 billion years ago has made room for the possibility that life originated deep in the Hadean, i.e., at a time when magma and crust of the Earth were much hotter than today and much more reducing. As a consequence, the water gas equilibrium of the volcanic gases

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad (1)$$

favors a high concentration of CO at the magmatic source. For example, under conditions of saturation with graphite, the molar ratio of CO/CO_2 is about 1:1 at 1200 °C and 2 kbar or at 900 °C and 0.1 kbar. If the rate of quenching along the flow path (by cold water or ice) is fast compared with the rate of equilibration, a significant disequilibrium concentration (chemical potential) of CO is the result (5). Quenched volcanic gases also contain H_2, which increases in concentration by water gas equilibration (2), as well as the sulfur compounds H_2S, CH₃SH, and COS; the nitrogen compounds NH_3, HCN, and HCN (5), and the volatile phosphorus compound P_2O_5 (7).

The formation of low-molecular-weight organic compounds by carbon fixation from volcanic nutrients involves primarily electron transfer (redox) reactions. Redox reactions are the most common reactions, catalyzed by metals that can have different oxidation states, with the catalytic activity of a transition metal center dependent on the nature of the ligands bound to the metal center. It is well established that the catalytic activity of a transition metal center is dependent on its ligands and that organic ligands may increase the catalytic activity by a factor of up to 10^4. This ligand feedback is unpredictable theoretically and idiosyncratic (8). Therefore, it is suggested that certain products of the carbon fixation pathways may increase the catalytic activity of transition metal centers as ligands. This result may well be the simplest and earliest form of autocatalytic product feedback. Autocatalytic product feedback (ligand-accelerated autocatalysis) constitutes “metabolic reproduction,” which gives rise to growth and inheritance by lateral spreading.

In broad terms, the proposed origin of life by autocatalytic ligand feedback is characterized by an extreme paucity of chemical possibilities because the starting materials are inorganic and the organic products have a low molecular weight. At the same time, the chemistry is highly selective because of the chemical energy in combination with specific transition metal catalysis. As a consequence, only one possibility may exist for a chemo-autotrophic origin of life.

Flow setting of the pioneer organism

Precious little geological evidence exists concerning the geochemical conditions of the Hadean Earth. Cockell (see Further Reading) has proposed that heavy bombardment caused impact cratering, rock fracturing, and deposition of debris creating a bed with myriads of diverse flow ducts for volcanic, hydrothermal fluids. Such a flow bed would undergo several characteristic transformations. The primal crust beneath the Hadean ocean may well have consisted of ultramafic rocks (e.g., komatiite), with
Ni(OH)₂ generates Ni-hydrogenase: source, which is correlated with the activity of extant Fe, Ni(CN)₂ having a layered crystal structure (9), and organic slowest travelers.

Differential ligand bonding strength would exhibit differential re-

The alkaline (Mg, Ca)(OH)₂ buffers the pH. After exhaustion of the buffer capacity by reaction with CO₂ and with the acidic products of carbon fixation, the flow ducts undergo a pH development from alkaline toward neutral (pH zoning). (Fe, Ni)(OH)₂ undergoes cyanidation, carbonylation, sulfidization, and pyritization. Specifically, cyanidation of Ni(OH)₂ generates Ni(CN)₂, having a layered crystal structure (9), and organic compounds may become trapped in the interspaces between the Ni(CN)₂ layers to form clathrates.

As a consequence of the notions of surface catalysis and volcanic flow (2), the flow ducts would operate like a chromatographic reactor with interacting reactive and chromatographic processes. Organic constituents of the superstructure with differential ligand bonding strength would exhibit differential retention or residence times, with the best surface bonders being slowest travelers.

**Toward the (Bio)Chemistry of the Pioneer Organism**

We now turn to the (bio)chemistry of the pioneer organism and begin with reactions that provide reducing power and nutrients for the pioneer metabolism. The (bio)chemistry of the pioneer organism is testable experimentally with the aim of correlating extant metal/o-enzymes with inorganic transition metal catalysts and with the ultimate goal of establishing a reproducing and evolving pioneer organism.

**Sources of reducing power for the pioneer organism**

CO is a source for reducing equivalents by reaction with H₂O but also with H₂S:

\[
\begin{align*}
(Mg, Ca)\text{SiO}_4 + H_2O &\rightarrow (Mg, Ca)(OH)_2 + (Mg, Ca)\text{SiO}_3 \\
(Fe, Ni)\text{SiO}_4 + H_2O &\rightarrow (Fe, Ni)(OH)_2 + (Mg, Ca)\text{SiO}_3
\end{align*}
\]

Inorganic nutrient interconversions in the surfaces of the inorganic substructure

Ammonia formation by nitrogen fixation under moderate temperature and pressure has been demonstrated unequivocally with ¹⁵N₂ in conjunction with reaction (7) as an electron source (10). It has been suggested as a precursor of extant nitrogenase activity:

\[
N_2 + 3 FeS + 3 H_2 \rightarrow 3 FeS_2 + 2 NH_3
\]

Hydrogen cyanide (HCN), the nitrogen analog of CO, forms by catalytic conversion of CO with NH₃:

\[
CO + NH_3 \rightarrow HCN + H_2O
\]

and it may be trapped as ligand by transition metals under conditions of moderate temperature and pressure (5). In the presence of catalytic NiS or FeS, COS, produced by reaction (5) (11), or from CO₂ by equilibration with H₂S (12), respectively, is reduced via Ni-CH₃ to methanethiol, which also becomes a ligand:

\[
COS + 6H^+ + 6(e^-) \rightarrow CH_3SH + H_2O
\]

α-Hydroxy acids and α-amino acids

Amino acids are among the most important products of extant intermediary metabolism. Therefore, it is of importance that volcanic conditions have been found to generate a suite of α-amino acids and a corresponding suite of α-hydroxy acids at 1 bar CO and around 100°C with (Mg, Ca)(OH)₂ as pH buffer and with an Ni-precipitate with CN-ligands. The CN-ligands serve as C-source and N-source. CO serves as reducing agent. The products satisfy the formula R–CHA–COOH, whereby A

...

The above experimental results suggest that among the most prominent pioneer amino acids of life were glycine \((R=H)\), alanine \((R=CH_3)\), serine \((R=HO-CH_2)\), and valine \((R=CH_2-CH)\). Moreover, the facile formation of serine suggests that cysteine originated early on by sulfidization of serine.

\(\alpha\)-keto acids and \(\alpha\)-imino acids

In the above reactions under modest conditions, traces of pyruvate have been found (13), which indicates that \(\alpha\)-keto or \(\alpha\)-imino acids are intermediates in the pathways to \(\alpha\)-hydroxy or \(\alpha\)-amino acids. The reductive amination of \(\alpha\)-keto acids has been demonstrated experimentally under alkaline conditions with \(\text{FeS/H}_2\text{S}\) or \(\text{Fe(OH)}_2\) as catalyst and reducing agent (14). The formation of significant amounts of pyruvate from \(\text{CO}\) and \(\text{FeS/nonylmercaptan} \at 2000\text{ bar}\) has been reported (15). Remarkably, pyruvate is stable under these conditions and apparently not reduced to lactate.

Sugar alcohols and sugars

In the above reactions, significant amounts of ethylene glycol have been detected (13), which is the simplest sugar alcohol and may form by reduction of glycolaldehyde, the simplest sugar. It is well established that sugars and sugar alcohols are excellent ligands for transition metals, which suggests that we may trace the biosynthetic roots of nucleic acids down to the pioneer metabolism.

Acetyl-thioester

Experiments provide evidence that in the presence of \(\text{Ni(OH)}_2\), \(\text{FeNiS}\), \(\text{NiS}\), or \(\text{CoS}\), the volcanic reactants \(\text{CO}\) and \(\text{H}_2\text{S}\) (or \(\text{CO}\) and \(\text{CH}_3\text{SH}\)) are converted to activated forms of acetic acid \((\text{CH}_3\text{CO-Ni}, \text{CH}_3\text{CO-SNi}, \text{CH}_3\text{CO-SH}, \text{CH}_3\text{CO-S-CH}_3)\) as evolutionary precursor of acetyl-CoA (16):

\[
\begin{align*}
\text{CH}_3\text{Ni} + \text{CO} & \rightarrow \text{CH}_3\text{CO-Ni} \quad (13) \\
\text{CH}_3\text{CO-Ni} + \text{H}_2\text{S} & \rightarrow \text{CH}_3\text{CO-SNi} + (\text{H-Ni}) \quad (14) \\
\text{CH}_3\text{CO-Ni} + \text{CH}_3\text{SH} & \rightarrow \text{CH}_3\text{CO-S-CH}_3 + (\text{H-Ni}) \quad (15)
\end{align*}
\]

Activated amino acids and peptides

The condensation of \(\alpha\)-amino acids to peptides in dilute aqueous systems is endergonic and requires energy coupling. Under alkaline, volcanic conditions \(\text{CO}\) in the presence of \(\text{H}_2\text{S}\) (or \(\text{CH}_3\text{SH}\)) and \((\text{NiFeS})\) is an efficient energy source for the formation of peptides, notably \(\text{CO}\) and \(\text{CH}_3\text{SH}\). The latter operates as a short-lived coupling intermediate, and therefore, the hydrolysis problem associated with theories that require accumulation and/or transport of \(\text{CO}\) is obviated. Interestingly, the addition of \(\text{Na}_2\text{MgPO}_4\) to the system \((\text{FeNiS})/\text{CH}_3\text{SH}\) broadens the \(\text{pH}\)-range of efficient peptide synthesis (16), which indicates nucleophilic catalysis with a phosphorylated intermediate.

As a mechanism of peptide formation, it is suggested that a \(\text{CO}\)-derived activated intermediate \((\text{e.g., COS})\) is formed, which subsequently suffers a nucleophilic attack by the free amino group of an amino acid \((\text{aa})\) followed by ring closure to activate the carboxyl group of the amino acid in a five-membered cyclic intermediate, notably an amino acid \(N\)-carboxyanhydride \((\text{aa}*)\) (16):

\[
\begin{align*}
\text{CO} + \text{H}_2\text{N-CHR-CO-H} & \rightarrow \text{H}_2\text{O} + \text{H}_2\text{N-CHR-CO-NH-CHR-COOH} \\
\text{H}_2\text{N-CHR-CO-NH-CHR-COOH} & \rightarrow \text{CO}_2 + \text{H}_2\text{N-CHR-CO-H}
\end{align*}
\]

Subsequently the free amino group of another amino acid molecule \((\text{aa})\) reacts nucleophilically with the activated amino acid to generate a dipeptide \((\text{aa}-\text{aa})\):

\[
\begin{align*}
\text{R} - \text{H} & \rightarrow \text{R} - \text{H} + \text{H}_2\text{N-CHR-CO-H} \\
\text{H}_2\text{N-CHR-CO-H} & \rightarrow \text{R} - \text{H} + \text{H}_2\text{N-CHR-CO-NH-CHR-COOH}
\end{align*}
\]

Alternatively the free amino group of a peptide may react with the activated amino acid \((\text{aa}*)\), with the result that the peptide is extended by one amino acid unit at its N-terminal end. The reduct energy of \(\text{CO}\) is converted into group activation energy, which then drives the endergonic peptide synthesis pathway. The synthetic pathway is controlled kinetically, and it runs as long as the energy source lasts (16).

Peptide cycle

Surprisingly, it was found that peptide-forming conditions actually support a peptide cycle (18), whereby the peptides react with \(\text{CO}\) to acquire an N-terminal hydantoin ring, which hydrolyzes to an N-terminal urea group. Finally, the urea group is hydrolyzed. The net result is a removal of the N-terminal amino acid unit. The following scheme shows the simplest case of a dipeptide cycle for glycine \((\text{Gly})\) through activated glycine \((\text{Gly}*)\), glycyl-glycine \((\text{Gly-Gly})\), its hydantoin derivative \((\text{H-Gly})\), and its urea derivative \((\text{U-Gly})\):

\[
\begin{align*}
\text{R} - \text{H} & \rightarrow \text{R} - \text{H} + \text{H}_2\text{N-CHR-CO-H} \\
\text{H}_2\text{N-CHR-CO-H} & \rightarrow \text{R} - \text{H} + \text{H}_2\text{N-CHR-CO-NH-CHR-COOH}
\end{align*}
\]

Similar cycles pass through the tripeptide, tetrapeptide, and so on, which results in a concatenation of peptide cycles. The
segment of the cycle up to the urea derivative is anabolic. The hydrolysis of the urea group constitutes a catabolic segment. Because of this anabolic-catabolic character, the peptide cycle generates a dynamic library of peptides, their hydantoin derivatives, and their urea derivatives (“peptide library”). The members of the peptide library come and go, and all transient members of this library are candidates for transition metal ligands. With an increase of the number of amino acids, the peptide library increases exponentially, and so does the likelihood of autocatalytic feedback. It has been discovered that homochiral amino acids undergo slow racemization under the conditions of the peptide cycle, which may be attributed to the hydantoin stage. Racemic amino acids are actually an advantage from the point of view of ligand feedback, because racemic amino acids generate a greater variety of peptides than homochiral amino acids. Remarkably, the dynamic peptide library is also self-selecting because differential bonding of its members as ligands to transition metal centers causes differential stabilization against hydrolysis. It means a self-selection of stable metallo-peptide structures.

The peptide cycle is driven by energy coupling from the redox energy of CO to group activation. The peptide cycle as a whole may be viewed as a catalytic cycle for the conversion of CO to CO$_2$. It may therefore be viewed as a functional evolutionary precursor of extant carbon monoxide dehydrogenase. The Ni-catalyzed hydrolysis of the urea group is the functional evolutionary precursor of the extant Ni-enzyme urease. It is also related to the imidazol ring of the purine bases, which suggests a synthetic as well as an evolutionary precursor relationship.

In the course of evolution of the metabolism by peptide feedback, the system will sooner or later become homochiral because of symmetry breaking by autocatalytic feedback. At this level it has been shown that α-amino acids undergo a facile enantioselective conversion to thermostable β-p-sheets (19).

Energy metabolism of the pioneer organism

In extant metabolism, it is expedient to distinguish an “energy metabolism” that generates a pool of energetic products, like ATP, for energy coupling with endergonic reactions. The initial chemical energy of a volcanic setting is redox energy, and it is surprising that the pioneer metabolism is engaged from the start in various forms of energy conversion and energy coupling. A pool of redox energy originates, if the reducing equivalents of the oxidative half reactions (4) and (6)-(9) are transferred to the transition metal centers in the inorganic substructure, and subsequently, they may undergo lateral transfer through suitably spaced transition metal centers in the substructure. Such a redox pool would form the basis for redox-to-redox energy coupling between exergonic redox reactions and endergonic redox reactions, notably as the beginning of chemiosmotic energy coupling.

Turning finally to redox-to-condensation energy coupling, the simplest case is shown in reaction (8). COS in turn will be the substrate for various forms of condensation-to-condensation energy couplings. For example, C–C-bond formation by carboxylation with COS will generate thioamides (3). Similarly, carboxylation of α-amino acids by COS generate thioamides en route to aminocarbonyl N-carboxyhydrazides:

$$\text{H}_2\text{N}-\text{CHR}-\text{COOH} + \text{COS} \rightarrow \text{H}_5\text{S}-\text{NH}-\text{CHR}-\text{COOH} \quad (16)$$

Cascades of condensation-to-condensation energy coupling are among the hallmarks of extant metabolism. They proceed through compounds of decreasing hydrolytic energy and usually increasing reaction selectivity and constitute a form of chemical energy conservation. Arguably the most important forms of energy coupling and energy conservation involve phosphorylation energy. Volatile phosphoric anhydride P$_4$O$_{10}$ in volcanic exhalations reacts hydrolytically on contact with liquid water through a cascade of oligophosphates (7). Pyrophosphate and other oligophosphates may have provided the group activation for the formation of phosphorylated organic compounds in early evolution (20). It has been proposed by de Duve (see Further Reading) that primordial acetate phosphate originates by reaction of carbon dioxide with phosphate ions. According to the Iron-Sulfur World theory, phosphorylation energy may also result from a transfer of the redox energy of CO/H$_2$S in the presence of amino acids (5), which received its first support by phosphate catalysis of peptide formation (16). Subsequently, it was supported by the discovery of a formation of aminocarbonyl phosphate in reaction of phosphate with aminocarbonyl N-carboxyhydrazide (22):

$$R_2\text{N} + \text{H}_2\text{PO}_4^- \rightarrow \text{H}_2\text{N}-\text{CHR}-\text{COOH} + \text{PO}_4^{3-}$$

and by the discovery of a transfer of the condensation energy of COS to aminocarbonyl phosphate or aminocarbonyl adenylate and further to pyrophosphate (22). The above-reported chemical reactions proceed under conditions that are compatible with an origin of life under the locally and temporally coherent conditions of a volcanic flow system. Therefore, the discovered reactions may well be components of the metabolic system of the pioneer organism. As additional components come into experimental view, the theory is expected to evolve. So far we have addressed the notions of growth and reproduction as aspects of one unitary chemical system. We now show that this unitary system is also the physical basis for the earliest mechanism of evolution and that it constitutes in fact the evolutionary “Anlage” for the emergence of the cellular and genetic features of extant forms of life.
Evolution from the Vantage Point of the Pioneer Organism

In extant forms of life, we distinguish a (fast) process of development over the lifetime of an organism from a (slow) process of evolution over many generations of reproduction. Going back in time we see these two processes fuse into one unitary process, which constitutes at the same time growth by the accumulation of carbon fixation products, reproduction by positive ligand feedback of some products, inheritance by rapid multiplication and lateral spreading of ligands with positive feedback, and metabolic evolution by variation of the set of ligands and by pathway extension. We now treat metabolic evolution in greater detail.

Ligand evolution

Let us assume that a member of the set of products of the pioneer metabolism binds as a ligand to a catalytic metal center and thereby increases the catalytic activity for a rate-determining step in the pioneer metabolism, which in turn increases the steady-state activities of all products downstream from said rate-determining step. This process induces additional ligand feedback effects and so forth, whereby a cascade of ligand feedback effects comes into play. It has the important consequence that the pioneer metabolism is self-expanding because of an avalanche of ligand feedback effects.

Under chromatographic flow conditions, strong bonding ligands have a long retention time and travel slowly in flow direction. Therefore, the strongest ligands (and the best candidates for ligand feedback) tend to become concentrated in upstream flow zones. These ligand zones are areas of ligand feedback concentration and thus constitute the spatial precondition for a local metabolic evolution. This catalyst evolution by ligand feedback is the most basic mechanism of evolution and the source mechanism for the emergence of a mechanism of pathway evolution to which we turn next.

Metabolic self-expansion

The evolutionary increase of catalytic activities and the concomitant widening of the spectrum of catalytic abilities elicit ever more complex expansions of the metabolism. These expansions may be from terminal extension or lateral branching of preexistent pathways. They lead to new synthetic products of the metabolism, which increase the set of potential ligands for the catalysts. Therefore, the catalysts evolve from strictly inorganic catalytic transition metal centers with inorganic ligands to hybrid inorganic–organic catalyst centers with a combination of inorganic and organic ligands. Within the latter category of catalysts, metallo-peptides become more and more dominant. Finally, with the emergence of the genetic machinery, primitive noncoded metallo-peptides are replaced by coded metallo-peptides and ultimately by folded metallo-proteins.

So far we have considered an evolving pioneer metabolism based on the full complement of primordial inorganic constituents of volcanic flow localities. With the expansion of the metabolism, new pathways come into play, which are independent of one or the other inorganic compound. This process allows the descendents of the pioneer organism to venture into spaces that are devoid of such inorganic compounds. For example, the coexistence of amino acid synthesis from CN-ligands and from CO/NH2-ligands allows the gradual colonization of spaces devoid of cy-anide sources. Here the amino acid synthesis becomes restricted to CN-ligands as carbon source and indispensable CN-ligands (in hydrogenases) come to be biosynthesized. In subsequently conquered realms, the CO-ligands become replaced by carbon dioxide as carbon source and by FeS/H2S or H2 as reductant. In new chemical spaces, the catalysts that are dependent on unavailable or extremely depleted inorganic starting materials will be opportunistically lost. Therefore, newly conquered spaces turn from optional habitats into obligatory habitats and metabolic evolution leads by necessity to a biosphere with diverse variants in a diversity of habitats.

In the course of early evolution, the metabolism becomes more and more integrated and centralized, so much so that a chemical conversion of highly integrated constituents tends to weaken the metabolism. The more central the constituent, the more severe is this effect of metabolic decay by chemical conversions, which leads us to yet another strategy of metabolic evolution: the strategy of dual feedback.

Metabolic evolution by dual feedback

So far we have looked at early evolution from the point of view of ligand function for catalyst promotion. Now, we look at it from the point of view of ligand synthesis. Let us assume that a fluctuation in the chemical environment expands the metabolism by the synthesis of a new ligand with new positive feedback into the metabolism from a preexistent ligand. However, if the chemical environment returns to the original state, the new ligand will disappear again, unless it exhibits also a positive feedback into its own synthesis. More generally speaking, duality of feedback of new metabolic products (atrociously into the metabolism and egotistically into their own branch pathways) seems to be essential for metabolic evolution. Extant cellular organisms are replete with dual feedback catalysts. Ribosomes produce all proteins, including ribosomal proteins. Coenzymes catalyze many synthetic pathways, including coenzyme biosynthesis. Protein translocases transport many proteins, including translocase proteins. DNA codes for all genes, including DNA polymerase genes.

Autocatalytic metabolic cycles

According to a general rule of organic chemistry, reactions involving the smallest molecules are catalytically the most restrictive. This rule holds notably for the build-up of carbon skeletons with the arithmetic C1 + C1 = C2 (e.g., C2 = gly-cine or acetyl thio-ester). Therefore, it may not come as a surprise that in the course of metabolic evolution, these most "simple" carbon fixation reactions may fail by the wayside. Under these conditions, an autocatalytic carbon fixation metabolism can only be maintained by a "metabolic cycle," which multiplies the C2 unit autocatalytically in the absence of its de novo synthesis. A prominent example is the reductive citric acid cycle (C2 + C1 =
for a tremendous expansion of the spaces inhabitable by life.}

**Thermal evolution of life**

By a reasoning that goes back to Thomas Brock, evolution can proceed only thermally downward from extremely thermophilic organisms to mesophiles by opportunistic losses of individual thermal stabilities. Such a view of evolution is compatible with an origin of metallo-peptide-protein folding structures with extremely thermostable covalent transition metal-sulfur cross-links, many of which were later replaced opportunistically and irreversibly by the cooperation of a multitude of weak, noncovalent group interactions in concert with an increase of fidelity of the emergent genetic machinery (4, 5).

**Genetization and enzymatization**

According to the RNA World theory, life begins with RNA replication, but the replicated RNA has no function other than its replication. According to the Iron-Sulfur World theory, the early novelties in the pioneer metabolism develop with the catalyst-promoting function of transition metal ligands. Therefore, early on, sugars, nucleosides, and (oligo)nucleotides must have evolved as ligands. Later oligonucleotides folded and were associated by base pairing, which ushered in a new type of catalysis: base pair-assisted positioning of RNA-bonded amino acids for the formation of peptides. This catalysis was of immediate benefit because the function of peptides as ligands was preordained in the universal laws of physics and chemistry. Remarkably, up to the level of the origin of speciation, the iron-sulfur cluster was the only noncovalent group interaction between molecules that is chemically restricted, preordained pathway possibilities. It is this association of chemical necessity and genetic chance that preordained the universal laws of physics and chemistry.

**Cellularization**

The emergence of any multicomponent genetic machinery depends on the existence of compartments. It has been suggested that the chemo-autotrophic origin of life occurred in primordial compartments that were bounded by FeS-membranes. The proposal was based on experimental FeS-precipitations (22). However, the detected cell structures are likely an artifact of sample preparation by freeze-drying of an FeS-hydrogel. Freeze-drying usually produces a porous product, in which the rate of freezing determines the size of the ice crystals, which in turn determines the size of the pores (5, 6). Instead of such chemically unsupported FeS-cells, it is proposed that a cellularization of life occurred step by step with autoformaly generated lipids (2). The first lipids may have been long-chain ω-hydroxy and ω-amino acids (6) to be replaced later by products of Classen condensation of acetyl thioesters fatty and isoprenoid acids (4). The first lipid function must have been lipophilization of the substructure by lipid accumulation. This process had the effect of lowering the H₂O and H₃O⁺ activities near the surface, thereby disfavoring hydrolytic reactions and favoring all kinds of condensation reactions. Because lipid syntheses from acetyl thioesters are condensation reactions, surface lipophilization is seen as collectively autocatalytic. The accumulating lipid molecules will inevitably organize into mineral-supported bilayer membrane patches, which ultimately will coalesce into an extended membrane that covers the mineral surface or surrounds a mineral particle (2).

Semicellular structures, which are bounded partly by a mineral surface and partly by a lipid membrane, form automatically for energetic reasons, when a lipid membrane spans over a small cavity in the mineral surface. The volcanic nutrients can pass freely through the membrane, whereas detached organic constituents are retained under the membrane. Now multicomponent systems can emerge within the incipient cytosol underneath the membrane along with the step-by-step enzymatization of the metabolism. Finally, self-supporting membrane envelopes appear as a result of the wedgeshaped transition metal-sulfur-phospho-glycerol lipids. These lipids are chiral, but they must have appeared first as a racemate. The self-supporting lipid membrane formed the basis for chemiosmosis by harnessing all kinds of environmental redox energy. Ultimately, the lipid racemate underwent chiral symmetry breaking by giving rise to the divergence of the domain Bacteria with one lipid enantiomer and the domain Archaea with the other lipid enantiomer (5, 6). Remarkably, up to the level of the origin of speciation, decisive events in the early evolution of life seem to have been preordained in the universal laws of physics and chemistry.

**Concluding remarks**

From the vantage point of a chemo-autotrophic origin of life, the overall mechanism of evolution is seen as a composite affair. It comprises a primal, direct, chemically deterministic, directional mechanism of evolution and a later, indirect, genetically stochastic mechanism of evolution. The latter is a derivative of and complementary to the primal mechanism of evolution. With the indirect mechanism of evolution appears the need for Darwinian selection, which gives rise to adaptations to the outer conditions of the environment and to the inner conditions of chemically restricted, preordained pathway possibilities. It is this association of chemical necessity and genetic chance that generates biochemical continuity between the origin of life and the extant biosphere. In fact the Iron-Sulfur World theory is not just a theory on the origin of life. It is also an explanatory theory of biochemistry.
Acknowledgments

Support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged. My gratitude goes to Claudia Huber and Wolfgang Eisenreich for seminal experiments and fruitful discussions, and to Adelbert Bacher for generously supplying laboratory facilities and material support.

References


Further Reading


See Also

Origins of Life: Emergence of Amino Acids
Origins of Life: Emergence of an RNA World
Origins of Life: Emergence of Nucleic Acids
Prebiotic Membrane Formation
Transition Metals in Biology
Origins of Life: Emergence of Amino Acids

Robert Pascal* and Laurent Boiteau, UMR5247, CNRS–University of Montpellier 1 and University of Montpellier 2, Montpellier, France
doi: 10.1002/9780470048672.wecb423

Amino acids can be synthesized from inanimate matter through processes independent of living organisms. An overview of these processes occurring in various environments, including interstellar media, is given with a special emphasis on Strecker synthesis and related processes compatible with primitive earth conditions. When analyzing their relevance to the origin of life, chemical pathways leading to α-amino acids from activated precursors may be considered independent from further polymerization into peptides promoted by activating agents. Alternatively, both processes may also be connected to each other so that the emergence of amino acids and the prebiotic formation of peptide may have been closely associated. The last possibility is that prebiotic chemistries of amino acids and nucleotides may have been cooperating with each other in a way that led to the emergence of the genetic code and to a rudimentary translation apparatus.

The emergence of life has become a scientific field of interest since Darwinian theory was introduced, presuming the evolution of living organisms. Strong evidence exists that all living beings on Earth have a common ancestor. However, this Last Common Ancestor (LCA) (1) was most probably a living entity involving nucleic acid-based information storage (RNA and probably DNA) and ribosomal translation machinery. It lived probably several hundred millions years, and possibly more than one billion years, after life originated. We have no clues to the evolution pathway that led from the emergence of life to this LCA and on the preceding chemical evolution. These processes are the field of investigation of researchers in prebiotic chemistry and in early evolution; they try to build self-consistent hypotheses for the development of life. Some information can be gained from the analysis of biochemical pathways, keeping in mind that new pathways may have evolved replacing extinct ones. Other indications can be obtained by reproducing the conditions assumed to have occurred on the early Earth (the panspermia hypothesis will not be considered here). Indeed, the first scientists who analyzed the origin of life process (2, 3) proposed that early life forms were dependent on abiotically formed organic molecules as carbon and energy sources for their growth (the heterotrophic hypothesis). The formation of amino acids by electric discharges in reducing mixtures of gases (CH₄, NH₃, H₂O) through the Miller experiment (4) has given experimental support to this hypothesis by demonstrating that some essential building blocks of life can be efficiently formed abiotically. The presence of amino acids in several extraterrestrial samples of matter delivered to the Earth as meteorites is also clearly established. Amino acids, and more generally organic matter, must not be considered rare in the universe, which is also independently supported by the spectroscopic analysis of interstellar molecular clouds. In the following sections, we will describe different processes leading to α-amino acids under potentially prebiotic conditions, as well as analyze their relevance to the origin of life and their connection with the role of amino acids at the early stages of life evolution.

α-Amino Acids and the Origin of Life Process

Proteins are one of the main classes of biopolymers found in living organisms. They can adopt a large variety of structural patterns, which makes them essential cell constituents mostly involved in recognition processes. Their molecular recognition ability is remarkable in catalysis because enzymes have evolved to accommodate transition states in their active site to lower the kinetic barriers of most biochemical reactions. This ability is connected to appropriate primary sequences, and it allows the proteins to fold into well-defined tertiary structures. However, the accuracy of the primary sequence, which is ensured by the translation of genomic sequences in
current-day living organisms, seems out of reach of prebiotic systems. Therefore, the actual role of amino acids at that time is not clearly known. Nevertheless, if we still assume that their prebiotic role was the catalytic activity of polypeptides, another unresolved question lies in the prebiotic pathway of synthesis of these polymers. It is believed that prebiotic peptides have been formed from monomeric $\alpha$-amino acid through physico- or chemically induced processes. This assumption is not supported by strong evidence and alternative processes may have taken place. From a more general point of view, the biochemical route to biopolymers represents one kind of disconnection among the various possible retrosynthetic pathways that can be conceived starting from highly activated simple organic compounds (5). In this view, the stepwise formation of biopolymers by adding monomeric building blocks using condensing (dehydrating) agents in (most probably) diluted aqueous solution seems unlikely compared with a self-assembly that could take place directly from activated simple organic precursors with no need of further activation. As a result, the current discussion of the emergence of $\alpha$-amino acids also includes the possibility that peptides may have been formed directly from activated precursors through processes bypassing monomeric $\alpha$-amino acids.

**An amino acid and peptide world?**

In biology, the catalytic role is mostly played by enzymes. It is tempting to speculate that life may have emerged in a peptide world (6). Short peptides and even isolated $\alpha$-amino acids are capable of catalyzing prebiotic processes (7). Various physical or chemical pathways are available for the formation of peptides from inactivated or activated monomers. However, with respect to the sequence of peptides derived from common $\alpha$-amino acids, this kind of chemistry still fails to provide an efficient replication system. For instance, peptide segments (activated as thioesters) can be self-replicated by ligation if selected residues (cysteine) are present at the N-terminus (8, 9), but the replication of a complete peptide sequence from monomers seems unattainable. The only remaining possibility is an early system of information storage based on unnatural amino acids that is capable of base pairing or related molecules such as peptide nucleic acids (PNA). Although a transfer of information from PNA to RNA is materially feasible (10), it is unclear why this system could have been replaced by nucleic acids rather than being improved by evolution. Moreover, no remnant of this system has been preserved in current-day living organisms.

**The role of $\alpha$-amino acids in the RNA world scenario**

The RNA-related nature of most coenzymes and cofactors, the role of RNA in the translation process, and the catalytic abilities of RNA show that RNAs have played a crucial role in early biochemistry and have supported the hypothesis of an RNA world (11). In this scenario, living organisms primarily used RNA for both catalysis and information storage before the advent of coded peptides and of the translation apparatus. But most supporters of this scenario believe that amino acids and peptides may have played a part in catalytic activity, for instance as ribozyme cofactors (12). This provides a rationale for the evolution of the genetic code. In any case, covalent interactions of RNA with amino acids and peptides were probably required early to initiate the development of translation, so that hypotheses considering that polypeptides cooperated with RNAs since the early beginning of the RNA world have been presented (13, 14).

**A coevolution process toward the genetic code**

Because the translation apparatus is among the oldest components of living beings, it is tempting to consider the emergence of life and of translation to be closely related. In other words, we can make the hypothesis that nucleic acids and translated peptides may not be considered as two independent subsystems but as one subsystem, the RNA-coded peptides (5). Experimentally, reactions starting from $\alpha$-amino acids activated as N-carboxyamides, and nucleotides have recently demonstrated the possibility of obtaining covalent conjugates (15) under prebiotically relevant conditions. Similarly, reactions of N-phosphoryl $\alpha$-amino acids can result in both oligopeptides and oligonucleotides (16). The chemistry involved in such systems may have served as a starting point for the evolution toward the translation apparatus. Subsequently we must determine how the current-day set of coded residues has been selected among many possible structures and how they have been allocated a triplet codon to allow the translation of the genetic information into protein sequences. The coevolution theory of the genetic code (17, 18) that may also be independent of the coevolution process mentioned above) states that codons have initially been selected for a subset of prebiotically synthesized $\alpha$-amino acids. Then codon allocation was guided by potential biochemical conversions between the amino acids.

**Abiotic and Prebiotic Chemistry of Amino Acids**

Amino acids have probably formed in different places of the Solar system through processes independent of living organisms as shown by their occurrence in carbonaceous chondrites, a particular class of meteorites. Among these processes, only a particular subset of them can be considered to be truly prebiotic. It is the processes that make amino acids available in environments capable of developing a complex chemistry compatible, or at least presumed to be compatible, with the emergence of life.

**Formation on the early earth**

Several processes may have been responsible for the local formation of amino acids. In fact, energy release (electrical discharges, ultraviolet light, or impact shocks) into gas mixtures containing elementary carbon sources (CO$_2$, CO, CH$_4$), water and nitrogen produces organic molecules, including amino acids, but significant yields require a reduced overall redox state.
Figure 1: Strecker and related systems: complex network of equilibrated reactions from aqueous solutions of cyanide, ammonia, and aldehyde. Pathways for exiting the equilibrated network correspond to (a) cyanohydrin hydrolysis, (b) aminonitrile hydrolysis, and (c) Bucherer–Bergs reaction. Adducts 6 and 7 and other hydrolysis products can be formed at high concentrations of reactants.

This requirement is fulfilled for electric discharges in a reduced atmosphere containing methane, ammonia, and water, as in the original Miller experiment. It has also been observed for atmospheres based on N₂ and CO or CO₂ on the condition that H₂ or methane is also present in sufficient amounts (19). A neutral atmosphere (based on N₂, CO₂, and water) would produce much lower yields of organics (by several orders of magnitude). In the absence of other species to be oxidized, the reduction of CO₂ requires the concomitant thermodynamically unfavorable conversion of water into O₂ (as in photosynthesis). However, even if the atmosphere was neutral when life arose, as usually believed, the Earth was not uniform with respect to redox state simply because the reduced state of the mantle and the high volcanic activity favored the occurrence of locally reduced environments (for instance, in hydrothermal vents in the oceans). Then, a preservation of the hydrogen content of the early atmosphere or the diversity of environments on the early Earth is likely to have made amino acid formation possible, at least at specific places.

The strecker synthesis and related processes

The formation of α-aminonitriles: a set of equilibrated reactions

A strong argument in favor of the involvement of Strecker synthesis is that aldehydes and HCN are among the initial species formed in electric discharge experiments leading to α-amino acid formation. Mechanistic and thermodynamic studies (20) of systems derived from aqueous mixtures containing an aldehyde, ammonia, and hydrogen cyanide have disclosed the network of equilibrated reactions involved in these systems (Fig. 1). In the first stage, cyanide is rapidly added to the carbonyl compound and cyanohydrin 2 is produced. Then, the α-aminonitrile 5 is formed more slowly through cyanide attack on an iminium ion intermediate 4. The composition of the equilibrium mixture formed from equimolecular concentrations of aldehyde and cyanide is strongly dependent on the pH and of the initial concentration of ammonia. Then, under conditions simulating a presumed prebiotic aqueous environment (with low concentrations of organic precursors and ammonia and in a 5–7 pH range brought about by a CO₂-rich early Earth atmosphere), α-aminonitriles are formed in very low equilibrium yield (21). However, since hydrolysis pathways (Fig. 1) are likely to proceed at different rates, there is no reason for the final hydroxy acid/amino acid ratio to match the large predominance of cyanohydrin over α-aminonitrile at equilibrium. Actually, an important, often overlooked, feature of the Strecker synthesis is that, under neutral or mild alkaline conditions, α-aminonitrile hydrolysis (or conversion into stable derivatives) proceeds through efficient pathways involving the specific participation of the neighboring amino group, which offsets their low thermodynamic stability.

Catalytic conversion of α-aminonitriles: the determining process

It turned out that α-aminonitrile hydration is subject to efficient catalysis by aldehydes (22), which are already reactants of the Strecker reaction (Fig. 2). This catalytic pathway is still prevailing at 20-μM formaldehyde concentrations (21) and...
half-lives shorter than a year are expected in the presence of 10-mM formaldehyde at pH 7 and 25°C. It proceeds through the addition of the aldehyde at the amino group and the subsequent intramolecular reaction of carbinolamine adduct 13.

**Bücherer-Bergs reaction**

This reaction (Fig. 2) is closely related to the above described catalytic pathway for α-aminonitrile hydration, but it involves CO₂ instead of an aldehyde (23). The main difference lies in the fate of isocyanate 19 (formed instead of imine 15), which easily undergoes the cyclization into hydantoin derivative 12 (instead of hydrolysis to α-aminamide 18). Considering the kinetic stability of 12, CO₂ is a reactant instead of a catalyst for the Bücherer-Bergs reaction; nevertheless, this pathway is an additional, very efficient process for selecting α-amino acid derivatives rather than those of α-hydroxy acids. Moreover, despite the very sluggish reactions of its products, the Bücherer-Bergs reaction must not be considered as a dead-end since the N-carbamoylamino acid 20 can be converted into the α-amino acid N-carboxyanhydride 21 (Fig. 3), readily under the effect of nitrogen oxides (24), but also slowly through a spontaneous pathway (25, 26).

**Hydrothermal formation**

The synthesis of amino acids in aqueous media is energetically feasible in reducing fluids and temperatures found in hydrothermal systems (27). Amino acids have indeed been produced, although in micromolar concentrations, by heating a mixture of methane and nitrogen in simulated hydrothermal fluids (28). A likely process for their production could be a Strecker reaction. Plausible chemical pathways have been thermodynamically analyzed (29), and a process involving reductive amination of keto acids (30) has also been proposed. On the other hand, high temperatures are also likely to damage organic molecules.
Origins of Life: Emergence of Amino Acids

**Formation in interstellar media**

Another potential source of organic compounds on early Earth is extraterrestrial bodies (31–33). This possibility is attested by the presence of α-amino acids in meteorites. The synthesis of organic compounds occurred in the presolar molecular cloud or in the protoplanetary disk that preceded the formation of planets (34). It is likely to have resulted from UV-irradiation of interstellar ices (35) with the formation of radicals and/or other highly activated species, which then randomly combined into more complex organic compounds at the low temperature of the molecular cloud. Then these ice particles were involved in the accretion of meteorite parent bodies, where organic matter could be preserved from the heat released by the process in small accreting bodies. Subsequently, reactions like those involved in Miller’s experiment may have occurred on meteorite parent bodies early in the history of the solar system. Then, these organic molecules could have been delivered to the Earth through impacts of meteorites, micrometeorites, and comets.

**Peptide formation**

Many processes have been proposed for the formation of prebiotic peptides. Peptide bond formation from free amino acids can become thermodynamically favorable using physical or chemical means for dehydration (6, 36). Alternatively, activated amino acid derivatives are capable of polymerizing into oligopeptides in aqueous solution. A pathway for the formation of α-amino acid thioesters starting from sugar precursors has been discussed (37). α-Amino acid N-carboxyanhydrides (NCAs) correspond to the most activated form of amino acids reachable in a prebiotic environment with a high content in CO$_2$ or bicarbonate (38), since carboxyl-activated derivatives are easily converted into NCAs through a very efficient CO$_2$-promoted process (39) (Fig. 3). The length of peptide chains formed by polymerization of NCAs is often limited by precipitation, but longer peptides (up to 55-mers) have been obtained in the presence of illite and hydroxylapatite (40). NCAs can also be formed through prebiotically relevant pathways starting from N-carbamoylamino acids (24–26) or through the activation of amino acids with carbonyl sulfide (41).

**Amino acids and the emergence of homochirality**

The configuration of natural amino acids has led to studies on the possibility that homochirality emerged at a prebiotic stage, which may be supported by the presence of amino acids as non-racemic mixtures in meteorites (32). This enantiomeric excess may have resulted from the exposition of extraterrestrial matter to circularly polarized light (42). Whatever the origin of this enantiomeric excess, it may have initiated stereo-selective processes through different catalytic pathways (7). Symmetry breaking may also have resulted from reactivity in connection with other processes such as crystallization or interfacial chemistry (43, 44) and polymerization of amino acids (45).
Relevance of Abiotic Processes to the Emergence of Life

The presence of amino acids in meteorites (chondrites) and their delivery on the surface of the Earth may have triggered prebiotic pathways starting from these building blocks. On the other hand, their occurrence of meteorite parent bodies may simply be considered as a manifestation of their easy synthesis through various abiotic pathways, and of their abundance in the universe and consequently on the early Earth. The answer to this alternative idea mainly depends on the rate of evolution of the atmosphere and on the development of conditions favorable to life on the planet. Most scientists consider that the atmosphere of the Earth was initially based on a mixture of N₂ and CO₂ with a significant content in H₂ and CH₄ (favorable to the production of organic molecules). Then, it evolved from this reduced state to a more neutral state (N₂, CO₂) as hydrogen escaped to outer space (46). The rate of this evolution depends on both the input of H₂ from the volcanic activity and its loss to outer space (47). Another problem is the presence of conditions that allow the growth of living beings. Studies on the origin of life processes are strongly dependent on the knowledge gained by other fields of science such as astrophysics, geology, and other planetary sciences. Since liquid water was already present on the Earth a few hundred million years after its accretion leading to the early formation of oceans, a major problem encountered by early forms of life would have been that of impact of asteroids or comets capable of vaporizing the oceans (at least in part) and thus of sterilizing the planet. Both issues are still debated, and many hypotheses are acceptable. Life may have emerged early on Earth in a globally reduced environment and may have survived impacts by asteroids or comets. A debatedly, synthetic processes occurring on early Earth with a reduced atmosphere may have no connection with a later emergence of the ancestors of all living beings that we currently know. Instead, it may have resulted from the delivery of organics or their synthesis in locally reduced environments. It is still possible to speculate that life emerged several times and then was annihilated by impacts.

Future Research Directions

More than 175 years after urea was synthesized from mineral precursors, abiotic processes seem to be capable of having made α-amino acids and various organic molecules available in many places of the solar system. The easy abiotic synthesis of α-amino acids may also be the consequence of their composition based on the most abundant reactive elements in the universe (H, C, N, O). Determining the actual pathway through which life emerged is currently unsolved considering the absence of remnants, and we have no indication that this situation may change. However, the study of nonlinear processes of complexification, involving biomolecules (including amino acids) and other prebiotic organic or inorganic derivatives, remains an important field of interest for chemical biology. The discovery of processes working under kinetic control and capable of connecting replication, adaptation and self-maintenance, which is one of the biggest problems in modern science, needs additional analysis of simple and robust pathways. Amino acid prebiotic chemistry, starting from activated species and/or energy sources and potentially capable of delivering energy to such systems, has brought about important pieces of information in this direction and will surely be helpful in the achievement of this goal.

References


Origins of Life: Emergence of the RNA World

Michael P. Robertson, Center for Molecular Biology of RNA, University of California, Santa Cruz
doi: 10.1002/9780470048672.wecb424

The RNA World hypothesis refers to a period during the evolution of life on Earth, preceding the DNA/RNA/protein-based systems of modern biology, during which RNA alone directed and catalyzed the chemistry of life. The concept is an elegant solution to the problem of devolving our intertwined 3-biopolymer biochemistry toward its origins as the simplest system imaginable. In support of this idea, a scattering of molecular clues within our cells point suggestively toward an RNA-only period leading up to the ribozyme-mediated invention of protein synthesis. Indeed, the RNA World hypothesis is a very sensible scenario when considered from the perspective of trying to simplify life’s complexity. The alternative perspective, however, imagining a sterile, primordial Earth and trying to assemble abiotically produced precursors into the RNA World is an extremely difficult problem. Certain steps along the logical chemical pathway to the RNA World have been demonstrated in the laboratory, but many others remain problematic. The exploration of RNA’s catalytic properties has been encouraging, but complex ribozymes are extremely scarce in RNA populations. If the RNA World did once exist, it came about by a mechanism that has not yet been adequately elucidated, including the possibility that a different, unknown, biopolymer system may have preceded the RNA World.

The mechanism of the emergence of life on Earth is one of the most fundamental mysteries of the biological sciences. Despite centuries of pontification and decades of serious experimental efforts, the details remain poorly understood. The most satisfying attempts to deduce the genesis and earliest evolution of primitive life have focused on what is known of modern biology and attempted to simplify it back to the most fundamental and primitive (1, 2). This simplification exercise quickly leads to a paradox in determining which of the essential biopolymers of life, proteins or nucleic acids, would have come first. Conventional logic indicates that it could not have been DNA either, which, although containing all the coded instructions to make copies of itself and the functional abilities to carry out those instructions that the theory gained wide acceptance. The assumption of a so-called ‘RNA World’ era that preceded the biological assimilation of DNA and protein, during which RNA would have been the primary facilitator of a metabolism of unknown complexity, offers an elegant and appealing solution to the challenge of inventing the system of multiple, interrelated biopolymers that exists today. Clues pointing to this bygone RNA World era are scattered throughout our biochemistry. Considered individually, they are not overwhelmingly convincing, but taken together, they make a compelling case that an RNA-based biology might have preceded what we have today. Despite the deserved enthusiasm that the RNA World hypothesis generates, there remain many puzzling difficulties and unanswered questions. An RNA World is quite appealing from the perspective of modern biology looking backwards in time, but is extremely difficult to envision from a perspective at the formation of the Earth looking forward in time. Bits and pieces of the components of RNA can be synthesized, sometimes efficiently, in prebiotic soup-type conditions, but there is not yet a plausible scenario to assemble them into a monomer or dimer of RNA, let alone a polymer of sufficient complexity and functionality to kick-start a living system.
series of mechanisms, including the possibility of a preRNA World, led to the formation of long RNA polymers and eventually ribozymes, the creation of a self-replicating system based solely on RNA, the simplest possible RNA World scenario, is still a daunting challenge that well over a decade of focused experimental attempts have so far failed to solve. Even so, the prospects remain optimistic. Each year brings new insight into the complex characteristics and abilities of RNA, as the boundaries of its capabilities are explored further.

Life, Dogma, and the RNA World

The term ‘RNA World’ (8) refers to a hypothesized period of the early evolution of life, during which RNA, in the absence of DNA or proteins, served as the sole biopolymer of life. Such a life form would likely be unfamiliar to what we are accustomed to, and so a thoughtful, general definition for what constitutes “life” becomes important in order to identify the boundary that separates non-living from what can be considered alive. However, drafting a fundamental, general, succinct definition of the processes and characteristics that represent life has been difficult, controversial, and often unsatisfying (9). The formulation that seems to enjoy the most widespread support is the “chemical Darwinian” definition, which states, simply, that “life is a self-sustained chemical system capable of undergoing Darwinian evolution” (10). The power of this statement is that the primary natural force that has shaped the diversification and evolution of all the organismal life that we know of, namely Darwinian evolution, is also applicable at the simplest, sub-organismal, molecular stages of life. Implicit in the chemical Darwinian model is a system that can sustain itself and create progeny copies. The copies were slightly less than perfect, such that natural selection could encourage a lineage of generations, improving and diversifying, evolving eventually into the entire biosphere. The “central dogma of molecular biology” (11, 12) describes the flow and directionality of information exchange among the three primary information-rich biopolymers of contemporary biology, namely, DNA, RNA, and proteins (Fig. 1). The biochemistry of life today is intertwined between these major, discrete components, each one dependent upon the others for its existence. However, a scenario in which two, let alone three, of these individual systems arose simultaneously and were able to function synergistically is very difficult to imagine. In theory, a system comprised only of RNA could direct and execute the replication of its own genetic information, producing progeny via a self-templated RNA polymerization mechanism catalyzed by the RNA itself, thus satisfying the minimal definition for life with a single biopolymer.

The central component of this RNA World scenario is a replication system that can make copies of the genetic material to grow and produce progeny. The most basic scenario is a simple templated ligation of small oligomers in a cyclical replication scheme (13-16), but ultimately a catalytic entity responsible for this crucial function of replication would be necessary. The simplest example is a single self-replicating ribozyme that can copy itself. More elaborate scenarios for a mature RNA World might also involve ancillary functions such as producing the nucleotides, or building blocks, for the replicase to work with. A case can be made for numerous other functions as well, but one that seems very probable is the invention of protein synthesis. The observation that a nearly protein-free preparation of ribosomal RNA can catalyze the crucial chemical peptidyl transferase step of ribosomal protein synthesis (17), and subsequent X-ray crystal structures of the ribosome showing the arrangement of RNA around the catalytic center (18), together indicate that the cellular machine for making protein is, at its core, a ribozyme. The invention of protein synthesis marks a boundary between the RNA World and the era that persists to this day, in which proteins have become the dominant agents of catalysis across all forms of life.

Prebiotic Synthesis of RNA Components

Nucleotides are the primary building blocks necessary to assemble an RNA world. Each nucleotide is comprised of a base attached to a sugar through an N-glycosylic bond, with a different site on the sugar phosphorylated (Fig. 2). The information content of a nucleic acid polymer is encoded in the identity and order of the bases—adenine, cytosine, guanine, and uracil in RNA, or thymine in place of uracil in DNA. D-ribose, the sugar moiety in RNA, bridges the backbone with 5′-phosphodiester linkages to adjacent nucleotides in addition to attaching to the base via the C1′ position.

Bases

The bases are nitrogenous aromatic heterocycles, bicyclic in the case of the purines, and pyrimidines with only a single ring. Both systems are decorated with complementary arrays of base-pairing interactions, usually between a purine and a pyrimidine, e.g. adenosine (A) pairs with uridine (U), and guanosine (G) with cytidine (C) (19) (Fig. 3). These canonical Watson-Crick pairings form the basis for the storage and copying of double-stranded DNA and the genetic transfer from DNA to RNA, although non-canonical pairings are common in nucleic acids that adopt structures more complex than a simple double helix.

Purines

Demonstrations of the abiotic synthesis of the nucleic acid bases has been a relative success among the steps necessary to understand the synthesis of RNA nucleotides. Beginning in 1953, when the Miller-Urey spark discharge experiments (20) triggered a shift towards experimental testing in origins of life research, the simplest subunits of life have been favorable targets for investigation. In 1960, O’o demonstrated the synthesis of adenine in yields of less than 1% from simple aqueous solutions of ammonium cyanide (21, 22). Stoichiometrically, adenine is a pentamer of HCN, and although
Origins of Life: Emergence of the RNA World

DNA replication

Translation

Protein

Reverse transcription

RNA replication

Figure 1: The central dogma of molecular biology. The central dogma describes the flow of information between the principle biopolymers of the cell. The primary exchange is shown horizontally, left to right, namely, DNA directs the synthesis of RNA, which in turn directs the synthesis of protein. Alternate observed exchanges are depicted off the horizontal with less prominent lines. The letters ‘P’ and ‘R’ indicate that protein or RNA, respectively, are required for particular exchanges.

Figure 2: The components of RNA. (a) The primary bases of RNA, which includes the purines adenine and guanine and the pyrimidines cytosine and uracil. (b) A purine nucleoside (adenosine) and a pyrimidine nucleoside (uridine). (c) The nucleotide adenosine 5’-triphosphate. (d) An RNA strand with the sequence AGCU.

Some uncertainty about the precise reaction pathway exists, the detection of 4-aminoimidazole-5-carboxamidine (AICA) and 4-aminoimidazole-5-carbonitrile (AICN) in the complex reaction mixture, suggests a role for either or both of these molecules as intermediates in alternate pathways to the synthesis of adenine from HCN (Fig. 4) (23, 24). Diaminomaleonitrile (DAMN), a tetramer of HCN and an abundant component of HCN polymerizations, can be converted to AICN through a photochemical rearrangement and may be an important precursor to adenine under certain conditions (25, 26). Alternatively, Ferris and coworkers have suggested that very little or no free adenine is formed in these discreet scenarios, but rather adenine is released only during the hydrolysis of the complex and insoluble cyanide polymers (27). The synthesis of guanine has also been detected under identical conditions at levels 10 to 40-fold less than adenine.
similar to the original Oró synthesis. Purine syntheses proceed through HCN-mediated mechanisms known to form under these conditions (32). It is likely that these were formed in methane/ammonia mixtures (29–31). Since HCN is in low yields in several discharge/irradiation experiments perpetrator. intermediate.

Figure 4 Prebiotic synthesis of adenine. This schematic shows potential pathways and intermediates in the synthesis of adenine from ammonium cyanide.

Pyrimidines A variety of routes for the synthesis of pyrimidines have also been demonstrated. Uracil is formed in high yields by heating malic acid and urea in sulfuric (33) or polyphosphoric acid (34). Although the simple starting materials are prebiotically attractive, the extreme conditions and anhydrous requirement of this reaction make it of questionable relevance for the prebiotic Earth in anything other than highly specialized local environments. A number of other more plausible pyrimidine syntheses have also been explored, but these tend to be much less efficient. Oró reported low levels of uracil synthesis by heating aqueous acrylonitrile with urea and ammonium chloride (35). A variety of pyrimidines, including oxonic acid, 5-hydroxypyrimidine, and 4,5-dihydroxypyrimidine have been detected at low levels in HCN polymerizations (27, 36), as has uracil (37). Schwartz and Chittenden demonstrated the synthesis of 5,6-dihydropyrimidinol (DHU), a constituent of transfer RNA, by condensation of β-alanine and urea through a β-ureidopropionic acid intermediate, which can then cyclize to form DHU (38, 39). Subsequent photolytic dehydration of DHU produces uracil in a reaction promoted by hydrated clay minerals, particularly montmorillonite. Orgel and coworkers produced reasonable amounts of cytosine and its hydrolysis product, uracil, from solutions of cyanocetylene and cyanoate or urea (40). Cyanocetylene is an attractive prebiotic molecule in the sense that it is a reactive compound formed abundantly in spark discharge experiments (41). However, its reactivity makes cyanocetylene short-lived in the environment, the chief byproduct being the hydrolysis product cyanacetaldehyde. Ferris combined cyanacetaldehyde with guanidine to form 2,4-diaminopyrimidine, which can in turn hydrolyze to make cytosine, and then uracil (42). Nearly 30 years after Orgel’s experiments with cyanocetylene and cyanate, Miller, starting directly with cyanacetaldehyde and exploiting the extreme solubility of urea, was able to produce cytosine and uracil in combined yields greater than 50% (43).

Concentration mechanisms The investigation of prebiotic chemistry often involves a compromise between extreme conditions that favor efficient product formation on a timescale amenable to laboratory investigation and milder conditions that are more plausible on the early Earth. The most optimistic estimates for the production of reactive biological precursor molecules, while substantial in their global amounts, when diluted throughout the volume of the Earth’s hydrosphere would result in only meager concentrations (44). In some cases low reactant concentrations simply result in slower reaction rates, which may not present serious difficulties when considered in the context of a geologic timescale. In other instances, however, reaction pathways at low concentrations may differ significantly from those at higher concentrations. HCN polymerizations, for example, produce biologically interesting materials only at HCN concentrations of about 0.01 M or higher. At HCN concentrations lower than 0.01 M, hydrolysis out-competes oligomerization, resulting in the accumulation of formamide and formic acid instead of HCN oligomers (45). Concentration mechanisms are often invoked to address these
difficulties and boost efficiencies, and, indeed, certain concentra-
tion processes seem likely to have played a role in nature.
Miller’s rationale for using such high concentrations of urea in
his synthesis of cytosine was the “drying lagoon” model of
prebiotic synthesis in which periodic evaporation of bodies
of water to near dryness, such as beaches, lagoons, and tide
pools, concentrate the most soluble compounds to extremely
high concentrations, even if they were initially very dilute (46).
In the case of HCN, its volatility precludes evaporative con-
centration mechanisms, but freezing can be an effective means
of concentration, as the HCN-water eutectic is 74.5% HCN at
−23.4°C (47). HCN polymerizations proceed effectively under
these conditions, the extremely high concentrations compensat-
ing kinetically for the low temperatures (47).

Extraterrestrial synthesis of bases

Aspects of every putative prebiotic synthesis are vulnerable
to critical analysis, and experiments are often open to questions
of interpretation and relevance. This is especially true in ori-
gin of life research where virtually everything is unknown and
will remain so, as no direct evidence survives from that time.
Uncertainties about specific conditions, sources, and reaction
pathways, however, in the context of the discussion of nucleic
acid bases, are secondary to the likelihood that these particular
fundamental building blocks of life were produced in ubiquitous
abiotic processes throughout our solar system and presumably
beyond. Analyses of carbonaceous meteorites have detected
amino acids (48, 49) and nucleic acid bases (50–52) that are ex-
ternal to Earth. However, the distribution of compounds and
their relative abundances are reminiscent of those observed
in known prebiotic simulation experiments. These findings sug-
gest that conditions favoring the synthesis of these fundamental
biological materials, whether they be through the mechanisms
reviewed above or others that have not been considered, were
at work during the formation and early evolution of our solar
system, and, consequently, the nucleic acid bases are very likely
to have been present on the prebiotic Earth.

In addition to invoking the evidence of extraterrestrial syn-
thesis of biomolecules as a means to corroborate that these
reactions could have occurred on the prebiotic Earth, they can
also be invoked directly as potential source material for the
origin of life. The earliest stages of Earth’s history are character-
ized by a steady bombardment of extraterrestrial objects such as
meteorites and comets that could have delivered large amounts
of biologically useful material to the Earth. Less exciting than
meteor or comet impacts, but much more significant in terms of
the amounts of extraterrestrial material delivered to Earth, are
interplanetary dust particles (IDPs). IDPs are small carbona-
ceous particles on the order of tens to hundreds of micrometers
in size that are formed by comet evaporation or asteroid colli-
sions and they “rain down” onto the Earth’s surface in amounts
estimated to be at least several tons per day (153, 54). This sig-
nificant source of extraterrestrially synthesized organic material
is expected to have supplemented, perhaps substantially so, any
endogenously produced biological precursors.

Sugars

A sugar moiety, specifically the pentose D-ribose, is attached to
the base as the other half of a nucleoside. The synthesis of sugars
from formaldehyde, a compound generated in spark discharge
experiments and frequently invoked as a prebiotic reagent, has
long been known (55), and its applicability as a prebiotic pro-
cess has been well investigated. The so-called formose reaction
produces a rich mixture of sugars from a solution of formalde-
hyde, but requires an unrealistically basic pH and improbably
high concentrations of formaldehyde to seem prebiotically fea-
sible (56). The reaction has been modified to proceed under
much milder conditions at neutral pH, and can be catalyzed
by certain common minerals, which is a more attractive pre-
biotic scenario (57, 58). However, a second problem with the
formose reaction exists—its lack of synthetic selectivity. RNA
nucleosides utilize a specific sugar, D-ribose, but the formose
reaction generates a complex mixture of trioses through hex-
oses, including branched molecules, without a preference for
ribose, and without enantiomeric enrichment (59). In contrast,
the base-catalyzed aldolization of glyceraldehyde phosphate in
the presence formaldehyde produces ribose-2,4-diphosphate as
the major product (60). Further refinements to this procedure
using a clay mineral catalyst allow the reaction to proceed ef-
ciently with dilute reactants at neutral pH (61). In addition,
other mechanisms to further enrich for ribose from a mixture
of sugars have been demonstrated, such as selective membrane
permeability (62) or selective reactivity with reagents that form
adducts preferentially with ribose (63). It is also possible that
unknown mechanisms or untested catalysts may exist that can
produce D-ribose more efficiently, but have not yet been dis-
covered. A third problem is the instability of the sugars, which
are very likely to have been present on the prebiotic Earth, but
also, the instability of the sugars make it unlikely that
ribose could persist in solution and accumulate on the prebiotic
Earth on time scales that are typically assumed for the origin of
life (64). However, the clay minerals that have been shown
to catalyze the synthesis of ribose phosphate have also been
demonstrated to provide a protected environment to sequester those
products, resulting in increased stability relative to free solution
(61). An alternative approach involves creating stable adducts of
ribose with other molecules. Recently, Benner and co-workers
have demonstrated that certain borate minerals can substantially
stabilize sugars, such as ribose, that have hydroxyl groups on
adjacent carbons, sequestering the sugar in a stable complex and
thereby prolonging the sugar’s availability for further assembly
(65). Similarly, the product of ribose and cyanamide forms an
insoluble adduct that crystallizes out of solution protecting the
ribose from solution-mediated degradation.
Nucleosides

A nucleoside is the combination of a base and ribose. Their formation, in theory, is most directly achieved by a reaction involving pre-formed bases and ribose, but this approach, so far, has shown little promise. The best attempts have been the demonstrations that the dry-phase heating of purine bases and ribose leads to the formation of purine nucleosides (66, 67). Further, the yields are enhanced several-fold in the presence of certain salts, including evaporated seawater. However, the same investigators were unable to substantiate an earlier report claiming the formation of adenosine from dilute solutions of adenine and ribose exposed to ultraviolet light, and were also unable to detect any nucleoside formation in aqueous solution under a wide variety of conditions. This dry-phase synthesis, while producing nucleosides in yields of several percent, is complicated by a side reaction in which bases bearing exocyclic amines (adenine, guanine) are preferentially ribosylated on the amine group in yields of up to 74% (66). Similar reactions with hypoxanthine, which does not have an exocyclic amine group, were not affected by this side reaction, and produced inosine in yields comparable to the adenine and guanine reactions. As was the case with a lack of selectivity between synthesis of D and L enantiomers, a similar lack of selectivity exists in the formation of α and β anomers.

The problem of pyrimidine nucleoside synthesis has been even more challenging than purines. In dry-phase heating experiments, the pyrimidines do not form detectable amounts of nucleoside (66). Alternative strategies include using prebiotic reagents to build a base on a pre-formed ribose (68-70), or the converse, building a sugar on a pre-formed base. Both tactics have had degrees of success, but neither seem to be the simple and robust processes that might be expected of a convincing prebiotic synthesis, especially of such a crucial class of biomolecules. Still, all of the above approaches use other pyrimidines (not C, U, or T) (71) or pyrimidine-like bases (72) that more readily form nucleosides with ribose. In a subsequent step, these non-nucleoside pyrimidines are then either chemically converted to a standard pyrimidine or else invoked as a precursor to conventional RNA.

Nucleotides

The last step to make an RNA monomer is to add phosphate. This step can be coupled with chemical activation of the nucleoside in order to make the polymerization of monomers energetically favorable. A phosphate links each nucleoside monomer together in an RNA polymer through phosphodiester bonds between the 5′ and 3′ hydroxyls on the riboses of adjacent nucleotides. The activated form of the nucleoside used by all RNA polymerases in contemporary biology is the nucleoside 5′-triphosphate (NTP). The energy that drives polymerization is accessed by breaking the bond linking the α and β phosphate groups, with pyrophosphate released as a byproduct. Attempts to create nucleotides by phosphorylating nucleosides under prebiotic conditions have had mixed success. The most successful approach involves the dry-phase phosphorylation of nucleosides with phosphate salts by heating a dried solution of nucleoside and inorganic phosphate. This reaction produces an assortment of different mono-phosphorylated nucleotides (NMPs) (73). The addition of urea to the reaction increases the yields substantially, but still produces a mixture of products, including 3′- and 2′-phosphates and 2′-3′-cyclic phosphates (NMPs), which, in many cases, are the major product (74). The relative yields of products are sensitive to reaction conditions, which can be fine-tuned to favor the accumulation of 5′-NMP (75). For instance, the temperature of the reaction affects the distribution of products, particularly the formation of NMP. The earliest experiments were carried out at temperatures of 100°C and resulted in the greatest proportion of NMP phosphorylated products. Simply lowering the temperature to 65°C greatly reduces the amount of NMP relative to the 5′-NMP product. However, lowering the temperature below 60°C inhibits all phosphorylation (75).

An alternative to the dry-phase phosphorylation schemes involves the use of a condensing agent in aqueous solution. Small yields of NMPs are produced from heated solutions of nucleoside, phosphate, and certain prebiotically feasible condensing agents such as cyanogen, cyanate, and cyanoacetylene (76, 77). The yields are low because none of the tested condensing agents provide any selectivity for the nucleoside over water. Consequently, water is able to out-compete nucleoside for phosphorylation due to its substantial concentration advantage. With what little phosphorylation of the nucleoside that does occur, as with the dry-phase reactions, NMP is found as a major product of these reactions. The observation that the unimolecular cyclization of 3′-NMP to NMP proceeds efficiently under these conditions is consistent with this model (78).

Other possibilities include adding or building a base onto a pre-formed sugar phosphate as mentioned in the previous section (69, 70) or utilizing phosphate derived from minerals or extraterrestrial sources. For example, the phosphorous-containing mineral schreibersite, found in many meteorites, decomposes in aqueous solution generating compounds containing phosphate, pyrophosphate, and phosphate (79). The proposed degradation pathway proceeds through a phosphate radical intermediate, which could potentially be used to form polyphosphates. In this regard, phosphate has been suggested as an attractive prebiotic alternative to phosphate due to advantages in solubility and reactivity (80).

The mono-phosphorylation discussed above completes the nucleotide unit, but does not create the high energy activated phosphate that drives polymerization, so an additional step must be invoked to complete the activated nucleotide. This could occur either by using a nucleoside with an activated phosphate group or in the form of using an external activating agent to effect the polymerization process. Polyphosphates can be made under similar conditions (81, 82), but generally require higher temperatures to achieve, and their utilization in a prebiotic setting has been questioned (83).

Cyclic phosphates

The emphasis in the area of nucleotide synthesis is almost always focused on 5′-activated phosphates, based in large measure, on nature’s (current) preference for NTPs. Alternatives to this model include 2′,3′-cyclic phosphates, which have a strong case for consideration. First, they are often the major products,
rillonite to otherwise similar reactions has a favorable effect. The distribution of products is highly sensitive to reaction conditions, and the behavior of non-enzymatic RNA polymerization is often bypassed in these types of non-enzymatic template-directed polymerization experiments.

Nonenzymatic Assembly of Oligonucleotides

The prebiotic utility of nucleotides lies in their ability to form long, polymeric chains that store genetic information and form complex and functional macromolecular shapes. Any initial assembly of oligonucleotides presumably would have been non-nucleotide-templated and occurred either in solution or perhaps on a clay or mineral surface. The energy to drive polymerization could come either in the form of a pre-activated nucleotide or by utilizing an external activating agent during polymerization (discussed above). The latter approach, which bypasses the pre-activation of the nucleoside phosphate, would balance that advantage in the inefficiency of polymerization. Typical experiments yield small amounts of mostly dimers and trimers, and among these are a mixture of the "natural" 3'5' internucleotide linkages with unnatural 2'5' linkages, as well as other complex products including 5'5'-pyrophosphate linkages (85). The alternative strategy of pre-activating the nucleoside has produced the most robust laboratory results. The choice of activating group is dictated primarily by practical considerations. Nucleoside triphosphates, nature's chosen activating group, react too slowly to be practical for the laboratory. Other activating groups have been explored, and among these, the most studied have been the phosphorylazidials. The prevalent use of phosphorimidazolides in these types of non-enzymatic polymerization studies is based on the simplicity of their synthesis and utilization, as well as their successful results. The steps and components of their synthesis can be considered plausibly prebiotic, but unlikely to persist and accumulate to reasonable levels. Their popularity as a polymerization substrate is less an endorsement of their prebiotic relevance and more the adoption of a well-behaved model system with which to address general and, hopefully, fundamental questions about the behavior of non-enzymatic RNA polymerization.

Polymerizations of this type in solution produce predominantly shorter than 5-mer oligonucleotides with a mixed and variable composition of 2'5' and 3'5' linkages (86). The distribution of products is highly sensitive to reaction conditions, including the presence of metal ions, modification of the leaving group, and sequence context. The addition of the clay montmorillonite to otherwise similar reactions has a favorable effect on polymerization (87). The effect is likely due to the concentration of monomers within the clay's cationic layers. These clay-catalyzed experiments have produced polynucleotides up to 55 nucleotides long with approximately 80 percent 3'5'-linked composition (88).

Template-directed synthesis

The appeal of RNA or any nucleic acid as a genetic material is due to its ability to base-pair and consequently serve as a template for the synthesis of complementary copies of itself. Eventually, random, non-templated RNA assembly would have had to give way to template-directed systems able to propagate beneficial genetic information from one generation to the next. When a template is added to prebiotic polymerization simulations, the effects can vary dramatically depending on the nucleotide composition of the template. In basic experiments with homopolymeric templates, the complementary strand is polymerized from monomers that base-pair with the template. The incorporation of purines, particularly guanosine, into the complementary strand is much more efficient than pyrimidine incorporation (89). These trends hold true for mixed sequence systems: good templates are generally those with a high percentage of cytidine, directing the incorporation of guanosine. Such a system might be a good template from which to make a reverse-complementary copy, but that copy would be an extremely poor template from which to regenerate the oligonucleotide. The complementary nature of nucleic acid replication cannot succeed with such an extreme templating bias unless another, currently unknown, rescue mechanism is postulated.

The increased efficiency of oligomerizing dimers, trimers, or longer oligomers instead of monomers is an attempt to address some of these difficulties. The energetics of base-pairing longer stretches of nucleotides increases binding efficiency, but unless the pieces are sufficiently long, the templating peculiarities just discussed can still apply to oligomerizations and severely restrict the number of replicable sequences. The feasibility of this kind of system has been characterized using a carefully chosen patroductom system with trimers ligating on a hexamer template (14). A modified system using a manual denaturation step to separate the product strands from the template strands during each generation of copies has demonstrated exponential replication of nucleic acids (15).

Enantiomeric cross-inhibition

A previously mentioned difficulty in prebiotic synthesis, now re-surfacing for this discussion of prebiotic template-directed oligonucleotide synthesis, is the problem of sugar handedness. Unless a reasonable mechanism or process is found that leads to the accumulation of one sugar enantiomer over another, the prebiotic synthesis of sugars must be assumed to produce an equal mixture of D and L ribose, and, consequently, also in the absence of a reasonable selectivity mechanism, lead to a mixture of D and L nucleotides. Studies of uncatalyzed templated polymerizations using a mixture of nucleoside enantiomers indicate that the non-biological "L" enantiomer acts as a chain terminator when incorporated against a D enantiomer template, blocking further templated extension of the oligonucleotide (90). This is another serious difficulty without a satisfactory
resolution with regard to the uncatalyzed template-directed polymerization of RNA nucleotides.

Ribozyme to Replicase

Ignoring, for the moment, the challenges associated with the abiotic synthesis of RNA polymers, the emergence of the RNA World still depends on the ability of RNA to catalyze its own replication in some fashion. An RNA-dependent RNA polymerase is the class of enzyme that modern biology would use for this purpose, and a ribozyme with this functionality is thought to be central to the beginnings and propagation of the RNA World. No such ribozyme has yet been discovered in nature, but a laboratory re-creation of a ribozyme with that functionality, though not proof of its historical existence, would show that it could once have existed, and validate a fundamental pillar of the RNA World hypothesis. The isolation of RNA ligase ribozymes from a population of random sequence RNA molecules was the start of an experimental progression that has come closest to realizing the goal of an RNA polymerase ribozyme (91). RNA ligase ribozymes catalyze the identical chemical reaction as an RNA polymerase, but instead of the triphosphate-bearing half of the reaction being an NTP, it is instead present as the 5' triphosphate of the 5′ terminus of the ribozyme itself (Fig. 5). In terms of in vitro selection targets, ligase activity appears to be relatively prevalent in RNA populations after having been isolated under different reaction conditions by several teams using different starting populations of RNA. The phrase “relatively prevalent”, in this case, means on the order of one ligase ribozyme in every 1011-1014 RNA molecules (91), only about an order of magnitude lower frequency than experiments isolating self-cleaving ribozymes from random sequence populations (92, 93). While these probabilities may seem relatively low, they correspond to finding a ~100 nucleotide ligase ribozyme in approximately every microgram of random RNA, a trivial amount of material assuming that mechanisms for the prebiotic synthesis of RNA oligomers existed at all.

A strategy to convert RNA ligase ribozyme into a general RNA polymerase involves separating the template and substrate from the ribozyme, optimizing the enzyme to “ligate” single nucleotides as substrates, and translocating along the template to the next position. Following a symbiotic regime of rational design and in vitro evolution, Bartel and his coworkers created a polymerase ribozyme able to polymerize along a template up to 14 nucleotides, the equivalent of a full turn of RNA helix (94). Technological advances in in vitro selection methodologies have led to a modest improvement of more than 20 polymerized nucleotides, which represents the current state of the art in polymerase ribozymes (95). This lineage of experiments has generated the most sophisticated artificial ribozyme known to science, but still falls well short of the presumed efficiency that would be needed in the RNA World.

Nucleoside-2′,3′-cyclic phosphates have been mentioned as candidates for activated monomers in the context of non-enzymatic polymerizations, and also deserve mention in the context of ribozyme-catalyzed polymerizations. The initial experimental verification of the possibility of a cyclic phosphate polymerase was the isolation of cyclic phosphate RNA ligase deoxyribozymes (96). Once again, in vitro selection was used to explore large populations of both DNA (97) and RNA (M.P. Robertson, J.E. Blaustein, and W.G. Scott, unpublished results) and isolate new ligases. Attempts to convert these ligases into polymerases have not been reported. RNA has proven to be a remarkably malleable tool in the hands of biochemists, and a satisfactory RNA-dependent RNA polymerase ribozyme will probably be developed in the laboratory eventually. When that happens, it will be the culmination of at least a decade’s worth of directed efforts, using complex combinations of rational design and in vitro evolution. It will be an important demonstration that further strengthens the case that an RNA World could have existed, but it will also show that RNA polymerase ribozymes do not appear to be a simple functionality that is likely to appear de novo from a naïve population of RNA molecules. The possibility of a self-replicating RNA system less sophisticated than a general RNA-dependent RNA polymerase ribozyme cannot be eliminated.

Ribozymic Metabolism

The minimal theoretical requirement to initiate the RNA World is a self-replicating RNA polymerase ribozyme. Realistically, supporting functionalities would soon be necessary to sustain an emerging RNA World. For example, a reliable supply of activated nucleotides would have been essential as the polymerases depleted the presumably scarce supply of abiotically produced material. To demonstrate that ribozymes are capable of catalyzing such complex chemistries, a nucleotide synthetase ribozyme was created in the laboratory (98). The ribozyme appends a base to an activated ribose to form a nucleoside. This is just one step in what can be imagined as an ever more complex network of metabolic pathways supporting the propagation of the RNA organism. But while the boundaries of the RNA World begin with an RNA polymerase, and begin to end with the invention of protein synthesis, less can be inferred about what happened in between and what level of complexity was achieved before the RNA World began to wane. Based on the defined starting and end points of the RNA World, a polymerase with a supporting metabolism to produce its activated nucleotide substrates would be necessary. Closer to the end, a ribosomal-type translation system with a supporting metabolism to produce its necessary substrates is believed to have been in place. Beyond these examples, modern biology may hold clues of extant molecular relics of RNA World functionalities. In fact, the case in favor of an RNA World is built, in part, on the ubiquity of nucleotide-based cofactors throughout biochemistry and the suggestion that they are remnants of an extensive RNA-catalyzed metabolism (99). While the range of chemical reactions that natural ribozymes are known to catalyze is very limited, in vitro selection has generated a rich collection of ribozymes with diverse functionalities indicating that, under the right conditions, RNA is capable of catalyzing a large subset of the reactions normally catalyzed by protein enzymes, though not as efficiently.
A key characteristic of living systems that is thought to have been utilized very early in the evolution of life, if not concurrently or even preceding RNA replication, is compartmentalization. All modern cells are enclosed by membranes that act as a barrier to regulate the intake and escape of molecules from the system. This functionality is necessary to contain the byproducts of metabolism and as a mechanism to link genotypes with phenotypes, which would be just as important for the earliest forms of life. For instance, an RNA replicase operating free in solution without boundaries might be able to replicate RNA molecules that it encounters, but the most efficient replicases would not benefit from an evolutionary advantage unless their systems and progeny were segregated from less efficient competitors. In simple systems involving vesicle-encapsulated RNA, the synthesis of additional RNA resulted in spontaneous vesicle growth at the expense of empty vesicles. Preliminary studies with these types of protocell systems have shown that basic properties like growth, encapsulation, and division occur as a consequence of purely physical functionalities in order to be operational.

**preRNA**

In most discussions, the RNA World is synonymous with the origin of life, but this need not necessarily be the case. The difficulties associated with the prebiotic synthesis of long RNA polymers have led many to speculate that an alternative, simpler biopolymer—that can somehow eliminate the problems associated with RNA—preceded the RNA World. Most of these “preRNA” candidates retain the familiar Watson-Crick base pairing between strands, but a number of different types of alternative backbone systems have been proposed, including a non-cyclic carbohydrate (102), peptide (103), and a variety of non-ribose sugars (104, 105). But unlike RNA, no recognizable molecular artifacts have yet been found in support of any of them. In the absence of direct evidence, for any particular candidate can only be judged against a number of practical factors, including the ease and likelihood of its prebiotic synthesis pathway and the ability to ultimately transfer its sequence information to an RNA polymer. With examples of the genetic transfer of sequence information from a non-RNA polymer to a strand of RNA having been demonstrated (106), the main hurdle, like RNA, is in the ease and likelihood of its prebiotic synthesis. Most preRNA candidates have one or more attributes that address a shortcoming of RNA, but none of the candidates are without flaws themselves. In this respect, no current preRNA candidate has distinguished itself as clearly superior to the others.

**Summary**

The mysteries of how life originated on this planet and came to achieve its current complexity hold a special fascination in the imagination of humankind. The earliest era of life’s evolution are very poorly understood, but fragments of evidence exist. The RNA World hypothesis outlines a description of how the earliest molecular forms of life may have arisen. The fragmented evidence that points most convincingly toward an RNA World relies almost entirely on what is known of modern biochemistry and the plausibility of reverse engineering a pathway from now to then. The converse, however, the period of time leading up to the emergence of RNA-based life, is more challenging to imagine. Difficulties and uncertainties exist at every step leading up to the emergence of the RNA World, from the inefficiencies of abiotic organic syntheses to the probability of inventing a ribosome more complex than any that has been seen today. Because of the difficulties of creating an RNA World from scratch, many believe that a preRNA World based on a simpler, nucleic acid-like polymer, preceded the RNA World as the first living system.

**References**


Origins of Life: Emergence of the RNA World


Further Reading


See Also

Origins of Life: Emergence of Nucleic Acids
In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes)
Catalytic Modes in Natural Ribozymes
S-adenosyl-l-methionine is a high energy compound and is the major source of methyl groups for a myriad of biologic transmethylation reactions. These highly specific single-carbon transfers onto diverse nucleophilic centers in biomolecules are catalyzed by methyltransferase enzymes and play important regulatory and structural roles in the cell. Here we discuss the chemical mechanism of the methylation reactions, including structural features of the methylsulfonium center in the cofactor molecule, enzyme-assisted activation of diverse nucleophilic targets by deprotonation or covalent catalysis, and spatial constraints of the reaction.

The first literature source describing biologic methylation refers to the methyl donor S-adenosyl-l-methionine (AdoMet) as “ATP-activated form of methionine” (1). The formation of AdoMet from methionine and ATP is catalyzed by AdoMet synthetase (MAT, methionine adenosyltransferase, EC 2.5.1.6) and occurs in a two-step reaction in which PPi and Pi are released along with the product (AdoMet) (2). Although almost any part of AdoMet molecule can be used by the living organisms (3), the most ubiquitous and important role of AdoMet is the participation in biologic transmethylation reactions (Fig. 1). Other methyl group donors such as tetrahydrofolic acid, betaine, and vitamin B12 are used in certain cases, but AdoMet is by far the most often used source of methyl groups and the second-most ubiquitous cofactor after ATP. The methyl transfer reactions from AdoMet are associated with very favorable enthalpies (∼−70 kJ/mol) compared with other methyl donors, which permits efficient and selective methylation of a large variety of biologic substrates in all living organisms from bacteriophage to humans (4).

### Methyl Group as a Biologic Mark

Methyl group transfer to particular targets is directed by enzymes called methyltransferases (MTases, EC 2.1.1.-). These enzymes catalyze more than 150 different reactions (according to SWISS-PROT database). AdoMet-dependent methyltransferases according to their substrates can be classified into four major groups: enzymes acting on small molecules, proteins, nucleic acids, and glycans. The atomic targets in these molecules can be carbon, oxygen, nitrogen, sulfur, or halides (4). Because of this enormous variety of substrates, methyltransferases play vital roles in many cellular processes, including cellular metabolism, signal transduction, storage, and processing of (epi)genetic information.

In a certain sense, the methyl group plays a comparable role with the phosphoryl group in biologic systems. From the chemical standpoint, the methyl group is a small (volume ∼20 Å³) and uncharged apolar group, which is usually added to replace a hydrogen atom (volume ∼5 Å³) in a target molecule. Therefore, depending on the role and the chemical context of its predecessor hydrogen, the structural content of the methylation signal will be different. If minimal chemical alterations in the target molecule are brought about, the methylation can be regarded as a “steric” signal. Such subtle additions of a small chemical group can lead to a highly specific recognition by cellular components and dramatic biologic consequences. Alternatively, when methylation leads to altered chemical-physical properties, such as tautomeric forms, H-bonding patterns, and new chiral centers, its role could be assigned as “chemical.” As an extreme example of a “chemical” role, the methyl group can serve as a chemical “activator” for a subsequent chemical conversion of a target molecule.

### Small-molecule methyltransferases

Out of more than 150 different reactions catalyzed by the methyltransferases more than 90 are carried on small molecules. The methyl groups serve here as important “chemical” building blocks required for the construction of essential cellular components and metabolites. N-methylation of the smallest amino acid glycine is responsible for regulating AdoMet/AdoHcy ratio in the eukaryotic cells (5). Catechol O-methyltransferase (COMT) is responsible for dopamine and other catecholamine
Chemistry of AdoMet-Dependent Methyltransferases

Figure 1
Action of AdoMet-dependent methyltransferases. (Top) Methyltransferases catalyze the transfer of the methyl group from the cofactor AdoMet onto defined nucleophilic targets (Nu = N, O, C, S) in various biomolecules (shown as gray balls). Activation of nucleophiles is often achieved by abstracting a proton (if present) from the target atom by a general base (B:). (Bottom) The methylation reactions proceed via a direct SN2 transfer, in which the attacking nucleophile and the substrate are involved in a single transition state structure. These reactions require a transient hybridization change from $sp^3$ to $sp^2$ and back to $sp^3$ with inversion of configuration at the reacting carbon.

neurotransmitters’ methylation hence, it influences nervous signal propagation (6). Another important enzymatic methylation converts norepinephrine to epinephrine (adrenaline) (7). Creatine, which is used by the cells for energy storage, is synthesized by methylation of guanidinoacetate (8). Methylated hydroxycinnamic acid derivatives are precursors of the plant cell wall esterified phenolic compounds, soluble sinapate esters, dimeric lignans, and the extensively cross-linked polymer lignin (9). Biosynthesis of vitamin B12 employs six different methyltransferases that add methyl groups to specific positions of the tetrapyrole ring (10). Finally, histamine N-methyltransferase inactivates inflammatory and allergenic mediator histamine (11). An example of a fatty acid methylating enzyme is cyclopropane-fatty-acyl-phospholipid synthase from M. tuberculosis. Its action is required for the long-term survival of nongrowing cells and is often associated with environmental stresses (12).

DNA methyltransferases

DNA methyltransferases modify nucleobases by depositing methyl groups onto exocyclic amino groups (N6 in adenine and N4 in cytosine) or the intracyclic C5-position of cytosine. Because these methylation reactions occur in the major groove of the DNA helix and without chemical consequences on the DNA structure, they can be viewed as “steric” signals designed for recognition by specialized proteins, enzymes, or large multicomponent complexes. All three types of DNA methylation found in prokaryotes and archaea occur sequence-specifically. A unique DNA methylation pattern (a combination of several methylated sequences) serves as a discriminatory species “self” code. In higher eukaryotes, the cytosine-5 methylation is solely known, which occurs in both a sequence-specific and a locus-specific manner. DNA methylation generally leads to a strong and heritable repression of gene expression and plays numerous essential regulatory roles in cellular differentiation and development, parental imprinting, X-chromosome inactivation, and silencing of endogenous retroviruses (13).

RNA methyltransferases

Methylation of RNA is even more diverse and abundant than that of DNA. RNA methyltransferases target nearly all chemically accessible sites on nucleobases and the ribose: N1, N2, and N7 in guanine, N1 and N6 in adenine, C5 in cytosine or uracil, and 2′O in ribose. Most of the known biologic methylation reactions seem to serve as a means for “chemical” tuning of RNA transcripts into biologically active species. For example, guanine-N1 methylation of tRNA prevents frame shifts during protein translation (14) in bacteria. 2′O-ribose methylations at specific positions are essential for stability of tRNA in both eukaryotes and prokaryotes (15). The same modifications guided by small, nucleolar RNAs to specific loci on rRNA is required for ribosome assembly (16). All mRNAs in eukaryotes are capped at their 5’ end, and at least two RNA MTases are required for the maturation of the cap structure; these modifications are required for the stabilization and efficient translation of the transcripts (17). Resistance to clinically important antibiotics (macrolide, lincosamide, and streptogramin B) is conferred by adenine-N6 methylation in 23S rRNA (18).

Protein methyltransferases

In proteins, a large variety of methylation targets has been identified: the carboxylate of aspartate and glutamate, the sulfur of cysteine and methionine, the imidazole of histidine, the amide of glutamine and asparagine, the guanidinium of arginine, the ε-amino group of lysine, and the terminal amino and carboxylate groups (19). Arginine and lysine can accept more than...
one methyl group producing symmetrical and unsymmetrical dimethylarginine, dimethyllysine, and trimethyllysine.

In the focus of a renewed interest are histone M-Tases. They produce monoarginine, symmetrical and unsymmetrical dimethylarginine, and all possible lysine ω-amino group methylation states. Such post-translational modifications of the histones determines whether chromatin adopts a compacted structure and is associated with silenced DNA–heterochromatin, or if it seems to be an extended structure and is associated with transcriptionally active DNA– euchromatin (20). Arginine M-Tases seem to methylate even more substrates, but the modifications effects are not well understood (21). As in DNA, the above-described examples of methylated residues can be viewed as “steric” marks designed for recognition by highly specific proteins.

In contrast, 3-isopropylamine T-MTase methylates the carboxyl group of 3-isopropylamine residues, which are spontaneously accumulated in proteins via inadvertent isomerisation of asparagines and asparagines. The resulting methylisoxa carbonyl group is reactive in the reverse intramolecular trans-esterification reaction, which regenerates an isopropylamine residue. Therefore, the methylation is the activating step on a chemical repair pathway of aged proteins and plays an important role in extending the life span of organisms from all domains of life (22).

The Methyl Donor: S-Adenosyl-L-Methionine

As discussed, AdoMet is a high energy compound. The methyl group is activated by the neighboring sulfonium center, and AdoMet serves as a co-factor during the M-Tase-catalyzed reactions. The sulfonium center induces a partial positive charge on all three adjacent carbon atoms (methyl group, ribose 5'-carbon, and methionine γ-carbon). This charge is clearly observed from a nearly 1 ppm high field shift of proton chemical shifts in AdoMet as compared with those in methionine and AdoHcy (23). Under physiologic conditions, both adjacent methylene groups can be attacked leading to slow decomposition to reactive species. AdoMet is particularly labile under alkaline conditions, forming adenine and S-ribosylmethionine readily. This reaction is initiated by the deprotonation at C-5′ (24). A competing pathway for AdoMet degradation, which is prominent at even lower pH values, involves an intramolecular attack on the methyl group with 2H0 r 13C result in slight positive and negative isotope kinetic effects, respectively, in the case of COMT enzyme. Based on these observations a “symmetrical” and “asymmetric” transition state has been proposed (28). It is now generally accepted that enzymatic methylation reactions proceed via a direct S2′ transfer, in which the attacking nucleophile and the substrate are involved in a single transition state structure (see Fig. 3). The catalytic power of many methyltransferases thus largely derives from their ability to bring the two substrates together in correct orientation.

Substitutions in the sulfonium center

AdoMet analogs, in which sulfur of the sulfonium center is replaced with selenium or tellurium, have been synthesized and used to study the reaction mechanism (Fig. 2a–c). Electronegativities of S, Se, and Te (2.58, 2.55, and 2.30 on the Pauling scale, respectively) suggest that carbons adjacent to telluronium groups are poorer electrophiles as compared with carbons adjacent to selenium or sulfur groups atom because of the reduced ability of the tellurium atom to induce a positive dipole at the adjacent carbon. On the other hand, the increase in atomic radii results in weaker heteroatomic bonds in the series S–CH3 > Se–CH3 > Te–CH3, thus making the higher analogs better leaving groups. These two opposing effects result in a slight chemical activation of the S-analog and substantial inactivation of the Te analog, as observed in the two enzymatic systems examined (29, 30). The same reasoning seems to be valid for explaining the relative effects on the heteroatoms on the two modes of decay of the analogs in water: deprotonation at C-5′ and intramolecular nucleophilic attack at C-γ. The deprotonation reaction is decreased, whereas decomposition via the C-γ attack is enhanced in the selenium analog as compared with AdoMet. This is understood taking into account that the first pathway involves the activation of the carbon atom, whereas the second requires both electrophilic activation of the carbon and a good leaving group.

AdoMet analogs with nitrogen replacements of the sulfur atom (N-adenosyl-L-azamethionine or azo-AdoMet; see Fig. 2d (31) seem even less reactive. These compounds are not susceptible to the decomposition reactions peculiar to AdoMet. They can act as charge-switchable mimics of AdoMet because the tertiary amino group (pKa = 7.1) can be protonated by adjusting the pH slightly below the physiologic values (32). However, because diakylamines are not nearly as good leaving groups as are dialkylsulfides, these compounds are not expected to be efficient methyl group donors in methyltransferase-catalyzed reactions. A further problem may derive from a low inversion barrier at nitrogen as compared with sulfur, leading to lower abundance of the correct epimeric form and thus poor overall positioning of the reactants in the enzyme pocket. Although initial testing suggested that NAM could serve as a substrate (31), this compound functioned as a substrate but not as a methyl group donor in the case of t-RNA uracil-5′-methyltransferase (33) and many other systems. Notably, a significantly improved reactivity is achieved in analogs carrying a sterically tense aziridine cycle (see Fig. 2e). Upon protonation of the ring nitrogen, enzyme-assisted nucleophilic attack on the ring carbon leads to the opening of the ring, which couples the whole cofactor to the target (Fig. 3a (34)). Alternatively, the aziridine ring can be generated in situ in
Nitrogen is a relatively good nucleophile but also a good base. In biologic substrates, nitrogen occurs in the form of alkylamines (glycine, lysine, or phenylethanolamine), aromatic amines (cytosine-N4, adenine-N6, or guanine-N2), heterocyclic systems (guanine-N1 or -N7, adenosine-N7, or histidine), or conjugated amines (glutamine or guanidine). Generally, an enzyme needs only to orient a lone pair of the nitrogen for \( \text{S}_{\text{N}}2 \) in line attack onto the methyl group of the AdoMet. If the nitrogen is protonated (alkylammonium), the proton needs to be removed before, in concert with, or after the methyl transfer; this step usually requires the presence of a general base in the active site.

### N-methylation

A well-studied example of aliphatic primary amino group methylation is provided by histone lysine MTases. The critical step in methylation of this amino group is its deprotonation, because a protonated lysine is a very bad nucleophile. The rate-determining step is the nucleophilic attack by the substrate on the electrophilic methyl group of the AdoMet. The reaction can be rescued by placing a double bond (allylic system) or a triple bond (propargylic system) next to the reactive carbon in the extended side chain (see Fig. 2k and 2l). The unsaturated bonds seem to stabilize the \( \text{sp}^2 \) transition state via conjugation of their \( \pi \) orbitals with the transient \( \rho \) orbital in the penta-coordinated transition state (37). These new synthetic cofactors are thus termed double-activated AdoMet analogs because the reactive carbon located between the sulfur and nitrogen is stabilised by resonance involving double bonds (sp2 orbitals) next to the sulfur atom. Consequently, the whole cofactor is covalently attached to the target molecule and irreversibly inhibits the enzyme (39).

### Substitutions in the transferable methyl group

AdoMet analogs with larger moieties replacing the methyl group (see Fig. 2k) have been obtained previously enzymatically (36) and later via regiospecific chemical S-alkylation of AdoMet. Such “reverse” chemistry additionally enhances the chemical stability of cofactor. Because of the positive charge of the protonated amine and correct chirality at the carbon center, sinefungin has an extremely high inhibitory potential for AdoMet-dependent methyltransferases.
Chemistry of AdoMet-Dependent Methyltransferases

Figure 3  Methyltransferase-directed coupling of extended groups to biomolecules using analogs of AdoMet. (a) Covalent coupling of N-aziridine and N-mustard cofactor mimics; (b) transfer of an extended aliphatic chain from a double-activated cofactor, S-adenosyl-L-propenthionine; (c) covalent coupling of an S-vinyl-analog of AdoMet.

are two possible ways to achieve this, as follows: “passive” catching of deprotonated species or “active” deprotonation of the lysine by a base in the active site of the enzyme. It turns out that both mechanisms are in use. In SET domain lysine MTases, the absence of an apparent general base in the active site and a very high and sharp pH optimum (∼9–10) (40) suggest a passive mechanism. In contrast, classic fold lysine MTases, Dot1p, are active in a broader pH interval (6–9.5). These enzymes contain a conserved essential Asn residue in the active site, which may facilitate the deprotonation of the target lysine (41, 42).

The methylation of exocyclic (aromatic) nitrogens proceeds in a different fashion, because pKa values of such nitrogens are usually low enough to stay unprotonated under physiologic conditions. The best representatives of this kind of N-methylation are cytosine-N4 and adenine-N6 MTases. These enzymes make hydrogen bonds to the target amino group directing its hydrogens to positions corresponding to sp3 hybridization. Such a “forced” hybridization change is thought to enhance the nucleophilicity of the nitrogen lone pair and to accelerate the reaction (43). A similar mechanism may also be used in nonaromatic conjugated amines, for example, in arginine MTases. Because of a very high pKa value of the guanidinium group (∼12), arginine methyltransferases are unlikely to achieve its complete deprotonation. Instead, they could polarize the guanidinium group and redistribute the positive charge away from the target nitrogen such that the nucleophilic attack on the methyl group of AdoMet et is accelerated (44).

O-methylation

Oxygen is less nucleophilic as compared with nitrogen. However, its nucleophilicity can be enhanced by generating a (partial) negative charge on it. In biologic substrates, modifiable oxygens are often found in three types of groups: phenolic, ribose hydroxyl, and carbonyl groups.

No general base catalysis is needed in the case of a carboxyl group, because it is usually deprotonated (pKa ∼4.5) under physiologic conditions. Indeed, a well-characterized example of such an enzyme, protein-isoaspartate O-MTase, contains no acidic or basic residues in the substrate-binding cleft (22). Phenolic hydroxyl groups have a markedly high pKa (∼10.5), and thus, they remain largely protonated under physiologic conditions. Enzymes can enhance the methylation reaction rate by abstracting a proton. For example, in the catechol MTase, the target hydroxyl group is coordinated with essential Mg2+, which apparently replaces the proton to generate a nucleophilic phenolate (6). Interestingly, electron-withdrawing groups in the phenol ring, such as nitro groups, lead to strong inactivation of the substrates converting them to potent inhibitors of the enzyme. Alternatively, caffeate, isoflavone, and chalcone O-methyltransferases use a well-positioned His residue for proton abstraction (45).

As in aromatic alcohols, ribose 2′-hydroxyls are unprotonated under physiologic conditions (pKa ∼14.5). However, it seems less likely that enzymes can catalyze the methylation reaction by abstracting directly the proton from the hydroxyl group.
Alternately, rotation freezing and steering an oxygen lone pair toward the methyl group could sufficiently enhance the reaction. Such a mechanism is exemplified in the VP39 cap-2 mRNA (nucleoside-2′-O-)methyltransferase. The enzyme is thought to form a nondeprotonating hydrogen bond between a Lys side chain and the 2′-OH proton. That would be possible if the lysine was unprotonated beforehand; i.e., its pKa value was much lower than usual. An 15N-labeled steric mimic of the lysine was employed to confirm that the pKa is indeed perturbed (~8.5) by adjacent Arg and Asp residues (46). The ultimate removal of the proton from the ribose oxygen onto the lysine occurs synchronously with or right after the methyl transfer. A similar mechanism may be valid in the case of tRNA guanosine-2′-O-methyltransferase, which uses an Arg residue to scavenge the proton (47).

C-methylation

Because of the low intrinsic nucleophilicity of carbon and a high energetic cost of generating an intermediate carbocation or its equivalent, the formation of a C–C bond in aqueous milieu seems a challenging task. However, single carbon transfers to carbon centers are a common event in biologic systems. Among the best studied examples are the methylation of the CS-position in pyrimidine nucleobases, cytosine and uracil, the methylation of tetrapyrrole system during synthesis of vitamin B12, and in pyrimidine nucleobases, cytosine and uracil, the methylation seems a challenging task. However, single carbon transfers to carbon centers are a common event in biologic systems. Among the best studied examples are the methylation of the CS-position in pyrimidine nucleobases, cytosine and uracil, the methylation of tetrapyrrole system during synthesis of vitamin B12, and in pyrimidine nucleobases, cytosine and uracil, the methylation.

The catalytic mechanism of the pyrimidine-5 methylation in nucleic acids is more complex as it involves covalent catalysis. The mechanism is common for numerous DNA/RNA cytosine MTases as well as for thymidylate synthase (although the latter uses tetrahydrofolate as the methyl donor) and has been studied in detail in several systems (50). Here, the cytosine-5 methylation in DNA is presented as an example (see Fig. 4a). The CS-position of cytosine, which is part of an aromatic ring, does not carry sufficient nucleophilicity for a direct methyl group transfer. The continuity of the aromatic system is disrupted by a nucleophilic attack of thiolate from a conserved cysteine residue in the enzyme on the carbon-6 (51), which is accompanied by protonation of N3 (by a conserved glutamate). The resulting 4-5 exomethylene structure provides sufficient electron density at CS (52) for a direct attack on the methyl group of AdoMet. The methyl transfer step is irreversible. The methylated intermediate 5,6-dihydrocytidine is resolved into 5-methylcytosine and free enzyme via deprotonation at CS and β-elimination of the cytosine residue. The nature of the base responsible for the CS-deprotonation in DNA cytosine MTases remains elusive, whereas an RNA uracil-5 MTase was shown to employ a second conserved Cys residue for that purpose (53). In the absence of cofactor, a proton from bulk water can reversibly bind to the CS of the target cytosine resulting in MTase-dependent exchange of the CS-hydrogen into solvent (54).

A series of mechanism-based analogs of cytosine was used to elucidate the mechanistic details of covalent activation. The significance of these analyses extends into the realm of MTase inhibitor design for anticancer therapies (54). Among the best known inhibitors of DNA cytosine-5 MTases are 5-aza-2′-deoxycytidine (5-aza-dC or decitabine), 5-fluoro-2′-deoxycytidine (5-F-dC), and 2-pyrimidinone-5,β-π(2′-deoxyriboside) (zebularine) (53). In 5-aza-dC, the nucleophilic attack of the thiolate on the ring system is strongly facilitated, because a higher electron negative character of nitrogen at position 5 increases the electrophilicity at C6. In the presence of AdoMet, the covalent complex can still accept the methyl group from the cofactor, but it cannot be resolved additionally because no proton is present at N5, which is required to regenerate the double bond and free the enzyme. A similar mechanism is proposed for zebularine. 5-fluoro-2′-deoxycytidine exerts its catalytic activity through formation of a covalent intermediate with bound AdoMet, which irreversibly traps the enzyme in a stable covalent complex (55). Such covalent complexes were purified and crystallized to reveal first the structural details of the reaction intermediates (56).

Spatial Control in Enzymatic Transmethylation

The catalytic power of AdoMet-dependent MTases to a large extent derives from their ability to bring the two substrates, the cofactor AdoMet and a target molecule, together in the right orientation. As most other biocatalytic enzymes, MTases accommodate their substrates next to each other in a concave catalytic pocket. Five different protein folds are known to date that are used to bind AdoMet; however, one of them, similar to the NAD(P)-binding Rossmann fold, is highly prevalent (4). Many of these enzymes contain flexible loops that close during catalysis to cover the bound substrates from bulk solvent and, in certain cases, bring in important catalytic residues. This unique arrangement serves to create a proper milieu for the reaction. However, binding a target molecule in a concave pocket is not always easily achieved, because access to a specific locus that
Chemistry of AdoMet-Dependent Methyltransferases

Figure 4  Catalytic activation of cytosine for C5-methylation by nucleophilic addition of a thiolate at the C6 position. (a) The chemical mechanism of enzymatic DNA cytosine-5 methylation. Mechanism-based inhibition of DNA MTases by cytidine analogs 5-fluoro-2′-deoxycytidine (b), 5-aza-2′-deoxycytidine (c), and 2-pyrimidinone-1-β-D-(2′-deoxyriboside) (d).

is deeply buried within a large macromolecule is a challenging task. The most extreme and elegant examples are found in DNA MTases, which methylate nucleobases in double-stranded DNA. The target positions for methylation are exocyclic amino groups or the intracyclic C5 position of cytosine, which are located in the major groove of the DNA helix. In the latter case, a covalent catalysis is required for the methylation, meaning that access to both faces of the ring and an edge of the cytosine base is critical (see Fig. 4). It is hard to imagine how that could be achieved in the framework of the DNA helix with nucleobases tightly stacked on each other. Remarkably, this problem is solved by completely rotating the target nucleoside out of the DNA helix and into the active site of a MTase, with minimal distortions to the rest of the DNA (56). Similar mechanisms are operative for many other enzymes acting on nucleic acids, including many RNA MTases.

Another catalytic challenge relates to product control in cases when several methyl groups can in principle be transferred onto the same residue or atom. As mentioned, lysine methylation can lead to monomethylated, dimethylated, or trimethylated products. Because, for example, secondary amines are better nucleophiles than primary amines, it is tricky to stop a chemical alklylation reaction at the monosubstituted product. This problem is solved by building in an appropriate number of bulky residues (tyrosines) in the active site such that only a defined number of methyl groups is accepted (57, 58). But this alone does not ensure processivity. The production of dimethylated or trimethylated lysine requires several cycles of AdoMet binding, methyl transfer, and AdoHcy release. The protein architecture with two substrates bound in a single active site bears the risk of releasing incompletely methylated products upon reloading the cofactor. If that is to be avoided for biologic reasons, different protein architecture should be used. Examples of such enzymes are SET domain lysine MTases, in which AdoMet and the substrate peptide are bound in two separate grooves located on different sides of the protein (57, 58). The successive cycles of methyl transfer are carried out in the cofactor pocket, whereas independent processing of the substrate and release of a properly methylated product occurs in the other.

Practical Implications

Because many MTases play important roles in biologic processes, these are potentially good target candidates for drug design. Inhibitors for numerous enzymes have been produced.
and studied. Among important examples are inhibitors of COMT (see above), which are used as therapeutic agents in Parkinson’s disease (53). 5-aza-2′-deoxycytidine (5-aza-dC) and related analogs of cytidine are metabolically incorporated into DNA leading to mechanisms-based irreversible inhibition of DNA M Tases (54). Despite their high toxicity, these DNA-demethylating drugs can be applied in combination therapies with conventional chemotherapies.

Another interesting application for M Tases is the transfer of larger chemical entities from engineered cofactors. Such enlarged AdoMet mimics, in combination with a myriad of AdoMet-dependent M Tases available in nature, would provide useful molecular tools for targeted functionalization of biomolecules (38). So far, two chemistries have shown a good promise. First, derivatives of N-azidinoadenosyl-adenosine (60) and mechanistically related nitrogen mustards (61) (Fig. 2e) are covalently coupled to their natural targets in the presence of DNA M Tases (Fig. 3a). By attaching chemical groups to these cofactors, they were shown to work as delivery systems for various functional or reporter groups. An inherent feature of this system is potent product inhibition (single-turnovers) by the covalent cofactor-substrate conjugate. A second class of AdoMet analogs circumvents the problem of catalytic product release. In these analogs, the methyl group is replaced with an extended carbon chain that contains an activating double or triple bond (allylic and propargylic systems; Fig. 3b) (37). These cofactors were shown to confer catalytic M Tase-directed transfer of their activated side chains (Fig. 3b), permitting sequence-specific functionalization and labeling of plasmid DNA (62).

References


Further Reading


See Also

DNA, Covalent Modifications of Enzyme Co-factors, Chemistry of AdoMet-Dependent Methyltransferases

Chemistry of AdoMet-Dependent Methyltransferases
Chemistry of AdoMet-Dependent Methyltransferases

Enzyme Catalysis, Chemical Strategies for Post-Translational Modification, Regulating Protein Function by Proteins, in Vivo Chemical Modifications of Protein-Nucleic Acid Interactions Tags and Probes in Chemical Biology
The mid-chain dehydrogenation of saturated fatty acyl derivatives is carried out by a large family of O2-dependent, nonheme diiron-containing enzymes known as desaturases. Both soluble and membrane-bound desaturases have been characterized. The mechanism of desaturation is thought to involve the stepwise synthesis of vicinal hydrogen atoms via a short-lived carbon-centered radical intermediate. The most common desaturase inserts a (Z)-double bond between the C–9,10 carbons of a stearoyl thioester; however, many variations of this prototypical reaction have been discovered. Accounting for this diversity in terms of subtle alterations in active-site architecture constitutes a new frontier for research in this area.

The regioselective and stereoselective introduction of an olefinic link into fatty acyl side chains is catalyzed by a unique set of enzymes known as desaturases (1). The overall biochemical equation for this transformation can be depicted as follows:

\[
H^+ + NAD(P)H + O_2 + R-\text{CH}_2-\text{CH}_2R' \rightarrow NAD(P)^{+} + 2H_2O + R-\text{CH} = \text{CH-R}'
\]

Unlike other dehydrogenases, desaturases attack unactivated C–H bonds with the concomitant reduction of molecular oxygen; two reducing equivalents are derived from NADH, and two are derived from the substrate. Most fatty acid desaturases are integral, membrane-bound proteins found in the endoplasmic reticulum and accept substrates bearing phospholipid or Coenzyme A headgroups. In plants, a soluble, plastidial desaturase operates exclusively on acyl carrier protein ACP substrates. Common to all of these enzymes is a nonheme diiron-containing catalytic core (vide infra) that is used to generate the active oxidant in situ. Considerable insight into the detailed chemical mechanism of desaturation has been gained in the last 15 years (2). This knowledge has lead to a more sophisticated understanding of lipid biochemistry in various contexts, including chemical ecology, plant biotechnology, and medical lipidology.

Biologic Context

Fatty acid desaturation is a ubiquitous lipid modification that is critically important to aerobic life forms in two major ways:

1) adjustment of lipidic biophysical properties and 2) biosynthesis of chemical messengers. Saturated fatty acids produced by the fatty acid synthase assembly line, typically are dehydrogenated first at C–9,10 to give an oleate derivative that is additionally desaturated to generate the ω-6 and ω-3 group of the so-called “essential fatty acids” (Fig. 1a). Various highly bioactive signaling molecules (Fig. 1b) containing unsaturated sites are biosynthesized from this collection of primary fatty acids. In addition, desaturase-derived natural products with antifeedant properties have been discovered (Fig. 1c). It should be noted that the introduction of double bonds into biomolecules also dramatically enhances their susceptibility to free radical autoxidation with all attendant deleterious effects (3).

Unsaturated lipids and membrane fluidity

To function properly, cell membranes must exist primarily in the liquid-crystalline state. Model studies using synthetic lipids have demonstrated that the presence of a (9Z)-olefinic fatty acyl side chain in a diacylphospholipid lowers its gel-liquid phase temperature \(T_c\) by some 50°C relative to its fully saturated analog (4). The presence of the \(\Delta^9,12\) fatty acyl side chain lowers the \(T_c\) even more and constitutes an important adaptation to chilling for plants. The regulation of the response to chilling temperatures has been studied in some detail in various organisms (4).
Figure 1: Role of fatty acid desaturases in the biosynthesis of unsaturated fatty acids. Typical fatty acids found as components of cell membranes and storage lipids (a); signalling agents (b); and plant-derived antifeedants (c).

pheromone biosynthesis (5). Species-specific signals are generated by using a chemical language dictated by the number, position, and stereocchemistry of internal double bonds located in a hydrocarbon backbone of varying lengths. Increased differentiation is achieved through additional functionalization and the use of multiple components in specific ratios. The blend of (Z)- and (E)-11-tetradecenoes shown in Fig. 1b is a case in point. In the area of sphingolipid biochemistry, it has been discovered that the position and stereocchemistry of the double bond in ceramide is critically important to biologic function (6).

Plant defense

The involvement of desaturase-type enzymes in the biosynthesis of some important lipidic antifeedants now has been firmly established (Fig. 1c) (7). An instructive example of this phenomenon is the production of ricinoleate (a purgative) by a desaturase homolog found in the castor plant. Here, a subtle variation in mechanistic pathway is responsible for the introduction of a C-12 hydroxyl group rather than a 12,13-double bond (7). Another interesting case features the putative dehydrogenation of a cyclopropyl fatty acid to produce sterculate, a potent inhibitor of the mammalian but not the plant \( \Delta^9 \) desaturase (8). Finally, the biosynthesis of polyacetylenes with antifungal properties such as falcarindiol is thought to involve a sequence of unique desaturase-mediated oxidations (9).

Chemistry of Fatty Acid Desaturase

The soluble castor stearoyl-ACP \( \Delta^9 \) desaturase has served as an important enzyme for detailed structural and biochemical studies (10). A crystal structure of this protein confirmed the presence of a nonheme, carboxylate-bridged diiron cluster in close proximity to a narrow, curved hydrophobic channel (11). The latter can accommodate a stearoyl-ACP substrate that adopts a gauche conformation around the 9,10–carbon–carbon single bond. Delivery of electrons from nicotinamide adenine dinucleotide phosphate (NADPH) via ferredoxin as required by Equation 1 is gated by substrate binding—an event that alters the redox properties of the catalytic center and ultimately triggers oxygen binding. Generation of a potent oxidant then can proceed (12). It is assumed that a similar diiron-based cluster exists in membrane-bound desaturases; Detailed comparison of amino acid sequences for this class has revealed the presence of a highly conserved, catalytically essential 8-histidine motif that is capable of binding two iron atoms. In addition, M Gsobauer studies of a closely related alkane \( \omega \)-hydroxylase revealed the presence of two catalytically active iron atoms (13). However, reconstitution of purified membranous desaturases to permit more detailed structural study has been extremely difficult. Initial design and testing of mechanic probes was carried...
The cryptoregiochemistry of fatty acid desaturation

The term “cryptoregiochemistry” was coined to describe which substrate hydrogen is removed first in fatty acid desaturation (19). In the case of membrane-bound desaturases, the site of initial oxidation could be determined by examining the magnitude of the primary deuterium kinetic isotope effect (KIE) on C-H cleavage as a function of isotopic substitution. According to the mechanistic model (Fig. 2), abstraction of the first C-H bond should be energetically more difficult, and hence more sensitive to isotopic substitution, than the subsequent C-H bond cleavage step. Thus, a large KIE was observed at C-9 but not at C-10 in the desaturation of regiospecifically deuterated substrates by a yeast stearoyl CoA \( \Delta^9 \) desaturase (Fig. 3a) (19). Corroborating evidence for initial attack at C-9 by this desaturase was obtained by observing preferential sulfur oxidation of an S-9 substrate analog (Fig. 3b) (20). A additional proof for a C-9-initiated \( \Delta^9 \) desaturation was the observation of low level (1%) regioselective 9-hydroxylation along with the production of the major 9,10-olefinic product (14). The former pathway presumably occurs by hydroxyl trapping of the putative C-9 radical intermediate (Fig. 2).

The cryptoregiochemistry of a large number of membrane-bound desaturating systems with varying positional specificities has been determined (2), which include desaturases found in species of bacteria, blue-green and green algae, fungi, nematode, plants, insects, and mammals. The somewhat surprising trend that has emerged is that the carbon closest to the C-1 (acyl) terminus always is attacked first during these dehydrogenation reactions. This result points to a highly conserved active-site architecture that is used by desaturases from a wide range of life forms. The corresponding hydroxyl by-product (Fig. 2) with the predicted regiochemistry is observed routinely (14). The determination of the cryptoregiochemistry for the soluble castor \( \Delta^8 \) desaturase has been more difficult because the C-H bond cleavage is masked kinetically by other events in the catalytic cycle, which nullifies the use of an approach that relies on the measurement of an intermolecular KIE. Nevertheless, the results of experiments that use sulfur, oxygen, and fluorine-labeled substrates suggest that, in this case, the initial attack on C-H bonds is at C-10 rather than C-9 (10).

Variations on an oxidative theme: the oxygenation/dehydrogenation connection

Perhaps the most important mechanistic question that remains unanswered with respect to desaturases relates to the switch that controls the choice of dehydrogenation/hydroxylation pathways (Fig. 2). The study of the FAD2 subgroup of plant desaturases has been particularly instructive in this context. Pair
Chemistry of Fatty Acid Desaturases

Figure 3 Evidence for a stepwise removal of hydrogen by membranous stearoyl Δ⁹ desaturase (SCD) initiated at C–9: Primary deuterium isotope effect observed at C–9 but not at C–10 (a); Sulfoxidation of S–9 preferred over S–10 thio substrates (b).

Regiochemistry of desaturation

Regioselective remote functionalization of unactivated C–H bonds has long been a goal of synthetic chemists. Indeed, much early work in this area (26) was inspired by the ability of fatty acid desaturases to introduce double bonds in a regioselective manner. Three modes of regiocontrol have been identified (27, 28). The positional specificity of the Δ⁹ class of desaturases is determined by the location (carbons n, n + 1) of the incipient double bond relative to the acyl head group (C–1) independent of the chain length of the substrate. For example, both yeast and rat liver membrane-bound Δ⁹ desaturases dehydrogenate a range of fatty acyl CoA thioesters (C–15 to C–19) at the 9,10-position with no measurable regiochemical "error" and at comparable rates. Similar regioselectivity is observed for the corresponding soluble castor enzyme; but in this case, the C–18 substrate is a highly preferred substrate (10). Both the regioselectivity and the chain-length specificity of the latter enzyme can be altered through protein engineering experiments (1). The ω-n group desaturases inserts a double bond "n" carbons from the methyl terminus of substrate. Yet a third group of enzyme homologs that are tuned to catalyze either dehydrogenation or oxygenation as the major pathway have been identified (7). Thus, oleate Δ¹² desaturase introduces a double bond at C–12,13 by initial hydrogen abstraction at the C–12 position, as determined by a KIE study (21). However, an enzyme homolog, oleate 12-hydroxylase, forms 12-hydroxyoleate (ricinoleate), a major component of castor oil (Fig. 4a). The results of site-directed mutagenesis experiments that use a bifunctional Δ¹² oleate desaturase/12-hydroxylase demonstrated that relatively conservative changes in a few amino acids have a major impact on the hydration/desaturation ratio (7, 14). An even more dramatic example of this mechanistic dichotomy is the discovery of FAD2-type enzyme homologs in the Crepis species that can execute either the dehydrogenation or epoxidation of the C–12,13-olefinic bond (Fig. 4b) (22). Interestingly, the former reaction is initiated also at C–12, as indicated by a large primary deuterium kinetic isotope effect at this position (23). A similar trend in kinetic isotope effects is exhibited by an insect acetylenase (24). It is tempting to account for these observations in terms of "fine control" of substrate position relative to oxidant in the active site. Alternatively, subtle changes in the coordination chemistry at the catalytic centers, similar to that postulated for clavaminate synthase, could be responsible for the observed divergence in reaction outcome (25).
Chemistry of Fatty Acid Desaturases

Figure 4  Variations on an oxidative theme: the dehydrogenation/oxygenation connection. C-12-initiated oxidation catalyzed by oleate \( \Delta^{12} \) desaturase and its homolog oleate 12-hydroxylase (a) and \( \Delta^{12} \) acetylenase and its homolog \( \Delta^{12} \) epoxidase (b).

of desaturases, the \( \nu+n \) class uses an adjacent double bond as the primary reference point and introduces unsaturation "n" carbons from a preexisting double bond.

Chemical Tools and Techniques

The design of probes for the mechanistic study of fatty acid desaturases had to take into account that this class of enzymes cannot tolerate large alterations in substrate structure. Consequently, use of isotopic labeling and the isomorphic replacement of the -CH2- unit by the sulfur atom or CHF moiety have been the most versatile approaches. Additional challenges were the high endogenous lipid content of membrane-bound desaturases and the need to enzymatically prepare substrate ACP thioesters in the study of soluble desaturases (16). To understand the bioinorganic chemistry of desaturases, a new methodology for probing nonheme iron-containing enzymes also was required (29).

Stereospecific and regiospecific isotopic labeling

To determine the stereochemistry of desaturase-mediated hydrogen removal, the synthesis of stereospecifically monodeuterated fatty acid substrates and a suitable gas chromatography–mass spectrometry (GC–MS) or liquid chromatography–mass spectrometry (LC–MS) method of product analysis is needed (2). The most common method of introducing deuterium stereospecifically is through deuteride displacement of suitable activated chiral alcohols. Occasionally, the latter compounds are available from natural sources; however, when they are not available from natural sources, total synthesis of substrate is necessary. In addition, the frequent problem of mass spectral interference because of high endogenous di-product content must be overcome through additional remote mass labeling of substrate using deuterium, sulfur, or fluorine substitution.

An examination of the primary deuterium kinetic isotope effect on each C–H bond cleavage involved in desaturation can reveal the site of initial oxidation, provided other kinetically more important enzymic steps do not mask these effects. This methodology relies on monitoring the loss of deuterium from regiospecifically deuterated fatty acids via GC–MS. General strategies for the synthesis of the required labeled compounds were devised during early work on assignments of 13C nuclear magnetic resonances (NMRs) for various fatty acids (30).

Use of fluorine as a probe and as a tag

Monofluorinated fatty acids function as fatty acyl substrates for desaturases and can be synthesized in enantiomerically enriched form via diethylaminosulfur trifluoride (D.A.S.T.) treatment of the suitable chiral alcohols. An example of the use of fluorine-substituted fatty acids as mechanistic probes was demonstrated by the use of chiral 9-fluorostearoyl substrates to induce latent stereoselective 10-hydroxylation by the castor stearoyl ACP \( \Delta^{9} \) desaturase (16). In addition, the advantages
of $^1$H-decoupled $^19$F-NMR, such as wide chemical shift range, high sensitivity, and lack of interferences, allow one to monitor desaturase-mediated transformations at $\mu$Molar fluoro-substrate concentrations. In this context, $\omega$-fluorine-tagged fatty acid analogs have proven extraordinarily useful in tracking remote functional group transformations (2).

Use of sulfur as a methylene isostere

This fatty acid analogs are synthesized easily by thiolation of the corresponding bromoacids. When the sulfur atom in this substate is remote from the site of desaturation, normal olefinic this products are produced, with the observed chemoselectivity being a consequence of a strict, desaturation-imposed, regiochemical imperative. Thia compounds also can be used to determine the site of initial oxidation in desaturase-mediated reactions because oxo transfer occurs most efficiently when the sulfur atom is located at the site of initial oxidation (20). A as a bonus, the enantiospecificity of the oxidant can be ascertained by determining the absolute configuration of the resultant dialkyl sulfides (20). It was found that the stereochemistry of sulfoxidation matched that of hydrogen removal for the parent substrate. These analyses can be carried out on a microscope by using $\omega$-fluorine-tagged thia derivatives and suitable Pirkle-type NMR shift reagents (31).

Future Directions

Currently, efforts are underway to obtain more structural information on desaturases to address the mechanistic issues that have been raised through substrate-based studies. The need for more detailed 3D active-site information is acute, particularly in the case of membrane-bound desaturase for which only hypothetical models currently are available. Design of mechanism-based inhibitors of medically relevant desaturases such as stearoyl-CoA $\Delta^\text{9}$ desaturase (SCD) (metabolic syndrome) (32), DesA3 (tuberculosis) (33), and dihydroceramide desaturase-mediated transformations at $\mu$Molar fluoro-substrate concentrations. In this context, $\omega$-fluorine-tagged fatty acid analogs have proven extraordinarily useful in tracking remote functional group transformations (2).

Use of sulfur as a methylene isostere

This fatty acid analogs are synthesized easily by thiolation of the corresponding bromoacids. When the sulfur atom in this substate is remote from the site of desaturation, normal olefinic this products are produced, with the observed chemoselectivity being a consequence of a strict, desaturation-imposed, regiochemical imperative. Thia compounds also can be used to determine the site of initial oxidation in desaturase-mediated reactions because oxo transfer occurs most efficiently when the sulfur atom is located at the site of initial oxidation (20). A as a bonus, the enantiospecificity of the oxidant can be ascertained by determining the absolute configuration of the resultant dialkyl sulfides (20). It was found that the stereochemistry of sulfoxidation matched that of hydrogen removal for the parent substrate. These analyses can be carried out on a microscope by using $\omega$-fluorine-tagged thia derivatives and suitable Pirkle-type NMR shift reagents (31).

References


Further Reading


See Also

Oxygen-Activating Enzymes, Chemistry of GC-MS of Lipids

Membranes, Fluidity of Natural Products: An Overview

Organic Chemistry in Biology
All cells must respond appropriately to external stimuli, including nutrient availability, hormones, and growth factors. A main focus of biologic research is aimed at elucidating the molecular mechanisms of how signals are relayed within cells. Protein phosphorylation is a central mechanism of signal transduction in eukaryotes and is involved in nearly all regulated cellular processes. Two families of enzymes control the phosphorylation state of cellular proteins: protein kinases catalyze phosphorylation of either serine/threonine or tyrosine residues of target proteins using ATP as the phosphodonor, whereas protein phosphatases catalyze removal of the phosphate group from phosphoproteins (1). Thus, phosphorylation of cellular proteins is a reversible switch that controls the enzymatic activity, cellular localization, conformation, or macromolecular binding capacity of the target protein, ultimately culminating in cellular responses to a variety of stimuli. The families of protein kinases and phosphatases are large: the human genome contains ~600 protein kinases and ~150 protein phosphatases. Similarly, the extent of protein phosphorylation is immense; it is estimated that one third of all human proteins are modified by phosphorylation (2, 3). Thus, understanding protein phosphorylation is a major challenge for biologic research and also has implications for developing therapies to treat the many human diseases that originate from misregulated signal transduction. In this article, I will discuss the chemistry and enzymology of the phosphorylation/dephosphorylation reactions catalyzed by protein kinases and phosphatases as well as describe some chemical tools that have been developed to study features of protein phosphorylation that otherwise resist analysis.

### Enzymology of Protein Kinases

**The phosphotransfer reaction mechanism of protein kinases**

Protein kinases catalyze transfer of the \( \gamma \)-phosphate from ATP to hydroxyl side chains of target proteins. This phosphotransfer reaction lies somewhere on a continuum between two limiting mechanistic possibilities (4). The **associative** mechanism is analogous to the \( \text{Sn2} \) reaction, where the hydroxyl group of the phosphoacceptor protein attacks the \( \gamma \)-phosphate of ATP, followed by the formation of a \( -3 \) charged trigonal bipyramidal transition state, which is resolved then into ADP and phosphoprotein (Fig. 1a). This trigonal bipyramidal transition state has two positions around the phosphorous center. \textit{Apical} positions are long, weak bonds \( 180^\circ \) apart from each other where groups enter or depart, and \textit{equatorial} positions are shorter, stronger bonds \( 120^\circ \) apart from each other. A dissociative phosphotransfer reaction can be \textit{in-line}, where the incoming nucleophile approaches from the backside of the scissile bond (as in a classic \( \text{Sn1} \) reaction), or \textit{adjacent}, where the nucleophile approaches the backside of a nonscissile bond. The second limiting mechanism is dissociative phosphotransfer, where the \( \gamma \)-phosphate ionizes (as in an \( \text{Sn1} \) reaction) to form a \( -1 \) charged metaphosphate intermediate that subsequently is captured by the hydroxyl nucleophile (Fig. 1b).

Phosphotransferases are ubiquitous in biology and have varied mechanisms. To determine whether the phosphotransfer
Chemistry of Protein Kinases and Dephosphorylases

(a) The limiting case of purely associative phosphotransfer involves attack of the hydroxyl on the phosphoanhydride followed by release of ADP. (b) The alternative limiting case of purely dissociative phosphotransfer possible in kinase reactions involving ionization of the phosphoanhydride bond followed by capture of the metaphosphate intermediate by the hydroxyl nucleophile. (c) Structure of a transition state mimic of protein kinase A. The phosphoacceptor peptide is yellow with the nucleophilic serine in blue; the AlF₃ mimic of the γ-phosphate is gray. The structure shows that the β-phosphate–aluminum–serine internuclear geometry is in-line, and the internuclear distance suggests the mechanism is closer to the associative end of the mechanistic spectrum.

Catalysis by protein kinases

How protein kinases facilitate phosphotransfer is well understood from structural studies. The overall structure of the catalytic domain of protein kinases is bilobal, consisting of a small, mostly β-sheet subdomain at the N terminus connected by a short linker to a larger, mostly α-helical C terminus. The N-terminal domain binds Mg:ATP, whereas the C terminus binds the phosphoacceptor protein substrate (Fig. 2a). The detailed architecture and catalytic residues of the active site are very highly conserved among the family. Much structural biology analysis has been performed on protein kinase A (PKA), and its catalytic residues are conserved across the family (6, 7). In PKA, lysine 72 and glutamate 91 orient the γ-phosphate toward the protein substrate (Fig. 2b). Aspartate 166 acts as a catalytic base to accept the proton from the hydroxyl nucleophile, and Lys 168 acts as an electrostatic catalyst to stabilize the γ-phosphate during the reaction. Asparagine 171 positions a magnesium ion that coordinates the α/β phosphates (Fig. 2c).
Enzymology of Protein Phosphatases

Phosphatases hydrolyze phosphate groups from phosphoproteins, thereby reverting the action of protein kinases. Two classes of protein phosphatases exist with distinct structures, substrate specificities, and mechanisms: protein serine/threonine phosphatases and protein tyrosine phosphatases. A few dual-specificity phosphatases also are known that are similar in structure to protein tyrosine phosphatases.

Protein tyrosine phosphatases (PTPs) cleave phosphate groups from phosphotyrosine residues. As is the case with protein kinases, these enzymes are well characterized structurally.

Protein tyrosine phosphatases specifically recognize phosphotyrosine because of the presence of a P-loop binding motif, with the general sequence HCPxxxxxxR(S/T) (8). The P-loop arginine and several backbone amide nitrogens are the primary contacts to the phosphate oxygens through hydrogen bonds. The active site pocket is also deep, which disallows proteins phosphorylated on serine to access the P-loop. Hydrophobic side chains also pack the aryl moiety of phosphotyrosine to provide additional selectivity versus phosphoserine.

The mechanism of the PTP hydrolysis reaction has two steps. First, phosphate is transferred from tyrosine to the cysteine residue of the P-loop, which generates a phosphoenzyme intermediate with concomitant release of tyrosine. This process is followed by hydrolysis of the phosphoenzyme to free enzyme and inorganic phosphate. Two active site residues are of primary importance during the catalytic cycle: the nucleophilic cysteine and the general acid/base catalyst. After attack of the cysteine on phosphotyrosine, tyrosine can be expelled as the cysteine thiolate, with concomitant release of tyrosine. The aspartate also has a lowered pKa value, which amplifies its nucleophilicity. Measurement of the pH profile of the rate of alkylation of this cysteine with iodoacetate has determined its pKa to be as low as ~5 (normally 8.3), which makes the residue ionized at physiologic pH (10). The aspartate also has an anomalous pKa value and is believed to be protonated at the beginning of the reaction coordinate, as evidenced by cocrystal structures of multiple protein serine phosphatases.

Resolution of the phosphoenzyme intermediate into free enzyme and inorganic phosphate is accomplished by orientation and deprotonation of water by the aspartate residue and attack at the phosphophosphate. Mutagenesis of the aspartate residue to alanine (as well as the Arg in the P-loop) shows its importance in this second step, as it leads to accumulation of the phosphoenzyme intermediate (12).

As was the case with other phosphotransfer reactions, the hydrolysis reaction catalyzed by PTPs lies along a mechanistic continuum between the limiting cases of associative and dissociative mechanisms. In the case of Yersina PTP, a crystal structure was solved with nitrate, a mimic of the transition state with exogenous thiol, which explains the role of peroxide inactivating several PTPs and is reversible by treatment with exogenous thiol, which explains the role of peroxide use as a rapid diffusable signaling molecule (13).

Protein serine/Threonine phosphatases

Protein phosphatases that are specific for phosphoserine or phosphothreonine have a distinct reaction mechanism from tyrosine phosphatases. Protein serine phosphatases are transition metal-dependent, and the reaction mechanism does not involve a phosphoenzyme intermediate as in the case of PTPs. Crystal structures of multiple protein serine phosphatases have revealed how the enzymes catalyze hydrolysis of phosphoserine (14).

Serine phosphatases are metalloenzymes with two transition metals (Fe3+ and Zn2+–Mn2+) in PP2B and PP1, or 2 Mn2+ in PP1B.
Chemistry of Protein Kinases and Dephosphorylases

(a)

(b)

Figure 3 Structure of protein serine phosphatase PP2C with bound phosphate showing the binuclear metal site (purple spheres) and the nucleophilic water (red sphere). Also shown are residues that ligate the transition metals as well as two key asparagines that coordinate the phosphate group.

Figure 4 Protein kinases with the “gatekeeper” residue mutated to glycine that makes them recognize N6-substituted ATP analogs also programs them to be uniquely sensitive to designed inhibitors, allowing rapid generation of potent selective kinase inhibitors of any protein kinase.

PP2C) coordinated by Asp, Asn, and His residues to form a binuclear metal center in the active site. This metal center is bridged by a well-ordered water molecule. In the proposed hydrolysis mechanism, the metal-bound water acts as the nucleophile to attack the phosphorous center of phosphoserine/threonine. The metal centers contribute by Lewis acid catalysis to lower the pKa of the bound water and to enhance its nucleophilicity. The metal's in some cases (PP1 and PP2B) also coordinate the phosphate of pSer/pThr, which would serve to enhance the electrophilicity of the phosphate. A second metal-bound water has been proposed to perform general acid base catalysis to shuttle protons to the departing serine hydroxyl. The phosphate of the substrate is positioned by hydrogen bonding to Arg33 residue as well as metal bound waters (Fig. 5).
Chemical Tools to Study Protein Phosphorylation

Chemical tools have played a central role in elucidating the function of posttranslational protein phosphorylation. Kinase and phosphatase inhibitors have played a major role in the study of protein phosphorylation because they allow rapid, reversible inactivation of the target in meaningful contexts (cells and organisms) to study its biologic function. Chemical tools have also helped identify the direct substrates of individual kinases and phosphatases, and conversely, they have helped identify which kinase is responsible for phosphorylating a particular phosphoprotein. Here I will focus on chemical tools developed for protein kinases; several reviews discuss chemical tools for protein phosphatases (15, 16).

Engineering protein kinases to label direct protein substrates

A main question regarding protein phosphorylation is what the direct targets of any given protein kinase are, within a signal transduction event. If all phosphotransfers could be mapped, the sequence of events that lead to a host of important biologic functions would be known, which not only would be important for a full understanding of signal transduction but also would enable the rational design of drugs that target signaling systems with predictable outcomes. This information is very difficult to obtain because many kinases work in concert during a signaling event and phosphorylate many different proteins.

A chemical method was devised to address directly the issue of determining the direct substrates of any given protein kinase (17). The idea is centered on generating ATP analogs (A*TPs) that can function as phosphodonor in a kinase reaction only with a protein kinase that has been engineered properly to accept the modified nucleotide. Thus, the addition of radiolabeled A*TP to a cell lysate that contains one such A*TP-sensitive protein kinase would result in selective radiolabeling of its direct substrates (Fig. 6a).

Several requirements of an unnatural A*TP analog and engineered kinase exist to have substrate labeling capabilities. First, the A*TP must not be a substrate for any endogenous kinases. Second, a mutation in the ATP binding site of a kinase must be discovered such that A*TP is now used in the phosphotransfer reaction with good efficiency. Finally, the mutation that expands the nucleotide specificity of the kinase should be functionally silent, so the mutant kinase performs its roles normally, with the single exception of accepting A*TP for substrate labeling purposes.

Toward this goal, several radiolabeled N6-substituted A*TP analogs were synthesized and tested for their ability to label kinase substrates in a cell lysate rich in active kinases. Several bulky N6 substituents rendered the ATPs "orthogonal" to the set of all wild-type kinases, such as N6-cyclopentyl ATP and N6-benzyl ATP. Examination of available ATP:kinase cocrystal structures then allowed design of mutations that would allow these substrates to bind productively to the active site. One residue in the link region between the small N-terminal and larger C-terminal subdomains (M120 in protein kinase
Chemistry of Protein Kinases and Dephosphorylases

Figure 6  Scheme of the method to identify direct protein kinase substrates using the unnatural nucleotide N^6-(benzyl) ATP and kinases engineered to accept them by mutation of the gatekeeper residue (T338G in c-Src). Introduction of either the engineered kinase or γ-32P labeled ATP analog alone results in no substrate labeling, but in combination, the direct substrates of the engineered kinase are radiolabeled. (b) The crystal structure of wt c-Src and ADP shows direct contact between the Thr338 residue (red) and the N^6 position of ATP, precluding catalytic activity with N^6-substituted ATP analogs. (c) The crystal structure Thr338Gly c-Src and N^6-(benzyl) ATP shows the this mutant exposes a hydrophobic pocket that allows recognition of the unnatural nucleotide.

A) directly contacts N^6 of ATP and is conserved large hydrophobic residue. This hydrophobic residue serves as a “gatekeeper” to a larger hydrophobic pocket behind it (Fig. 6b). Point mutation, in which this residue is changed to glycine, allows the bulky N^6 benzyl/cyclopentyl substituent to access the hydrophobic pocket. This action rendered the mutant kinases efficient catalysts with the N^6-substituted ATPs and only modestly changed their catalytic efficiency with ATP, thereby making the mutation functionally silent. Most impressively, the mutation that sensitizes kinase toward ATPs is portable to nearly all kinases in the superfamily and in multiple organisms. Numerous groups have used this system to identify kinase substrates from many different signaling pathways in several organisms.

Specific kinase/phosphatase inhibitors as chemical tools

Perhaps the largest contribution that chemistry has made to the study of kinase-mediated signaling is the development of chemical inhibitors of protein kinases and phosphatases. Kinase inhibitors are under intense investigation as drugs after the spectacular success of the Abl tyrosine kinase inhibitor Gleevec as a treatment for chronic myelogenous leukemia. Because of this interest in drug development, many inhibitors of kinases are known and used as biochemical tools to inactivate rapidly kinases of interest. However, chemical inhibitors are useful to study the biology of kinases and phosphatases only if the compounds are selective for one member of the larger family. If compounds are monospecific, their biologic effects can be assigned unambiguously to the inhibition of a single enzyme. For protein kinases, however, truly selective inhibitors have been difficult to discover, and because the family is so large, it has been very difficult to confirm what the selectivity of commonly used inhibitors actually is. A recent comprehensive survey of the actual specificity of commonly used kinase inhibitors showed that many compounds are not as selective as previously thought and that many have more potent targets than the kinase they are commonly used against (18). This article highlights the need for strategies that preserve the advantages of using small molecule inhibitors to probe kinase function, such as speed, ease, and tunability, with stringent verification of target specificity. The difficulty in generating specific kinase inhibitors stems from the fact that most inhibitors target the kinase ATP binding site. This site is conserved highly among the family, which makes development of specific compounds difficult.

Toward this goal, a chemical-genetic strategy for developing specific kinase inhibitors was developed (19). The “gatekeeper” residue that when mutated to alanine or glycine sensitizes protein kinases toward unnatural ATPs also was exploited to generate kinases that are sensitive to inhibitors with similar bulky substituents oriented toward the gatekeeper residue. In this way, cells carrying the mutant kinase would be uniquely sensitive to inhibitors that require binding the hydrophobic pocket adjacent to the gatekeeper. The relatively nonselective inhibitor scaffold PP1 was synthesized with bulky substituents and ring expansions that were designed to contact the gatekeeper residue (Fig. 3). It has been shown that virtually any kinase can be “programmed” to be sensitive to these bulky inhibitors by mutating

Figure 7  Structure of ATP-peptide substrate linked “bifunctional” kinase inhibitors are a general strategy to generate specific inhibitors of any protein kinase whose peptide substrate specificity is known.
Chemistry of Protein Kinases and Dephosphorylases

a single amino acid. The selectivity of these compounds toward "analog-sensitive" kinase alleles is remarkable (∼1000 fold versus wild type kinases), as is their affinity, with IC50 values in the low nanomolar range. Furthermore, because inhibition requires a specific genetic background, for example, mutation of the gatekeeper residue, off-target effects can be controlled for by comparison with drug-treated cells with an intact gatekeeper. A more recent general strategy to generate specific kinase inhibitors involves generating compounds that probe the phosphoacceptor peptide binding groove between the small and large lobes of the catalytic domain (28). Oftentimes kinases are selective for specific peptide substrates, so exploitation of this selectivity can yield specific kinase inhibitors. Inhibitors were generated by linking the ATP moiety to kinase-specific peptide substrates to yield "bifunctional" inhibitors that capture interactions in the ATP and peptide substrate binding sites (Fig. 7). In one such study, targeted against the insulin receptor tyrosine kinase domain, γ-thio ATP was linked to an optimized IRK peptide substrate, which creates a nucleotide–peptide conjugate that has a submicromolar K_i (370 nM). The bifunctional inhibitor also shows selectivity toward IRK, as the K_i of the conjugate for the tyrosine kinase Csk was 100-fold higher. This strategy is suitable for any kinase whose peptide substrate specificity is known and has been applied for many diverse kinases.

Chemically cross-linking kinases and their substrates

Another chemical biology tool generated to study protein kinase signaling is a method to identify a kinase responsible for phosphorylating a particular phosphoprotein. Two methods have been devised by designing ATP molecules that form covalent bonds to both the kinase and substrate, which allows identification of the unknown partner by mass spectrometry.

Figure 8. Two strategies to cross-link kinases with their substrates, involving either photo-cross-linking groups (a) or a mechanism-based annihilation reaction (b).
In the first case, ATP analogs that have two azido groups attached at the purine and gamma-phosphate were synthesized (21). Such a molecule can bind the ATP binding site of a given kinase and allows its physiologic substrate to bind the bisazo-ATP kinase complex, which brings both kinase and target in close proximity to the azido groups. UV irradiation then uses each azide to form reactive nitrenes, which form a covalent ATP bridge between the kinase and substrate (Fig. 8a).

An alternate strategy to identify the upstream kinase responsible for phosphorylating a particular phosphoserine protein also was reported (22). In this scheme, a bis azide ATP analog was designed to form an imine with the lysine (Lys 72 in PKA) essential for orienting the αβ-phosphates of ATP. The substrate of interest, with its target serine mutated to cysteine, provides a nucleophile to attack the imine, which then cyclizes with a second azide to form a stable isoindole adduct that links kinase and substrate (Fig. 8b).

References


Further Reading

Walsh CT. Posttranslational Modification of Proteins: Expanding Nature’s Inventory. 2006. Roberts and Company, Greenwood Village, CO.

See Also

Cellular Communication by Signal Transduction Mitogen-Activated Protein Kinases (MAPKs): ERKs and JNKs Selective Inhibitors of Kinases Small Molecules to Elucidate Kinase Pathways
O-GlcNAcylation is a prevalent posttranslational modification that occurs when a molecule of the monosaccharide N-acetylglucosamine (GlcNAc) is adjoined to a serine or a threonine residue of a cellular protein. This modification can influence a protein in many of the same ways that a serine or threonine phosphorylation can. In addition, much like the kinases and phosphatases involved in phosphorylation/dephosphorylation, unique enzymes are responsible for the addition and removal of O-GlcNAc modifications, with an O-GlcNAc transferase that catalyzes the addition and an O-GlcNAcase that is responsible for the removal. This review will give a summative description of the diverse biochemical roles that this O-GlcNAcylation can play on individual proteins, and on whole animal systems, and will detail the current understanding of the enzymes involved in its maintenance. Finally, we discuss here the recent advances that have been made in the development of tools to study O-GlcNAcylation.

O-GlcNAcylation, which was first described just over 20 years ago, is a ubiquitous posttranslational modification, whereas a sugar molecule, N-acetylglucosamine (GlcNAc), is added to precise serine or threonine residues on cellular proteins. This modification can alter the physical properties of a protein in several ways, and it is highly analogous to posttranslational phosphorylation. The study of, and the understanding of, phosphorylation is well established, and in recent years, it has resulted in many advances in the comprehension of several proteins that are O-GlcNAc modified, as well as how this modification influences individual protein activities. The last five years have also brought about a wealth of biochemical information that concerns the enzymes accountable for the regulation of this modification, the O-GlcNAc transferase and the O-GlcNAcase, which includes their domain structures and the characterization of their catalytic activities. The mapping of O-GlcNAcylation sites has also experienced highly significant progress. Although O-GlcNAcylation still lags behind phosphorylation in popularity, this profusion of information is leading to the firm establishment of the importance of O-GlcNAc in cell and molecular biology.

Biologic Background

When glucose enters a cell, its primary fate is to be phosphorylated and converted to glucose-6-phosphate. The glucose-6-phosphate is then converted to fructose-6-phosphate, whereby it awaits one of several fates. The greater percentage will eventually be converted either to glycogen for storage in skeletal muscle and in liver or broken down for energy in the glycolytic pathway. A approximately 2-5% of the fructose-6-phosphate, however, enters the hexosamine biosynthetic pathway (HBP), where it ultimately will be converted to the high energy compound UDP-N-acetylglucosamine (UDP-GlcNAc). These sugar residues are linked commonly to proteins in two different ways. N-linked glycosylation occurs when the attachment is made to the amide nitrogen of asparagine side chains. N-linked glycosylations contain a minimum of four sugars in addition to the terminal GlcNAc, and this modification typically is implicated in the directing of proteins through the endoplasmic reticulum–Golgi–plasmalemma pathway. The second fate of the UDP-GlcNAc is its O-linkage to proteins (Fig. 1).

O-GlcNAc as a posttranslational modification

First described in 1984, O-linked GlcNAcylation is a form of posttranslational modification where a single sugar, the monosaccharide GlcNAc, is affixed via a beta linkage to the hydroxyl group of specific serine and threonine residues on target proteins (1). In rare cases, O-linked N-acetylglucosamine (O-GlcNAc) linkage can occur on lysine side chains. Although this modification has not been witnessed in bacteria, it is otherwise ubiquitous in all eukaryotic cells that have been studied.
When entry into a cell occurs, glucose is converted to glucose-6-phosphate (Glc-6-P) and then into fructose-6-phosphate (Fruct-6-P). Approximately 5% of the fructose-6-phosphate enters the hexosamine pathway, where the enzyme GFAT converts it to glucosamine-6-phosphate (GlcN-6-P), in a process that requires glutamine’s (Gln) conversion to glutamate (Glu). Eventually, the glucosamine-6-phosphate is converted to UDP-GlcNAc, and this can be added to protein serines or threonines by the enzyme O-GlcNAc transferase (OGT). O-GlcNAcase is the enzyme that removes this modification.

O-GlcNAcylation can occur on both extracellular and intracellular proteins and can be found on both cytosolic and nuclear members of the latter (1). O-GlcNAc additions are reversible, highly dynamic, and occur on proteins involved in most, if not all, processes of fundamental importance to a cell, including transcription, DNA replication, protein-folding, cytoskeletal regulation, translation, metabolic and signaling pathways, nuclear import-export, protein degradation, and vesicular trafficking, among many others. In addition to these downstream roles, the HBP is speculated to act as a global cell nutrient sensor. Given the makeup of the substrate UDP-GlcNAc, its levels are governed by the availability and the metabolism of amino acids (glutamine is used as the amine donor), fatty acids (which supply the eventual acetate), nucleotides (for UDP), and glucose in particular (for the sugar backbone). The global accumulation of O-GlcNAcylated proteins, as a result of an increase in glucose uptake and therefore UDP-GlcNAc availability, may act as a negative feedback, directly or indirectly, to downregulate proteins involved in additional glucose import (2, 3). This may be the case with catabolism and anabolism of the other constituents of UDP-GlcNAc as well.

O-GlcNAcylation in disease states

As a result of all of the above, the deregulation of protein-O-GlcNAc homeostasis has been linked to several disease pathologies, particularly diabetes, cancer, and Alzheimer’s disease. For example, increases in protein-O-GlcNAc—either by increasing the flux through the HBP, by overexpressing the enzyme responsible for its addition to proteins (N-acetylglucosaminyl transferase or OGT), or inhibiting the enzyme responsible for its removal (O-GlcNAcase)—can recapitulate several hallmark features of type II diabetes (OGT and O-GlcNAc case will be discussed). In fat or muscle tissue, increased O-GlcNAc leads to insulin resistance and to the hyperinsulinemia associated with increased fat synthesis and storage (4). In liver cells, an increase in O-GlcNAc results in an upregulation of glycogen storage, impaired glucose tolerance, and whole animal hyperlipidemia and obesity (5). In pancreatic β-cells, inflations in O-GlcNAc produce apoptosis (6), hyperinsulinemia, and insulin resistance (7); this inflation is postulated to be primarily because many proteins involved in the insulin-signaling pathways are O-GlcNAc-ed, including IRS-1, Akt, and PI-3 kinase. Finally, the O-GlcNAcase gene is located in a characterized diabetes susceptibility locus, and a naturally occurring single nucleotide polymorphism discovered within this gene was found to correlate with an increase in diabetic risk (8).

Second, because several proteins have been demonstrated to be O-GlcNAc-ed in a cell-cycle-dependent manner, and because cellular O-GlcNAc levels are highly correlative to the cell cycle (9), as well as in response to mitogens, growth factors, and cell stresses (see review by Zachara and Hart (10)), the disruption of O-GlcNAc pathways can be tumorigenic. Altering flux through the hexosamine pathway, providing more O-GlcNAc substrate for target proteins, has been shown to modify the growth rate of various cell types (11). In addition, overexpression of OGT, which results in more protein-O-GlcNAc, can result in polyplody and defects in cytokinesis. Overexpression of the reverse enzyme, O-GlcNAcase, influences cell cycle, which results in mitotic exit delays, altered cyclin expression, nuclear morphological differences, and delayed mitotic phosphorylation (9). Thus, cancerous phenotypes can result from faulty O-GlcNAcylation.

Finally, much evidence indicates that O-GlcNAc signaling is involved in neurodegenerative diseases. The potential role of...
O-GlcNAc in the pathogenesis of Alzheimer’s disease, in particular, has drawn much attention. The OGT and O-GlcNAcase enzymes are highly enriched in the brain and are particularly high in the Purkinje cells. The pathogenesis of Alzheimer’s disease involves, among other events, the accumulation of β-amyloid and tau aggregates, both of which have been shown to be O-GlcNAc modified and increasingly are modified in Alzheimer’s striken cells (12). Their ability to be degraded and cleared by the proteosomes in hippocampal cells is inhibited by increased O-GlcNAc levels (13, 14). In addition to being a hot spot for diabetes susceptibility genes, the chromosomal locus for the O-GlcNAcase gene is a candidate locus for late-onset Alzheimer’s disease. Interestingly, the locus for the single OGT gene is very near the locus responsible for dystonia-Parkinsonism syndrome.

The Roles of an O-GlcNAc Modification

Although the explanation of O-GlcNAc as a posttranslational modification has taken time (it has been just over 20 years since its discovery), the rate of its appreciation is similar to that of phosphorylation after its discovery. Hundreds of proteins have been demonstrated to be O-GlcNAc modified (Table 1), with a great many more as yet untested that possess potential O-GlcNAcylation sites. The roles that an O-GlcNAc modification can play on a protein have been found to be comparable to those of O-phosphorylation. Although O-GlcNAc is a neutral molecule, and does not possess the strong negative charge of a phosphate, its possession of numerous negatively polar hydroxyl groups allows it to perform many of the same reversible functions.

The addition of O-GlcNAc can alter protein-protein interactions. For example, adding it to the ubiquitous transcription factor Sp1 can inhibit its binding to TATA110 and can prevent gene transcription (15), and its addition to the chaperone HSP60 can obstruct HSP60-Bax interaction, which results in Bax translocation to the mitochondria in initiating steps of the apoptotic response (16). Conversely, O-GlcNAc modification on the signal transducer Stats can enhance the interaction of Stats with the transcriptional coactivator CBP (17). Protein-DNA interactions can also be influenced by an O-GlcNAc addition or loss, as is observed, for example, on the PDX-1 transcription factor, where increased O-GlcNAcylation on PDX-1 increases its DNA binding affinity (18). Within the transcription factor Oct-2, O-GlcNAcylation likewise regulates DNA-binding specificity (19).

O-GlcNAc modifications can also lead to the stabilization of a target molecule. O-GlcNAcylation of piklgothion, which is a component of adherens junctions and desmosomes, increases the stability of the protein and thereby promotes cell-cell adhesion (20). Another example occurs with eff2-p67 complexes, where an O-GlcNAcylation event on p67 leads to an increased half-life of these components, which allows the subsequent activation of the eff2 subunit required for protein translation (21). Furthermore, O-GlcNAc can contribute directly to global protein stability by modifying the ATPase subunit Rpt2 within proteasome caps, which results in the inhibition of proteosomal degradation (14).

Like phosphorylation, O-GlcNAc modification can activate or inactivate a protein through conformational changes. It is suggested that O-GlcNAc modifications on Stats are involved in its transient activation (22), and its addition to p53 is expected to increase the protein’s outside-in signaling necessary for cytoskeletal rearrangements and cell spreading (23). Conversely, O-GlcNAc modification can inhibit the signaling activity of PLC-β1, which is an enzyme involved in the generation of important second messengers that trigger intracellular Ca2+ release (24) or can result in the inactivation of the nitric oxide synthase eNOS (25).

Along with the above phenomena, the addition of O-GlcNAc influences a target protein’s cellular localization. According to several studies, it seems that O-GlcNAc may act as a nuclear localization signal. O-GlcNAcylated proteins seem to be more prominent in nuclear over cytosolic fractions (26), and proteins such as Akt1, mTOR, α4, and Sp1 translocate to the nucleus after O-GlcNAc modification. Exceptions exist, as witnessed with a viral protein Vpr, where O-GlcNAc added near the NLS influences nuclear import (27) negatively. The nuclear pore is considerably O-GlcNAc modified, and this can influence the transport into or out of the nucleus. The direct influence of O-GlcNAc on pore components is still being explored. Finally, because this modification occurs on serine and threonine residues, it can also compete directly with an O-phosphorylation event at an identical or nearby site to provide an additional level of regulation on a specific protein. To date, all described O-GlcNAc-modified proteins have been found to be phosphoproteins (28, 29, 30). Such competition between O-GlcNAc and O-phosphate binding occurs with many examples listed above as well as with others.

The Enzymes that Regulate O-GlcNAcylation

O-GlcNAc transferase

The enzyme that catalyzes the reaction between the target protein and the UDP-GlcNAc, with UDP being the leaving group, is OGT. OGT is a highly conserved, 1036-amino-acid protein found in all metazoa, and it is expressed ubiquitously in all tissues, although mRNA and protein levels can vary. OGT can be found in both the nucleus and the cytosol; however, localization studies have revealed a slight preference for the nucleus (31). The OGT protein is encoded by a single, highly conserved gene mapped to a location on human chromosome Xq13.1. Murine as well as cell knockouts have proven to be lethal, which indicates that the ability to glycosylate proteins is essential to complete embryogenesis and overall individual cell viability (32, 33). Conversely, overexpression of OGT is also toxic to cells, which suggests that, in addition to the necessity of the presence of O-GlcNAc modification, the maintenance of proper protein-O-GlcNAc levels is also essential for cellular function.
Cytosolic Glycosylated Proteins, Chemistry of

Table 1 Current listing of identified O-GlcNAc modified proteins

<table>
<thead>
<tr>
<th>Transcription Factors</th>
<th>Enzymes</th>
<th>Cellular/Vesicular Transport</th>
<th>Cytoskeletal/Structural Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>survivin, p53, P21, GFP</td>
<td>OSGT</td>
<td>cell-surface HLA, CCL-5, CR2, CD45</td>
<td>Vimentin, tubulin, fascin, filamin</td>
</tr>
<tr>
<td>O-GlcNAc modified c-Jun</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Survival.** O-GlcNAc transfer is mediated by two C-terminal catalytic domains (CD I and CD II) that form a pocket together with both UDP–GlcNAc binding sites and a catalytic center (34, 35). In addition to its catalytic domains, OGT contains 11.5 tetratricopeptide (TPR) motifs in the protein’s N-terminus. TPR motifs are 34-amino-acid structures with several electrostatic surfaces that function primarily in protein–protein interactions, and this allows OGT to interact with a large number of proteins and with itself (36). OGT has at least three known isoforms with different cellular localizations (37). These isoforms all contain the catalytic domains; however, they differ in their number of TPR motifs, which suggests that the variations in cellular localization are in part caused by the targeting of its binding partners (37).

Because OGT has several binding partners, substrates, and varying expression levels and localization, it is not surprising that it is regulated tightly. To date, OGT has been shown to be O-GlcNAc modified as well as tyrosine phosphorylated (35, 38). The roles and sites of these modifications have not yet been elucidated, but studies suggest the tyrosine phosphorylation may have an activating effect on the enzyme’s catalysis (39). OGT substrate specificity and localization are regulated by its binding partners as well as by its own multimerization (35). Finally, OGT activity can also be influenced by feedback mechanisms. OGT is highly responsive to intracellular UDP–GlcNAc levels (35), and free UDP is a potent inhibitor of this substrate recognition (35). OGT catalytic activity has also been demonstrated to be inhibitable by the drug compound alloxan, which is a uracil analog (40). Naturally, neither of these inhibitors is specific for OGT activity within a cell. Currently, the development of OGT-specific inhibitors has only been preliminarily characterized (41).
O-GlcNAcase

β-O-linked N-acetylglucosaminidase, or O-GlcNAcase, is a 917-amino-acid enzyme responsible to remove O-GlcNAc modifications on a protein selectively. Like OGT, O-GlcNAcase can be found in the nucleus and in the cytosol of all tissues and organisms higher than yeast. Also, like OGT, O-GlcNAcase is expressed by a single gene product at a chromosomal locus at 10 q24-1 3 (in humans) that encodes at least three alternative splice variants in addition to the full-length product. The O-GlcNAcase active site, a TIM barrel (42, 43), resides in the protein’s N-terminus (44, 45) and is part of a larger enzyme that contains an acetyltransferase (AT) domain in the protein’s C-terminus. O-GlcNAcase has been shown to have acetyltransferase activity, at least for histone substrates, and therefore, it is sometimes referred to as NCOAT (nucleus-cytoplasmic O-GlcNAc and acetyltransferase) to reflect both activities (46). Two naturally occurring splice variants of this enzyme lack a portion of the O-GlcNAcase active site (exons 8 and 9, encoding amino acids 250-345 and 250-398, respectively), and therefore, they are defunct for O-GlcNAcase activity (46). The third variant lacks the AT domain (47) completely. O-GlcNAcase/NCOAT has been shown to be a substrate of the pro-apoptotic protease caspase-3. The cleavage site flanks the C-terminus of the O-GlcNAcase domain, and because the cleaved products retain their respective activities (46, 48), this may regulate/deregulate these activities as well as disconnect their locality within a cell (48).

Although it is speculated that 0-GlcNAcase activity must be regulated highly, little progress has been made to reveal posttranslational modifications that may occur on the protein. O-GlcNAcase has been shown to be O-GlcNAc modified, although it has not been elucidated as to whether this would have an activating or inhibiting effect on the enzyme (37). Like with OGT, O-GlcNAc modification on O-GlcNAcase may represent a unique feedback mechanism to help regulate cellular protein O-GlcNAc levels by regulating the enzymes responsible for the modification themselves. In addition to the variants and caspase cleavage products mentioned above, it is interesting that O-GlcNAcase exhibits strong direct interaction with OGT. It does so via a domain in the middle of the enzyme and through the N-terminus and first six TPRs of OGT (49). To prevent a futile cycle of glycosylation/deglycosylation that surrounds these complexes, the regulation of these enzymes becomes of paramount importance.

Several O-GlcNAcase inhibitors have been described, the most popular are streptozocin (STZ) and PUGNAc. Both of these compounds are substrate mimetics that inhibit O-GlcNAcase activity by resembling the natural substrate’s oxazoline transition state after its entrance into the active site (50, 51). Unfortunately, the above inhibitors are nonselective and can inhibit other glycosyl hydrolases; therefore, they are a detriment to a multitude of cell pathways. A more potent transition state analog, N-acetylglucosamine-thiazoline (NAG-thiazoline), has been described recently (5). The increased potency likely is because the compound already resembles the oxazoline intermediate before it is exposed to the enzyme. Macauley et al. (52) have begun recently to generate NAG-thiazoline derivatives with augmentations to the thiazoline ring. In particular, the addition of a butyl chain to the ring caused a dramatic increase in potency over the NAG-thiazoline parent compound and exhibited a selectivity toward O-GlcnAc case over lysosomal hexosaminidases. In addition, Stubbs et al. (53) and Kim et al. (54) have begun to characterize PUGNAc derivatives. These analogs involve alkyl extensions, and whereas the extensions are the same and in a position analogous to those of the NAG-thiazoline derivatives, the extensions are somewhat less potent and specific for O-GlcNAcase, although their usefulness may yet be evident in different contexts. The increased potency and selectivity of these compounds over their precursors represent important advances and will make them invaluable tools in the study of O-GlcNAcylation. Furthermore, Kim et al. (55) have exploited this butyl-extension to increase the sensitivity of a fluorogenic O-GlcNAc substrate, giving the field a promising high-throughput imaging tool for the analysis of O-GlcNAc functions.

Emerging approaches for the Study of O-GlcNAcylation

Until the last decade, proficient methods to study O-GlcNAcylation on a target substrate had been slow to develop. In whole-cell systems, the use of OGT or O-GlcNAcase inhibitors—or the overexpression of these proteins—is toxic or affects too many cellular pathways to interpret results directly. In isolatable systems, recombinant OGT or O-GlcNAcase, together with their respective inhibitors, can be used to identify target substrates, either by radiolabeling using UDP-[3H]GlcNAc or by Western blotting using protein-O-GlcNAc specific antibodies such as RL2 or CTD110.6. In several cases, the subsequent effect of the O-GlcNAc modification on the protein’s function could be determined in isolatable in vitro assays. More commonly, the effect of the O-GlcNAc modification on the protein was never fully understood, because the sites of O-GlcNAcylation could not be detected. This limitation prevents mutational analysis, and the direct investigation of the role of the O-GlcNAc in purified assays, in signaling pathways and in whole-cell and animal contexts.

Traditional methods to map posttranslational modification sites, like those of phosphorylation, have been anchored by protein digest and mass spectrometric (MS) approaches (for a review on the classic evaluation and for MS analyses of O-glycans, see Reference 56). Unfortunately, like many posttranslational modifications, O-GlcNAcylation occurs routinely on a protein population with substoichiometric frequency, which results in a very small detectable fraction of the O-GlcNAc-modified product. Also, much like O-phosphorylation, the protein-O-GlcNAc bond is labile and is detached by cell-permeable dissociation (CDD) during MS analysis. Often, the bond is lost before it can be detected on the peptides analyzed (57, 58). Phosphate modifications, however, can overcome this limitation by enriching the peptide mixtures via collision-induced dissociation (CID) during MS analysis.
through chromatographic approaches, for example, by using immobilized metal ion affinity columns. Such methods to enrich O-GlcNAc-modified peptides specifically are beginning to be optimized.

Recent advances have been developed to enrich O-GlcNAcylated peptides by using a lectin weak affinity chromatography (LWAC) approach. In the LWAC approach, a column of wheat germ agglutinin (WGA), which has binding affinity for GlcNAc as well as sialic acid, is used to separate and enrich GlcNAcylated peptides from the non-GlcNAcylated population (57). Although the binding affinity is weak, with dissociation constants in the 10-millimolar range, the interaction is sufficient to retard the rate of flow of GlcNAcylated peptides by HPLC enough to separate them for MS analysis and for sequencing. It has also been shown that such WGA chromatography is sufficient enough to purify proteins for 2-D gel electrophoresis and for subsequent MS proteomic analyses (59). These protocols can be used both on purified protein digests as well as in cellular extracts, and they represent a key advance in identifying O-GlcNAc modification sites.

A second promising method for mapping of O-GlcNAcylation sites involves a β-elimination/Michael addition with the DTT (BEMAD) approach (57, 58). This technique can be used alone, or in conjunction with the above LWAC method. Briefly, the use of a strong base such as sodium hydroxide can result in the β-elimination of the C–CH2–O–GlcNAc to C=CH2, where the first C is within the protein backbone on the digested peptides. This technique is followed by a Michael addition of DTT to the CH2 group (other “tags” can be added through this technique as well). The result is a formerly β-elimination susceptible site modified so that it can be enriched by chromatography (and is less labile), as mentioned above for O-phosphate peptide enrichment. In the case of BEMAD, an activated thiol-Sepharose can purify the DTT-tagged peptides, although a variety of columns can be used depending on the tag chosen. These affinity-tag methods have advanced enough so that they can be used to label O-GlcNAc sites selectively, whereas other β-elimination susceptible sites, like those of O-phosphate, are left undisturbed.

Affinity tags have also been developed to circumvent the β-elimination. The first uses an engineered galactosyltransferase to apply a UDP-ketone analog in the place of O-GlcNAc, which can be biotinylated for purification (60, 61). This strategy has been implemented successfully in both pure protein preparations and whole cell lysates. The second strategy exploits a phenomenon first reported by Vojdani et al., in which N-acidic glucosamine can be used as a substrate for both OGT as well as O-GlcNAcase (62). This method can be applied to cell culture, where a peracetylated GlcNAc azide analog is incubated with the cells (along with an O-GlcNAcase inhibitor to prevent deacetylation). After the incorporation of the analog at O-GlcNAc sites, the mixture can be treated with a biotinylated phosphate reagent to tag the sites of incorporation for streptavidin purification (63).


Cytosolic Glycosylated Proteins, Chemistry of


See Also

Chemical Tools for Studying Glycans
Chemistry of Glycans
Cytosolic Glycosylation in Signaling
Identification of Post-Translational Modifications
Regulating Protein Function by Post-Translational Modification
Glycosylation of Proteins in the Golgi Apparatus

Marguerite M. Desko and Jennifer J. Kohler,, Stanford University, Stanford, California,
doi: 10.1002/9780470048672.wecb212

Oligosaccharides are essential for interactions of cells with their environments. These complex carbohydrates are often found covalently attached to proteins embedded in eukaryotic cell membranes. Protein glycosylation is heterogeneous; this heterogeneity stems from the biosynthesis of these polymers. As proteins destined for secretion or cell-surface presentation travel through the endoplasmic reticulum and the Golgi apparatus, they are modified with sugars in a stepwise fashion by enzymes called glycosyltransferases. The differential expression of these enzymes leads to a multiplicity of specific oligosaccharides both among and within cells because not all cells contain all enzymes and because not all substrate proteins will encounter every enzyme. Although myriad oligosaccharides are found attached to proteins, most of these diverse structures can be grouped into several classes of glycans. In this article, we will discuss some of the most common forms of Golgi protein glycosylation: mucin-type O-linked glycosylation, N-linked glycosylation, and the formation of glycosaminoglycans. In addition, we will briefly consider some less common, but essential, forms of glycosylation.

The Organelles of the Secretory Pathway form an Assembly Line for Glycoprotein Biosynthesis

A large percentage of eukaryotic proteins have oligosaccharides covalently linked to certain amino acid side chains post-translationally. These posttranslational modifications are essential for cell–cell recognition, cell–matrix interactions, and cell–pathogen recognition. The protein-linked oligosaccharides are known as glycans. Biosynthesis of glycans occurs in the secretory pathway of eukaryotic cells.

Cellular glycans are biosynthesized in eukaryotic cells by sugar-transfer enzymes called glycosyltransferases. These enzymes reside in the secretory pathway, which comprises the endoplasmic reticulum (ER) and the Golgi apparatus. Glycosyltransferases transfer sugars from activated nucleotide sugar donors to a protein or nascent oligosaccharide substrate. Most Golgi-resident glycosyltransferases are type II membrane proteins, which are characterized by a short N-terminal cytoplasmic tail and a single-pass transmembrane region, followed by a variable-length stem region and catalytic domain, both of which extend into the Golgi lumen. Glycosyltransferase enzymes are localized to different parts of the Golgi, enzyme localization influences the order in which substrate proteins encounter different glycosyltransferases and thus the order and arrangement of sugar attachment. The cytoplasmic tail, transmembrane domain, and stem region have all been implicated in controlling localization of glycosyltransferases within the Golgi.

Nucleotide sugar donors, such as UDP-GalNAc, UDP-GlcNAc, GDP-fucose, and CMP-sialic acid, serve as substrates for the glycosyltransferases and are the source of the sugars that are added to substrate proteins (Table 1). Nucleotide sugar donors are synthesized in the cytoplasm and imported into the secretory pathway by membrane-resident transporters (1).

Glycan biosynthesis, unlike DNA, RNA, or protein biosynthesis, is not template directed. Rather, the secretory pathway functions as an assembly line that substrate proteins traffic through on their way to the cell surface. Substrate proteins enter the secretory pathway in the ER. As the proteins pass through the ER and the Golgi, they are modified by enzymes residing in these compartments. Glycosyltransferases that add sugars directly to the protein tend to be found earlier in the secretory pathway, particularly in the ER and the cis-Golgi. Other enzymes are responsible for further glycan modification. These elaborating enzymes tend to be localized in later Golgi compartments (medial and trans-Golgi) or in the trans-Golgi network (TGN) (Fig. 1). The exact modifications that occur depend on two factors: the availability of activated sugar donor.
Glycosylation of Proteins in the Golgi Apparatus

Table 1

<table>
<thead>
<tr>
<th>Common abbreviations and symbols</th>
<th>Glucose (Glc)</th>
<th>Mannose (Man)</th>
<th>Galactose (Gal)</th>
<th>Fucose (Fuc)</th>
<th>Sialic acid (Sia)</th>
<th>N-acetylglucosamine (GlcNAc)</th>
<th>N-acetylgalactosamine (GalNAc)</th>
<th>Glucuronic acid (GlcA)</th>
<th>Iduronic acid (IdOA)</th>
</tr>
</thead>
</table>

Here, we delineate the biosynthetic pathways of the most common types of protein glycosylation occurring in the secretory pathway: mucin-type O-linked glycosylation, N-linked glycosylation, and the formation of glycosaminoglycans. In addition, we will briefly visit the biosynthesis of some less-common varieties of protein glycosylation.

Secretory Pathway Enzymes Perform the Chemistry of Glycosylation

Mucin-Type O-Linked Glycosylation

Glycans can be attached via a glycosidic bond to the hydroxyl group of serine or threonine in a polypeptide. In these cases, the glycosidic bonds connect a sugar to an oxygen in the polypeptide; therefore, this type of glycosylation is termed O-linked. In mammals, the most common type of O-linked glycosylation is the mucin-type, in which N-acetylgalactosamine (GalNAc) is α-linked to either serine or threonine and is subsequently modified by additional sugars (Fig. 2). The name mucin-type derives from the initial isolation of glycoproteins containing these sugar moieties from mucus. Mucins are secreted by many types of tissues, including the linings of the digestive tract and airways.

Mucin-type O-linked glycans are synthesized in the Golgi apparatus of eukaryotic cells. Synthesis is stepwise, with individual Golgi-resident glycosyltransferases transferring sugars one at a time to the growing glycoprotein. A wide variety of final glycan structures is possible. The heterogeneity depends on the set of glycosyltransferases that are present in a particular cell and the order in which substrate polypeptide encounters its modifying enzymes. The sequence of steps required for biosynthesis of mucin-type glycans can loosely be divided into the following types of sugar transfer events: initiation, core formation, elongation/branching, and termination.

Initiation: The ppGalNAcTs

Polypeptide α-GalNAc transferases (ppGalNAcTs) initiate biosynthesis of mucin-type O-linked glycans by transferring GalNAc from UDP-GalNAc to an acceptor protein. Sequence analysis has revealed approximately 24 mammalian genes with high homology to known ppGalNAcTs. So far, 21 of these genes have been demonstrated to encode proteins that exhibit ppGalNAcT activity. Within the large ppGalNAcT family, individual isoforms display discrete patterns of tissue- and stage-specific expression (4), suggesting that they serve distinct developmental roles. However, there is some functional redundancy among family members as evidenced by the absence of discernable phenotypes in mice harboring targeted disruptions of individual ppGalNAcT genes.

Subcellular localization studies have been completed for some, but not all, of the ppGalNAcTs. Those enzymes that have been examined localize to positions within the Golgi; however, the exact distributions are isoform and cell-type specific. For example, α-GalNAcT3 (5) is localized to the cis-Golgi in...
Core Formation

Addition of a single GlcNAc to serine or threonine constitutes a structure called the Tn-antigen, often observed on tumor cells. However, in most cases, one or more additional monosaccharides are attached to the 3 or 6 positions of the initiating GalNAc, forming the core O-linked structures (Fig. 2). As with the GalNAc in low-molecular-weight polysaccharides, the 3 position of the initiating GalNAc can be modified by the 3- GlcNAc transferase (Fig. 2a), yielding the GlcNAc2,3GalNAc structure. The core 3 structure is not as ubiquitous as core 1 and has been found only in mucin-secreting tissues.

Despite the fact that only certain serine and threonine residues are glycosylated into the Golgi in other types of tissue. Other ppGalNAcT isoforms localize to the trans-Golgi compartments. Examples include human ppGalNAc-T2 and -T3, which localize to the trans-Golgi in HeLa cells (6). The localization of the initiating ppGalNAc-T determines which other enzymes the substrate protein can encounter as it completes its transit through the secretory pathway. Therefore, the localization of the initiating ppGalNAc-T affects the final structure: If the ppGalNAc-T is localized relatively late in the Golgi, fewer additional modifications are possible.

Porcine submaxillary gland, although it is distributed throughout the Golgi in other types of tissue. Other ppGalNAcT isoforms localize to the medial and trans-Golgi compartments. Examples include human ppGalNAc-T2 and -T3, which localize to the trans-Golgi in HeLa cells (6). The localization of the initiating ppGalNAc-T determines which other enzymes the substrate protein can encounter as it completes its transit through the secretory pathway. Therefore, the localization of the initiating ppGalNAc-T affects the final structure: If the ppGalNAc-T is localized relatively late in the Golgi, fewer additional modifications are possible.

A notable example is the porcine submaxillary gland, although it is distributed throughout the Golgi in other types of tissue. Other ppGalNAcT isoforms localize to the medial and trans-Golgi compartments. Examples include human ppGalNAc-T2 and -T3, which localize to the trans-Golgi in HeLa cells (6). The localization of the initiating ppGalNAc-T determines which other enzymes the substrate protein can encounter as it completes its transit through the secretory pathway. Therefore, the localization of the initiating ppGalNAc-T affects the final structure: If the ppGalNAc-T is localized relatively late in the Golgi, fewer additional modifications are possible.

Despite the fact that only certain serine and threonine residues are glycosylated into the Golgi in other types of tissue. Other ppGalNAcT isoforms localize to the medial and trans-Golgi compartments. Examples include human ppGalNAc-T2 and -T3, which localize to the trans-Golgi in HeLa cells (6). The localization of the initiating ppGalNAc-T determines which other enzymes the substrate protein can encounter as it completes its transit through the secretory pathway. Therefore, the localization of the initiating ppGalNAc-T affects the final structure: If the ppGalNAc-T is localized relatively late in the Golgi, fewer additional modifications are possible.
Glycosylation of Proteins in the Golgi Apparatus

The core 2 and 4 structures are synthesized by the addition of GlcNAc to the 6 position of GalNAc in the core 1 (Fig. 2c) and core 3 structures (Fig. 2e), respectively. These structures are produced by a family of β1-6GlcNAc-transferases that catalyzes the formation of the GlcNAc-α1-6GalNAc linkages. At least three β6GlcNAcT isoforms exist; they vary in their preference for core 1 or core 3 substrates (15). For example, the L-type (leukocyte) core 2 β6GlcNAc-T accepts only the core 1 substrate, producing core 2 antigen. The M-type (mucin) isoform, expressed in mucin-secreting cell types, exhibits broader substrate scope and is capable of synthesizing both core 2 and core 4 structures, as well as the I antigen (discussed below) (16).

Core 2 β6GlcNAcT localizes to the cis- and medial Golgi (17).

The addition of GlcNAc to core 1 by β1-3 GlcNAc-transferase 3 (β3GlcNAcT-3) results in the formation of extended core 1 structures (Fig. 2b). These glycans are frequently modified with additional sugars to form sialyl Lewis x or 6-sulfated sialyl Lewis x; this elaboration is essential to leukocyte rolling and L-selectin binding. Extended core 1 oligosaccharides may be further elaborated by the addition of GlcNAc to the 6 position of GalNAc, forming branched glycans containing core 2 (18). A dition of GlcNAc in a β1-6 linkage to Tn antigen generates the putative core 6 structure (Fig. 2g). Core 6 structures have been reported to occur in human ovarian tissue, but an enzyme responsible for core 6 formation has not been identified (19).

Less common core structures are formed when the initiating GalNAc is modified by the addition of α-linked sugars at either the 3 or 6 positions. Core 5 is formed by the addition of a second GalNAc in an α3-3 linkage to the first (Fig. 2f) (19–21). This enzymatic activity has been detected in biological samples, but the responsible gene has not been identified. The core 7 (GalNAc-β1-3GalNAc-c) (Fig. 2h) (22) and core 8 (Galα1-3GalNAc) (Fig. 2i) structures (23) are produced in restricted tissues and the responsible genes are not yet known.

Branching and Elongation

A common motif found in larger O-linked glycans is poly-N-acetylactosamine (Galβ1-4Galβ1-3), also called poly-LacNAc or the type 2 backbone (Fig. 3a). PolyLacNAc repeats are commonly found on fetal erythrocytes, where they constitute the blood group i antigen. PolyLacNAc chains of varying length are synthesized by the alternate action of two enzymes, a β1-4galactosyltransferase (β4-GalT) and a β1-3GlcNAc-transferase (β1-3GlcNAcT). A family of at least five β4-GalTs has been discovered and at least eight β1-3GlcNAcTs have been identified; these biosynthetic enzymes are widely expressed. The most well-characterized of the β1-3GlcNAcTs is known as iGnT because of its role in synthesis of the i antigen (24).

PolyLacNAc can be extended from sugars attached to either the 3 or 6 position of the initiating GalNAc; however, addition to the 6 (upper) branch is more common (Fig. 3c). A ditional branching can be introduced into the LacNAc backbone by the action of β1-6 GlcNAc-transferases, including the previously
and Sia produced by a family of six known members of this family can distinguish between Gal (FucT3, FucT4, FucT5, FucT6, FucT7, and FucT9) (26). Some β both. The resultant structures (Gal modified by glycosylation, and sulfation. Terminal sialic acids can also be further events. The most common modifications are fucosylation, sia-

Further Modification and Termination

Mucin-type O-linked glycans are decorated with a variety of capping structures. Given their prominent location, these structures play critical roles in cell-cell and cell-matrix recognition events. The most common modifications are fucosylation, sia-

N-Linked Glycosylation

The best known form of protein glycosylation is the asparagine- or N-linked variety. These large, branched structures contain a conserved core structure that is produced in the majority of eukaryotes, including yeasts, plants, and mammals. Other eu-

Assembly of the Dolichol Oligosaccharide Donor

The dolichol oligosaccharide donor (35) is composed of a tripyrophosphate that is attached to an oligosaccharide composed of 14 individual sugars: dolichol-P-P-GlcNAc5Man9Glc3. The dolichol donor is assembled by the action of the A1G fam-

Transfer of the Oligosaccharide to Polypeptides

The oligosaccharyltransferase (OT) transfers GlcNAc5Man9Glu3 from the dolichol donor to an asparagine (Asn) of the newly synthesized polypeptide, where a se-

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc. - 1461 -

mentioned core 2 β-GlcNAc α-M is a isoform. This type of branched structure, the I antigen, is commonly found on adult erythrocytes. Members of the β-GlcNAc α-transferease family that produce branched structures are known as IGnTs. Synthesis of branched structures is developmentally regulated, and members of IGnT enzyme family have differing substrate specificities. Some IGnTs select for terminal Gal whereas others modify internal Gal residues. The type 1 (Galα(1→3)Galβ(1→3)-β, polymer, also called lacto-N-biose, is a related but less common, backbone motif (Fig. 2b). Lacto-N-biose is synthesized by a β3-galactosyl-

Glycosylation of Proteins in the Golgi Apparatus
Glycosylation of Proteins in the Golgi Apparatus

Figure 4  Biosynthesis and structure of N-linked glycans. (A) The ER-resident Alg glycosyltransferases are responsible for addition of the individual monosaccharides to assemble the dolichol oligosaccharide donor. Initial steps are accomplished by Alg7 (a), Alg13/14 (b), Alg1 (c), Alg2 (d), and Alg11 (e) on the cytosolic face of the ER. The oligosaccharide is “flipped” from the cytoplasm to the ER lumen by Rft1 (f). Inside the ER lumen, additional sugars are added by glycosyltransferases Alg3 (g), Alg9 (h), Alg12 (i), Alg6 (j), and Alg10 (l). The assembled dolichol donor is transferred to the nascent polypeptide by the oligosaccharyltransferase (m). Trimming of the dolichol donor is catalyzed by glucosidase I (n), glucosidase II (o), and α-mannosidase I (p). In the Golgi, additional trimming is performed by Golgi-resident α-mannosidases (q) and GlcNAc is added by GlcNAcT-I (r). In the synthesis of typical complex glycan, two more mannoses are removed by α-mannosidase II (s), followed by the addition of GlcNAc by GlcNAcT-V (t) and GlcNAcT-V (u) and peptide residues by α1-4GalT (v). Extension of polylactosamine chains is performed by α1-3 GalNAcTs (w) and α1-4GalTs (x), whereas terminal sialic acid residues are added by fucosyltransferases (y) and sialyltransferases (z). (B) Subtypes of N-linked glycans include (a) mannan-type glycan; (b and c) monoantennary hybrid glycans; (d) a biantennary complex glycan; (e) a triantennary complex glycan; (f) a tetraantennary complex glycan; and (g) an N-linked glycan with a “bisecting GlcNAc.”

the ER through the translocon (40, 41). The heteromeric OX complex is composed of at least nine different polypeptides, including the STT3 subunit that provides the active site (42–44). The oligosaccharide is transferred to asparagines within the minimal consensus sequence Asn-X-Thr/Ser, where X is any amino acid except proline (45). The amino acids surrounding the consensus sequence also affect whether a particular Asn is a substrate for glycosylation. In addition, Asn-X-Cys sequences are occasionally glycosylated. After the oligosaccharide is transferred to a substrate protein, it is processed by ER- and Golgi-resident glycosidases.

Exit from the ER

The ER-resident glucosidase I removes the terminal α1-2 linked glucose from the oligosaccharide (Fig. 4A, n). Subsequently, glucosidase II removes the α1-3 glucose-linked glucose and then, more slowly, removes the α1-3 mannose-linked glucose (Fig. 4A, o) (46, 47). The presence of this mannose-linked glucose is intimately associated with protein folding. If a protein is folded improperly, the mannose-linked glucose is reinstalled by a glucosyltransferase (UGGT) and the protein remains in the ER to complete its folding (48, 49). Once the protein is properly folded and all three glucoses removed, α-mannosidase I removes the terminal α1-2 linked mannose from the middle chain (50) (Fig. 4A, p). The properly folded protein with its remaining Man0GlcNAc2–Asn glycan is now able to exit the ER and traffic to the cis-Golgi.

Trimming in the Golgi

Proteins destined for the cell surface or secretion are processed by Golgi-resident α-mannosidases, which add additional mannoses to produce Man2GlcNAc2–Asn (Fig. 4A, q).
Biosynthesis and structure of glycosaminoglycans. Synthesis of the core linkage tetrasaccharide is accomplished by XylT, GalT-I, GalT-II, and GlcAT-I. The repeating disaccharides of heparin/HS, CS, and DS are shown.

Chemical tools can be used to interfere with cellular glycosylation. Tools to interfere with mucin-type O-linked glycosylation include (a) α-benzyl GalNAc; (b) Galβ1-4GlcNAcβ-O-naphthalenemethanol; (c) Galβ1-3GlcNAcβ-O-naphthalenemethanol; (d) 1-6A; and (e) 2-6A. Molecules used to disrupt N-linked glycosylation at different steps include (f) 1-deoxymannojirimycin; (g) deoxynojirimycin; (h) tunicamycin; and (i) swainsonine.

\[ \text{Man}_{5}\text{GlcNAc}_{2} \] is also referred to as high-mannose (51) (Fig. 4B, a), and it serves as the starting point for the synthesis of a variety of other N-glycan subtypes, described below.

An alternative processing pathway is used by proteins that will traffic to the lysosome. GlcNAc-phosphotransferase adds phosphate ester-linked GlcNAc residues to two mannoses in the oligosaccharide (52). A GlcNAcase then removes the GlcNAc sugars (53), revealing mannose-6-phosphate, which serves as a signal for the protein to be shuttled to the lysosome.

**N-Glycan Subtypes**

Despite the conservation of the core structure, the ways in which N-linked glycans are elaborated vary among organisms and cell
types. In yeasts, Man\(_{n}\)GlcNAc\(_{2}\) is elaborated by several mannosyltransferases to form large mannan-type structures. In metazoan organisms, there are 3 N-glycan subtypes: high-mannose, hybrid, and complex. Invertebrates produce high-mannose and hybrid-type N-glycans, but not complex structures (54). In vertebrates, Man\(_{n}\)GlcNAc\(_{2}\) can be modified to form hybrid and complex N-glycans. The central \(\alpha\)1,2-linked mannose bears two mannoses, which are \(\alpha\)1,3- and \(\alpha\)1,6-linked. Addition of GlcNAc to the \(\alpha\)1,3-linked mannose yields the hybrid structures (Fig. 4B, b–c), whereas complex structures have GlcNAc added to both the \(\alpha\)1,3- and \(\alpha\)1,6-linked mannoses (Fig. 4B, d–f). Hybrid and complex glycans can be described by the number of GlcNAc-containing branches, or antennae, they possess. Six different GlcNAc-transferases can initiate branch formation, generating various antennary structures (55).

Synthesis of both hybrid and complex N-glycans begins with the addition of GlcNAc \(\alpha\)1,2 to the \(\alpha\)1,3-linked mannose, forming a monosialo hybrid glycan (Fig. 4A, a). GlcNAc addition is catalyzed by the GlcNAcT-I enzyme encoded by the Mga1 gene (56). The resultant glycan is a substrate for the medial Golgi enzyme, a-mannosidase II, which removes two mannoses to generate GlcNAcMan\(_{n}\)GlcNAc\(_{2}\) (Fig. 4A, a, b). This same glycan can be produced by an alternative route: a-mannosidase-II catalyzed removal of two mannoses from the high-mannose glycan and subsequent addition of GlcNAc by GlcNAcT-I. Synthesis of biantennary hybrid glycan is also possible, through the action of GlcNAcT-V, which adds a GlcNAc in a \(\alpha\)1,4 linkage to the \(\alpha\)1,3-linked mannose (57).

Complex glycans (Fig. 4B, d–f) are produced when GlcNAcT-II adds GlcNAc in a \(\alpha\)1,2 linkage to \(\alpha\)1,6-linked mannose (Fig. 4A, c). Once this linkage occurs, an additional antennae can be produced by GlcNAcT-V, which adds GlcNAc in a \(\alpha\)1,4 linkage to the \(\alpha\)1,6-linked mannose (Fig. 4B, e–f) (58). Finally, addition of GlcNAc in an \(\alpha\)1,4 linkage to the \(\alpha\)1,6-linked mannose is a rare modification catalyzed by GlcNAcT-VI.

Glycosaminoglycans and Proteoglycans

Glycosaminoglycans (GAGs) are defined by their composition; they are composed of long chains of repeating disaccharides. Alternating amino sugars (GlcNAc or GalNAc) and uronic acids (glucuronic acid or iduronic acid) comprise their disaccharide building blocks. The exact sugar composition and modifications to the sugars determine the classification of the GAG. Commonly occurring GAGs include heparin, heparan sulfate (HS), chondroitin sulfate (CS), and dermatan sulfate (DS). GAGs contain a common core tetrasaccharide linking them to a protein. Proteins or polypeptides with GAG chains attached are known as proteoglycans.

Biosynthesis of the Linkage Tetrasaccharide

Heparin, HS, CS, and DS share a common serine-linked core tetrasaccharide, GlcA\(\beta\)1,3-Gal\(\beta\)1,4-Gal\(\beta\)1,4-GalNAc Ser. Biosynthesis of this core tetrasaccharide, known as the linkage tetrasaccharide, is initiated in the ER by the transfer of xylose to serine by a xylosyltransferase (63, 64) (Fig. 5). This enzyme prefers to transfer sugars to serines followed by glycine and flanked by one or more acidic residues (65). The xylosylated protein is then transported to the Golgi to undergo further modification by three Golgi resident enzymes. GaT-I adds galactose in a \(\alpha\)1,4 linkage (66, 67). The resulting disaccharide is then elongated by GaT-II (\(\beta\)1,3-GalT6), a \(\beta\)1,3 galactosyltransferase. Finally, GlcAT-I, a GlcA transferase specific for this trisaccharide precursor, adds \(\beta\)1,3 glucuronic acid to complete synthesis of the linkage tetrasaccharide (68–69).

Sugars in the linkage tetrasaccharide can be modified. Common modifications are phosphorylation of xylose at c2 (70, 71) and sulfation of the second galactose at C4 (72). The function of these modifications is not known; not all proteoglycans containing the linkage tetrasaccharide are modified (73, 74).

The Branching Point: Addition of GlcNAc or GalNAc

Once the synthesis of the linkage tetrasaccharide is complete, it can be further modified by one of two Golgi-resident glycosyltransferases. The enzyme α-GlcNAcT-I can add \(\alpha\)1,4-linked GlcNAc, thereby initiating the heparin/HS biosynthetic pathway. In vitro activity assays indicate that α-GlcNAcT-I also appears to be capable of adding an \(\alpha\)1,3-linked GalNAc to cap the linkage tetrasaccharide, resulting in a pentasaccharide that cannot be further elongated (75). However, the biological relevance of this modification is not known. If the protein is not modified by α-GlcNAcT-I, ji-GalNAcT-II can add GalNAc \(\beta\)1,4 to the glucuronic acid, initiating the CS/DS pathway (76). The CS/DS pathway is suggested to be the default pathway for linkage tetrasaccharides not intercepted by α-GlcNAcT-I.
Heparin and HS

After the initiation of heparin/HS biosynthesis by the addition of GlcNAc to the linkage tetrasaccharide, polymerization of the disaccharide β1-4GlcA–1-4GlcNAc begins (Fig. 5). The polymerization reaction is catalyzed by two bifunctional Golgi-resident glycosyltransferases, EXT1 and EXT2. Individual, these enzymes each exhibit GlcNAc- and glucuronic acid-transferase activity, but they are most active when they are physically associated with one another and localized to the Golgi (77).

As the β1-4GlcA–1-4GlcNAc, polymer is produced, it is modified by a number of Golgi-resident enzymes that deacetylate, epimerize, and sulfate the growing GAG. The GlcNAc N-deacetylase/N-sulfotransferases (NDSTs) catalyze both the deacetylation and N-sulfation of GlcNAc in the repeating disaccharide. The four different NDST isozymes display different substrate specificities and vary in their enzymatic activities (78–83). Glucuronic acid (GlcA) residues adjacent to GlcNAc can be epimerized to iduronic acid (IdoA) by GlcA-C5 epimerase. The newly formed IdoA can then be sulfated at the 2 position by the 2-O-sulfotransferase, which prevents reverse epimerization to GlcA. The epimerase and 2-O-sulfotransferase work in close concert and are known to physically associate with one another (84). Further O-sulfation of the GAG chain is accomplished by 6-O-sulfotransferases (6-OSTs) and 3-O-sulfotransferases (3-OSTs). The three 6-OSTs that add sulfate to N-sulfated GlcA (deacetylated GlcNAc) (85) and the six 3-OSTs that sulfate the 3 position of the hexosamine (86) exhibit varying substrate specificities (87, 88).

Heparin and HS are formed from the same disaccharide repeat and differ only in their degrees of modification. Heparin is highly modified, containing more Idas than GlcA and having many sites of N- and O-sulfation; the predominant carbohydrate motif is a repeating trisulfated IdoA (2-O-SO3)–GlcNAc(6-O-SO3)(6-O-SO3) unit. HS is considerably more heterogeneous; modifications are confined to distinct regions of the oligosaccharide chain. HS can be rather large, reaching up to 70 kDa, whereas heparin is generally only 10–12 kDa. Like other glycans, the structures of heparin and HS GAGs are ultimately determined by expression levels of Golgi-resident glycosyltransferases and the availability of activated substrate.

CS and DS

Linkage tetrasaccharides not modified by α1-3GlCNAcT-I in the Golgi will be converted to CS or DS in the TGN. The oligosaccharide chain is elongated by the bifunctional CS synthase, which produces the β1-3GlcA–β1-4GlcNAc–3Cl polymer (89–90) (Fig. 5).

As the oligosaccharide is elongated, it is modified by several sulfotransferases. At any point during these modifications, a CS-epimerase is able to convert GlcA to IdoA (91). The presence of IdoA indicates that the oligosaccharide is DS rather than CS (Fig. 5). Three GalNAc-C5-OSTs are involved in the transfer of sulfate to GlcNAc in GlcA-rich regions of the oligosaccharide (92). The CS-epimerase works in conjunction with a DS-specific GalNAc-C6-O-ST (DSTST1) (93). A 2-O-sulfotransferase known as C5DS2ST catalyzes the addition of sulfate to both IdoA and GlcA (94). In addition, two different 6-O-sulfotransferases can produce chondroitin 6-sulfate. These two sulfotransferases have different substrate specificities, with C6ST-II sulfating GlcNAc(4-O-SO3) and with C6ST producing GlcA(2-O-SO3)–GlcNAc(6-O-SO3). In addition, a DS-specific GalNAc-C6-O-ST (DSTST) sulfates C6 on GalNAc flanked by two IdoAs (95).

Keratan Sulfate

Keratan sulfate (KS) is often grouped with the GAGs even though it does not meet the definition of a GAG. KS consists of polyN-acetylglucosamine sulfated at C6 on both hexoses. Three distinct types of KS exist that are differentiated by the way the sulfate is attached; these types are designated KS1, KSII, and KSIII (96).

KS biosynthesis is initiated in the same way as N-linked glycan biosynthesis, with the addition of an oligosaccharide to Asn from the dolichyl donor in the ER. The oligosaccharide is trimmed to form the KS linkage oligosaccharide, which is is of the complex biantennary type (97). The linkage oligosaccharide can be modified on either the C6 branch or the C3 branch to form KS structures (98–100). The nonreducing terminus is usually very highly sulfated, whereas the sugars nearest the reducing end are not sulfated (101). KS1 chains are frequently terminated with sialic acid and less frequently with GalNAc or Gal (97, 102).

KS1 biosynthesis is initiated by the formation of mucin-type core 2 disaccharide (103). KS1 is usually capped with sialic acid at either C3 or C6 of terminal GlcNAc. Sulfated GlcNAc in KS1 is often modified with α1-3-linked fucose, although not within four sugars of the terminus (104). KSII is the least-characterized member of the KS family. This polyN-acetylglucosamine polymer is extended from a serine-linked mannosyl and has been found in brain tissue (105).

Elongation of KS is not well understood. The activities of β4GalT-I, β3GnT, and IgNt have been suggested to fulfill this role, but none of these enzymes have been directly linked to KS biosynthesis (106–109). Sulfation of KS is catalyzed by at least two sulfotransferases. KSαGal6ST can sulfate terminal and internal galactoses of the elongating oligosaccharide (110); sulfation of terminal galactose is believed to block further KS polymerization (111). In contrast, GlcNAc6-O-sulfation occurs only on nonreducing GlcNAc and must be simultaneous with chain elongation. GlcNAc6ST-5 is the most likely candidate to provide the GlcNAc6-O-sulfation activity (112).

Other Types of Protein Glycosylation in the Golgi

O-fucose is an important modification that mediates cell-cell interactions and leads to intracellular signaling events. Fucose is linked to either serine or threonine found in the consensus sequence CXXGXXGXXFC3. The best-characterized modification sites are in the EGF domain of Notch (113). O-Fuc is added by protein O-fucosyltransferase (O-FucT-I) and may be either a standalone modification or further extended. If extended,
Glycosylation of Proteins in the Golgi Apparatus

Fringe glycosyltransferase, a β-3GlcNAcT, adds GlcNAc to fucose, followed by addition of galactose by β-1,4GalT-I. The single LeaNAc is terminated with α-2 6 sialic acid, although this modification is not essential to notch signaling. Glycans can also be associated with proteins via O-linked mannoside. Mannose is added in the ER and can be further elaborated in the Golgi to form a tetrasaccharide, 1,2-3Gal(1-4)GlcNAc(1-2)Man-Ser/Thr (114). O-mannose glycans are essential to proper brain and muscle function and their absence is associated with muscular dystrophy.

Polyasialic acid (PSA) is a linear homopolymer of α-2-8- or α-2 3-linked sialic acid attached to protein scaffolds including the neural cell adhesion molecule (NCAM). Two Golgi-resident polysialyltransferases, PST and STX, are responsible for its synthesis. Although PSA is widely distributed in embryonic tissues, it is normally found only in regenerating neural and muscle tissues in adults. Neo-expression of PSA is often associated with metastatic cancer (115).

Collagen is an important glycoprotein whose glycosylation does not fit into any of the categories previously described. In collagen, many prolines and lysines are posttranslationally hydroxylated and the hydroxyl-lysine residues further glycosylated. Hydroxyl-lysine may be modified by the presence of either a single galactose or Glc1-2Gal (116).

Chemical Tools to Manipulate Cellular Glycosylation

Protein glycosylation affects a large number of biological interactions, including developmental processes, cancer metastasis, and host-pathogen interactions. Small molecule tools allow researchers to control glycosylation in vivo and determine the roles of oligosaccharides in biological processes. Chemical tools provide advantages over genetic approaches because they offer time-dependent, dose-dependent, and reversible control of glycosylation events.

Small molecules such as brefendin A and nocardazole are molecules that disrupt the architecture of the Golgi apparatus, which results in a perturbation of the spatial organization of Golgi-resident proteins, including glycosyltransferases. Substrate proteins no longer encounter glycosyltransferases in the correct order, leading to gross changes in cellular glycans. Brefendin A reversibly inhibits vesicle trafficking from the ER to the Golgi; nocardazole depolymerizes microtubules and arrests the cell cycle, leading to improper formation of the Golgi.

Competitive primers function as alternative substrates for elaborating glycosyltransferases. These small molecules have provided a facile way to interfere with mucin-type synthesis and lead to truncation of the O-linked glycans found on cellular substrates. The competitive primer α-benzyl-α-galNAc (Fig. 6a) mimics the Tn antigen, whereas Galβ1-3GlcNAcβ1-2D-naphthalenemethanol (Fig. 6c) and Galβ1-4GlcNAcβ1-4N-p-naphthalenemethanol (Fig. 6d) are competitive with type 1 and type 2 backbone polymers. Inhibition of the initiating pgGalNACT is an alternative strategy for obstructing mucin-type biosynthesis. A uridine-based library has yielded several competitive inhibitors of the pgGalNACTs, including 1-6BA (Fig. 6d) and 2-6BA (Fig. 6e) (117). These molecules have been used successfully in cellular and organ culture (118).

Chemical control of the biosynthesis of N-linked glycans has been facilitated by the availability of cell-permeable natural products that interfere with various steps in oligosaccharide processing. Tunicamycin (Fig. 6b) impedes the first step of dolichol oligosaccharide donor synthesis, preventing transfer of GlcNAc from UDP-GlcNAc to Dol-P. Consequently, N-linked glycans are transferred to polypeptides. Once the oligosaccharide is added to the protein, desoxynojirimycin (Fig. 6g), nojirimycin, and related derivatives can be used to prevent further processing. These molecules inhibit trimming by ER-resident glucosidase (119, 120). Another azasugar, 1-desoxymannojirimycin (Fig. 6f), is a mannose analog that is able to inhibit the ER a-mannosidase I, keeping the substrate protein from being transported from the ER to the Golgi (121). Swainsonine (Fig. 6i) functions in the Golgi as a reversible inhibitor of a-mannosidase II. Treatment with swainsonine blocks elaboration of the oligosaccharide, precluding modifications by the GlcNAc-transferases (122). As each of these molecules functions at an early stage in N-linked glycan biosynthesis, they have a global effect on N-linked glycosylation.

Competitive primers can also be used as alternate substrates for GAG biosynthesis. β-D-xylolides containing two fused aromatic rings interrupt the galactosyltransferases involved in linkage tetrasaccharide biosynthesis (123).

Summary

The plethora of oligosaccharide structures synthesized by living organisms provides a tantalizing diversity of structures for chemists to explore. Although methods to analyze glycoconjugates are rapidly improving, we remain unable to predict or program cellular glycosylation events because of the challenges posed by the nontemplated nature of glycan biosynthesis. In the future, systems biology approaches may give a predictive understanding of the glycans produced by a given cell. Further exploration of these pathways will be enhanced by chemical biologists' development of additional chemical tools with targeted, rather than global, effects on cellular glycosylation.

Acknowledgments

We thank Dr. Danielle H. Dubé, Michelle R. Bond, Peter L. Lee, and Ethan J. Greenblatt for discussions and comments on the manuscript.

References

10. Elhammer AP, Kezdy FJ, Kurosaka A. The acceptor specificity of
11. Ten Hagen KG, Bedi GS, Tetaert D, Kingsley PD, Hagen FK,
14. Granovsky M, Bielfeldt T, Peters S, Paulsen H, Meldal M, Brock-
15. Dalziel M, Whitehouse C, McFarlane I, Brockhausen I,
4. Kingsley PD, Hagen KG, Maltby KM, Zara J, Tabak LA. Di-
8. Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, con-
N-glycosylation throughout the Golgi
2002;276(39):27867–27874.
Glycosylation of Proteins in the Golgi Apparatus


Lubas WA, Spro R. Gal-α1,3-Man disaccharide from rat liver, a novel N-linked carbohydrate unit processing enzyme. J. Biol. Chem. 1987;262(13):3771-3778.


Glycosylation of Proteins in the Golgi Apparatus


Further Reading

Glycosylation of Proteins in the Golgi Apparatus


See Also

Glycosyltransferases, Chemistry of Golgi Trafficking, Glycoengineering
Glycosyltransferases (GTs) form glycosidic bonds by catalyzing the transfer of saccharides from a donor to a wide variety of acceptors. The donors used by GTs are sugars conjugated to nucleotides, phosphates, or lipid phosphates; whereas acceptors consist of carbohydrates, proteins, lipids, DNA, and numerous small molecules such as antibioticonols, steroids, and so on. Together, the products of these reactions comprise the most diverse and abundant class of natural compounds found in nature. Numerous GTs are needed to synthesize these compounds because the formation of each distinct glycosidic linkage requires a different enzyme. The abundance of these enzymes is emphasized by the fact that GTs constitute 1% of the genes in all genomes sequenced. They are ubiquitous in every kingdom of life and in all compartments of the cell. Currently, over 32,000 GT ORFs have been classified into 90 families on the basis of amino acid sequence similarity. The structures for 64 GTs have been determined to date and generally reveal conserved architectures of a GT-A or GT-B fold, although other folds have been observed and are predicted. These crystal structures, together with biochemical data, have provided insight into the catalytic mechanism. GTs generally exhibit strict regio/stereospecificity and transfer with either retention or inversion of configuration at the anomeric carbon of the donor sugar. The importance of characterizing the precise activity of these enzymes is exemplified by the many genetic disorders that have been linked to aberrant glycosylation.

Glycosyltransferases (GTs) comprise a group of enzymes that catalyze the synthesis of glycosidic linkages by the transfer of a sugar residue, which generally is a monosaccharide, from a donor to an acceptor substrate (1–3). The products of these reactions are oligosaccharides, polysaccharides, and also numerous glycocompounds, which consist of carbohydrates linked to other noncarbohydrate molecules (Fig. 1). Together, these reaction products comprise the most diverse and abundant class of natural compounds found in nature. A different GT is required for each distinct sugar that is transferred and for each unique linkage that is formed. Because thousands of different linkages exist, this class of enzymes is enormous. In fact, GTs encode for 1% of the genes of all genomes sequenced to date. However, compared with well-characterized families of enzymes like the proteases or the protein kinases, considerably less is known about the GT enzyme family. Now the monumental task begins of identifying the precise function for each GT enzyme. This article will give an overview of GT mechanisms and structure/function relationships that have aided our understanding of the numerous biological roles of GTs. Furthermore, classification schemes as well as biotechnological and pharmaceutical aspects of GTs will also be discussed.

Glycosyltransferase Reactions

GTs catalyze the transfer of a monosaccharide or more rarely an oligosaccharide residue from a donor to an acceptor. The most common donors are sugar-nucleotides and GTs that use these donors are termed the Leloir-type GTs. These GTs comprise 90% of GT annotations, with UDP/TDP nucleotides accounting for more than 60% of these. However, dolichol-phospho-sugars, sugar-1-phosphates, and sugar-lipid-phosphates also serve as donors (Fig. 2). Acceptors cover every chemical class of compounds that include saccharides, polysaccharides, lipids, protein, DNA, and other natural products. GTs use these donor and acceptor substrates to form glycosidic linkages either through processive transfers with multiple additions of the same monosaccharide to the nonreducing end of a growing chain, as
Glycosyltransferases, Chemistry of

![Chemical structures]

Figure 1  The diversity of natural structures produced by GTs. These structures include carbohydrate energy sources such as sucrose, lactose, and amylose, as well as oligosaccharide structural elements such as cellulose, chitin, and peptidoglycan. Secondary metabolites like the flavonoid delphinidin 3,5,3'-tri-O-glucoside are modified by GTs, whereas the steroid estrogen is inactivated by conversion to estrone 3-glucuronide. Glycoconjugates important in therapeutics and disease such as antibiotics (streptomycin and oleandomycin), sialoside receptors for viruses, and blood group antigens are also produced by GTs. The synthesis of glycolipids such as GM1 ganglioside and galactolipid, is also dependent on GTs. Cellulose, chitin, and amylose are synthesized by processive enzymes that repetitively transfer monosaccharides, whereas the remaining structures are synthesized by nonprocessive enzymes each carrying out single addition of a different sugar.
Figure 2 Examples of glycosyl donors. Nucleotide donors are represented by UDP-galactose (UDP-Gal), GDP-fucose (GDP-Fuc), CMP-N-acetylneuraminic acid (CMP-sialic acid, CMP-Neu5Ac), GDP-galactose (GDP-Gal), dTDP-rhamnose (dTDP-Rha), and dTDP-daunosamine; lipid phosphate donors are represented by Lipid II and dolichol-phosphate-glucose (Dol-P-Glc); sugar phosphate donors are represented by glucose-1-phosphate. The saccharides are transferred from these donors by GTs to form oligosaccharides and glycoconjugate products.
Glycosyltransferases, Chemistry of

observed in the production of glycan biopolymers like cellulose and amylose, or through nonprocessive single transfer reactions, as exemplified by the blood group antigens, glycolipid GM1 ganglioside and streptomycin (Fig. 1). Because a different enzyme is required to synthesize each distinct glycosidic linkage, GTs are probably the enzyme class with the greatest chemical diversity of substrates.

Catalysis

With rare exceptions, the transfer of saccharides by GTs is regiospecific, such that the saccharide is transferred to only one of the many hydroxyl groups on acceptor molecules. The transfer reaction is also highly stereospecific. Thus, GT reactions follow two mechanistically distinct pathways that result in either inversion or retention of configuration of the anomeric configuration of the transferred sugar (Fig. 3). It is commonly assumed that the mechanisms of GT enzyme reactions are similar to those of the well-studied glycosidases that hydrolyze glycosides; one difference is that glycosidases transfer to a water molecule whereas GTs transfer to a hydroxyl group of an acceptor molecule. Although GT reactions in general favor the formation of glycosidic bonds biosynthetically, the reversibility of sucrose synthase and some natural-product enzymes has been useful for the production of nucleotide donors and natural product libraries (4).

Inverting mechanism

Inverting GT reactions are believed to follow a single displacement mechanism that involves nucleophilic attack of the OH-group of the acceptor on the anomeric center of the donor sugar (Fig. 3a). In this mechanism, a catalytic amino acid serves as a general base to deprotonate the reactive oxygen of the acceptor. Reaction occurs with formation of an oxocarbenium-ion transition state and is concomitant with departure of the nucleotide leaving-group. X-ray structures of several inverting enzymes (β1,4-galactosyltransferase, α1,3-galactosyltransferase, and the bacterial α2,3-sialyltransferases CstI and CstII) in complex with their donor and acceptor substrates, have revealed an amino acid residue (E317, E317, H202, and H188, respectively) that is in position to deprotonate the acceptor molecule or to stabilize the transition state (5–8). Other residues have also been found that determine the substrate specificity. In the β1,4-galactosyltransferase, a tyrosine residue (Y289) facilitates the use of UDP-Gal as a substrate but not UDP-GalNAc (5). A histidine (H308) plays a similar role in β1,3-glucuronyltransferase I by determining its specificity for UDP-Glca (9).

![Figure 3](image_url)

Figure 3  GT reactions are regiospecific and stereospecific that occur with either inversion (a) or retention (b and c) of configuration at the anomeric center of the donor sugar. Inverting enzymes are thought to follow a nucleophilic displacement mechanism in which a general base deprotonates the acceptor, which renders it nucleophilic so it can attack the donor sugar. In contrast, retaining enzymes are thought to proceed through a double displacement reaction via the formation of a covalent intermediate (b) or by the SN2-like mechanism (c) where the nucleophilic acceptor and departing donor are on the same face of the sugar ring.
Retaining mechanism

The catalytic mechanism for retaining GTs is unclear but is believed to occur via a double-displacement mechanism or a $S_{N}2$-like mechanism. The double-displacement mechanism (Fig. 3b) involves nucleophilic attack by an enzyme active site residue on the anionic center of the donor substrate. This mechanism leads to the formation of a covalent glycosyl-enzyme intermediate with inversion of configuration. In the second displacement, the glycosyl-enzyme intermediate is attacked by a hydroxyl group of the acceptor, after its deprotonation by a catalytic base. The configuration is inverted in the second step, which results in net overall retention of configuration in the product (Fig. 3a). For the $S_{N}2$-like mechanism (Fig. 3c), catalysis involves nucleophilic attack and departure of the leaving group through a concerted, asynchronous manner on the same side of the sugar ring. In early proposals, it was believed that a single catalytic residue on the retaining GT was required to act as the catalytic nucleophile by attacking the anomeric carbon of the donor. This mechanism was supported by the data obtained from structural elucidation of two retaining GTs, $\alpha$-lipo polysaccharide-$\alpha$-1,4-galactosyltransferase C (LgtC) from Neisseria meningitidis (10) and $\alpha$-1,4-N-acetyl glucosaminyltransferase (EXTL2) (11). Amino acids Gin189 and Arg 293 of LgtC and ETLX2, respectively, were the only functional groups positioned to contribute in the mechanistic process. Thus, it seems that (at least for these two enzymes) a single catalytic residue is likely, but more experimental evidence is required to corroborate these findings. Molecular modeling of various reaction processes with LgtC (12) as well as structural complexes of trehalose-6-phosphate synthase with nontransferable acceptors (13) suggested that the catalytic mechanism may actually be reminiscent of that proposed for glycosen phosphorylase (14). In this mechanism, it is thought that the active site constraints of the retaining GTs position the donor and the acceptor substrates in conformations that may not require a catalytic amino acid. When bound to the enzyme, the nucleotide sugar adopts a conformation where the sugar is folded over the pyrophosphate. The anomeric bond is elongated and weakened in this conformation thereby making the C1 position of the donor spatially accessible to direct attack by the OH nucleophile of the acceptor (13).

Glycosyltransferase Structure

Given the vast structural and functional diversity of GT products, together with the divergent evolution of the enzymes, it may be expected that numerous possibilities exist for GT folds. Indeed, GTs display a high level of diversity in their primary sequences, which indicates that multiple solutions may exist to the problem of how a protein can catalyze glycosyl-transfer. Furthermore, a large number of folds have been identified for the glycosyltransferases enzymes (15). Structural data indicate that GTs belong mainly in either the GT-A or the GT-B fold superfamilies or variants thereof (Fig. 4a-e 5, 8, 16, 17), although a lysozyme-like fold has been observed for peptidoglycan glycosyltransferase (18) (Fig. 4f), and a GT-C fold is predicted for integral membrane protein GTs that use lipid donors. Therefore, from an evolutionary standpoint, nature has settled on a limited number of protein folds to facilitate glycosylation events. It is noteworthy that the stereoselective outcome of glycosyl transfer reactions is not determined by fold because both retaining and inverting enzymes can belong to the same fold superfamily.

Glycosyltransferase-A fold

The first member of the GT-A superfamily fold was identified in 1999 when the three-dimensional structure of SpA was reported (19). This enzyme from Bacillus subtilis is involved in the formation of the spore coat and is a member of the Leloir type of GTs given its apparent use of nucleotide-diphosphate donor sugars as a substrate. Enzymes with the GT-A fold have an N-terminal $\alpha/\beta$ sandwich motif that resembles a Rossmann motif and is involved in nucleotide donor binding. Most GT-A fold enzymes also have a characteristic $\alpha$-$sp$-$\alpha$-$xx$-$\alpha$-$sp$ (DXD) or equivalent motif (EKD or TDD) near the center of the protein that coordinates to the phosphates in nucleotide donors via a divalent metal cation (M$^{2+}$) or M$^{4+}$. These motifs are present in the structure of the bovine $\beta$1,4-galactosyltransferase I (5) (Fig. 4a). Binding of the acceptor substrate does not have the same strong consensus character that has been noted for the donor site, but it has been noted that it takes place in the C-terminus. This acceptor-binding region generally contains motifs that consist of two flexible loops that undergo conformational changes after donor binds. In the retaining $\alpha$1,4-galactosyltransferase LgtC, the central flexible loop was shown in crystal structures to interact with both the donor and the acceptor sugar substrates (10, 20). An additional flexible loop involved in acceptor binding is located at the extreme C-terminus of these enzymes (20). Isothermal titration calorimetry studies with $\alpha$1,3-galactosyltransferase demonstrated that the disordered loop of the free enzyme has little affinity for the acceptor substrate (6). In this form, the enzyme possesses a wide-open active site cavity that serves to facilitate initial access of the donor sugar to the active site. Similar results have been noted for the crystal structure of $\alpha$1,4-GalT in complex with donor substrate (5) (Fig. 4a). It seems that residues in the flexible loop are fixed by their interactions with the donor sugar that leads to a semi-closed conformation of the active site that can bind the acceptor. It has been speculated that this semi-closed form is important for excluding solvent from the donor-binding site and thereby for preventing unwanted hydrolysis of the nucleotide-sugar substrates in the absence of acceptor. Once the acceptor binds, the loop adopts a closed conformation where donor and acceptor are positioned properly for catalysis (5, 6, 20).

Figure 4b shows a modular variant of the GT-A fold as reported for poly-epitope GaINAc transferase 10 (21). This enzyme has a C-terminal lectin domain that is thought to bind to GaINAc-containing peptides, which favors its substrate specificity for glycosylated peptides. Another variation of the GT-A fold is observed in the structures of $\alpha$2,3-sialyltransferases, CslI, and CslJ, from Campylobacter jejuni (7, 8). These enzymes use a monophospho-nucleotidyl sugar CMP-Neu5Ac as a donor. Crystal structures of these enzymes with substrate indicate that...
Glycosyltransferases, Chemistry of

this protein has an αβα sandwich motif reminiscent of the GT-A fold (Fig. 4c) and also contains flexible loops that undergo similar conformational changes on substrate binding.

Glycosyltransferase-B fold

The GT-B fold family includes most prokaryotic enzymes that produce secondary metabolites, like the antibiotics streptomycin, oleandomycin (Fig. 1) and vancomycin, and important bacterial cell wall precursors. It is also predicted to contain the vitally important O-GlcNAc transferase that modifies many nuclear and cytoplasmic proteins and influences gene transcription. The first glycosyltransferase structure reported in 1994 was for the GT-B fold enzyme, ji-glucosyltransferase (BGT) from bacteriophage T4 (22). This enzyme attaches glucose to modified cytosine bases on duplex DNA. Its low-sequence homology to other GTs made it difficult to draw comparisons between it and other GT-B enzymes. Only after the structures of two additional enzymes, MurG and GtfB, became available could consensus domains within the GT-B fold be verified (23, 24). MurG is an enzyme from Escherichia coli that catalyzes the transfer of GlcNAc from UDP-GlcNAc to the nonreducing end of a lipid-linked N-acetylmuramic acid acceptor to form the repeating unit of peptidoglycan. GtfB is a UDP-glucosyltransferase involved in the biosynthesis of chloroeremomycin (or vancomycin). More recently, the structure of another GT-B folded enzyme, oleandomycin glucosyltransferase from Streptomyces antibioticus that is involved in antibiotic synthesis was also reported (16) (Fig. 4d). All these enzymes, OleD, MurG, GtfB, and BGT, exhibit almost identical topology despite very little
sequence similarity. These enzymes and other members of the GT-B fold super family adopt a two-domain structure with a Rossmann-like fold in either domain. The predicted active site of the GT-B folded enzymes is located between the two Rossmann folds. Within the C-terminal portion of the cleft between the enzymes, another subdomain consists of an α/β structure that has a consensus glycine-rich pattern. This subdomain compromises the donor-binding domain. A crystal structure of MurG complexed with UDP-GlcNAc revealed that the first α-helix of this subdomain makes contact with the ribose moiety of the nucleotide, whereas the second α-helix interacts with the pyranose residue. Enzymes that use diphospho-containing sugar donors have to contend with the negative charge of the phosphates. The GT-B enzymes, unlike the GT-A enzymes, do not seem to use divalent metal cations for this purpose. Instead, the α/β subdomain of the Rossmann fold may account for this by creating a positively charged helix dipole with the first α-helix that acts to stabilize the α-phosphate of the nucleotidyl donor (2, 25), whereas other enzymes use basic residues to counteract the positive charge of the diphosphate moiety.

In addition to the conserved donor binding domains, the acceptor-binding site for GT-B folded enzymes has also been identified near the N-terminus. Analysis of GtfB-related sequences aided the identification of this site because all these enzymes are highly homologous at the primary sequence level; but functionally, they glycosylate different acceptors (2, 24). Because the acceptors are all from a similar group of peptide antibiotics, the variation in the sequence was presumed to reflect structural adaptations in the binding site to accommodate the slightly different acceptors.

A variant of the GT-B fold is shown in Figure 4d for an α1,6-fucosyltransferase (17). The structure shows an N-terminal coiled-coil domain, a catalytic domain that is similar to GT-B fold enzymes, and a C-terminal SH3 domain whose biological significance is currently uncertain.

Other glycosyltransferase folds

In addition to the variations of the GT-A and GT-B fold described above for sialyltransferases (7, 8) and fucosyltransferases (17), a distinct GT-C fold family has been predicted for GTs that use lipid linked donors. The crystal structures of the GT-domain of the peptidoglycan glycosyltransferase from Staphylococcus aureus (18) and Aquifex aeolicus (26) show structural similarity to the bacteriophage λ-lysozyme. These novel structures demonstrate the possibility of additional folds.
Glycosyltransferases, Chemistry of

Classification of Glycosyltransferases

The classic enzyme commission (EC) classification for GTs is on the basis of their donor and acceptor specificity as well as the product formed. Currently, 295 entries are in this database (http://www.chem.qmul.ac.uk/iubmb/). The distinction between these enzymes is noted by their ability to catalyze the transfer of hexoses (EC 2.4.9, hexosyltransferases), pentoses (EC 2.4.2, pentosyltransferases), or other glycosyl groups (EC 2.4.99, sialyltransferases). This classification is restricted to enzymes that are fully characterized, and it can be problematic for enzymes that act on several distinct acceptors but at different rates. It also does not take into account the origin of the enzyme or its three-dimensional structure.

A powerful, comprehensive classification of GT enzymes has been proposed by Henrissat and coworkers (http://www.cazy.org/). This classification scheme has been termed CAZy for carbohydrate active enzymes and delineates the enzymes into families based solely on amino acid similarity. This scheme has also been applied to the classification of other groups of enzymes, such as the glycosyl hydrolase, polysaccharide lyases, and carbohydrate esterases. These families are updated continually; currently, over 32,000 GT CRFs have been identified and classified into 90 families. Most enzymes in these families remain biochemically uncharacterized ORFs, but it has been demonstrated with both the GT and the glycosyl hydrolase families that this method of classification often leads to the grouping of enzymes with similar three-dimensional structures. However, as outlined in the structural fold section above, these three-dimensional structures do not necessarily translate into a predictable catalytic mechanism, substrate specificity, or the origin of the GT. For example, a few of the larger CAZy families, like GT2 with 8800 members, contain sequences that originate in bacteria, yeast, plant, viral, archaeal and animal species. This family also has at least 12 distinct GT activities, which include cellulose and chitin syntheses, mannosyltransferases, rhamnogalactosyltransferases and galactosyltransferases. Family 63 is mono-functional with a single entry for α-glucosyltransferase from bacteriophage T4. However, these families are not static. An example is the introduction of family GT78 when the three-dimensional structure of α-mannosylglycerate synthase (Mgs), from Rhodothermus marinus was reported (27). Despite the fact that this enzyme has high sequence similarity to members from GT2, a new family was generated because Mgs is a retaining enzyme whereas members of GT2 are typically inverting enzymes.

Glycosyltransferases in Biological Systems

The chemical diversity of glycosyltransferase reaction products is also reflected in their numerous biological roles (28). They serve as sources of energy exemplified by sucrose, lactose, and amylose (Fig. 1). They are also structural elements in cell walls and extracellular matrices. Oligosaccharides found on glycoproteins and glycolipids can be involved directly in the structure and the function of these molecules by affecting their stability, half-life, and activity within the cell. These oligosaccharides can also serve as receptors in biological recognition events such as signaling, development, and cell-cell adhesion or provide receptors for hormones, bacterial toxins and viruses. In bacteria and plants, oligosaccharides are also known to modulate the activity of secondary metabolites. Given these numerous biological roles of glycan products, it is important to understand the structure and the function of the GT enzymes that synthesize them. It is also pertinent to understand their overall function in the context of the biological system to which they belong. An overview of some important biological glycosylation reactions performed by GTs is described below. They are separated by taxonomy, but many occur throughout the kingdoms of life.

Eukaryotic glycosyltransferases

The best characterized eukaryotic glycosyltransferases are those from mammals. These GTs are important for the synthesis of N- and O-linked glycoproteins and glycolipids and are integral to the process of storing energy in the form of glycogen. Mammalian GTs use only nine nucleotide donor sugars, UDP-D-Galactose, UDP-D-Glucose, UDP-D-N-acetylglucosamine, UDP-D-Glucuronic acid, UDP-D-N-acetylgalactosamine, UDP-D-xyllose, GDP-D-mannose, GDP-D-L-Fucose, or CMP-D-N-acetylneuraminic acid. Mammalian GTs also use dolichol-phosphate-GlcNAc2Man9Glc3, dolichol-phosphate-mannose and dolichol-phosphate-glucose (Fig. 2).

In comparison with mammals, plants contain considerably more GTs because, in addition to the reactions carried out by mammalian GTs, they are required to convert the products of photosynthesis into diverse cell carbohydrates. For example, these GTs synthesize cell wall polysaccharides as well as secondary metabolites and xenobiotics. Plant GTs differ from mammalian GTs even more by their diversity of nucleotide donors. They use not only eight of the nine mammalian nucleotide donors, but numerous others such as UDP-L-rhamnose, GDP-L-glucose, GDP-L-galactose, UDP-L-arabinose, UDP-D-galacturonic acid, UDP-D-apiose, and so on. (29).

Despite the differences in roles, nucleotide donors, and number of GTs between eukaryotes, striking similarities exist between the processes by which these organisms synthesize glycan products. For example, the GTs responsible for the synthesis of N- and O-linked glycoproteins and glycolipids may differ in their substrate specificity, but they are all primarily located in the endoplasmic reticulum (ER) and/or Golgi apparatus. Furthermore, they follow similar steps for the assembly of their respective glycans. Although fewer in number, both plants and mammals do possess some important GTs that are not localized to the ER-Golgi synthesis pathways. Examples have been noted in both the cytoplasmic and the nucleoplasmic compartments as well as associated with the plasma membrane. To elaborate on some prominent GT reactions, the following descriptions have been separated based on their cellular location.

The GTs located in the ER-Golgi produce glycolipids and glycoproteins that are found on cell surface, in the extracellular matrix, or in fluids throughout the organism. These enzymes are usually type 2 transmembrane proteins composed of a short N-terminal cytoplasmic tail, a transmembrane domain,
Glycosyltransferases, Chemistry of

A through still associated with the ER-Golgi, the synthesis of O-linked glycoproteins differs greatly from the N-linked processes described above. Typically O-linked glycosylation of proteins occurs on serine and threonine residues, but it has also been identified on hydroxylated lysines (38). Semidineamine glycosylation is initiated in the Golgi by the action of retaining polypeptide GalNAc transferases (ppGalNAcTs, Fig. 42 (21, 39). The ppGalNAcTs represent one of the largest human GT families because they account for over 20 of the 250 known human GTs. Unlike N-linked protein glycosylation, the specificity of O-GalNAcTs does not seem to be determined by a known consensus sequence. This enzyme family is rather unique among GTs, in possessing a C-terminal, ricin-type lectin domain (39). This additional domain is believed to endow certain ppGalNAcTs with an even greater capacity to adapt to and capture glycosylated substrates, which ensures the high density of glycosylation characteristic of mucin domains (39). After ppGalNAcTs initiate glycosylation, other nucleotide GTs produce core 1 and core 2 glycan structures that can in turn be elaborated even more. These glycan structures impart unique physicochemical features to the proteins. For example, heavily glycosylated mucins facilitate the retention of water along environmentally exposed surfaces of the body.

Mammalian proteoglycans represent another class of heavily O-glycosylated proteins; however, they are synthesized by an alternative ER pathway. The protein attachment site for glycan synthesis initiation contains a Ser-Gly motif that is β-pyrosylated by the action of a xylosyltransferase (40). The β-pyrodase is extended by the action of additional galactosyltransferases to produce a core common to proteoglycans. This core is decorated by GTs even more through the sequential addition of alternating sugars of disaccharide repeating units to produce heparin, keratan, dermatan, and chondroitin (41). The biosynthesis of fully glycosylated proteoglycans involves different GT reactions that result in a heterogeneous product. Despite this heterogeneity, the glycans all contain the reproducible structural elements that allow them to fulfill their functions, such as in load-bearing portions of the joint that release water slowly under pressure then reuptake water when pressure is reduced.

Apart from GTs associated with the ER-Golgi, eukaryotes also possess biologically significant GTs that are associated with the plasma membrane. In particular, cellulose synthase is a complex of GTs that are responsible for the synthesis of the most abundant polymer on earth, cellulose. These GTs range in size from 988 to 1088 amino acids and have approximately eight transmembrane domains with large central cytoplasmic domain (42). Research on cellulose synthase has focused on understanding how these enzymes associate and how they coordinate the progressive addition of p-glucose to a growing polymer. This research is complicated by the fact that the formation of cellulose depends on several noncellulosic cell wall polysaccharides as well. Hemicellulose and pectin form the cell wall backbone on which cellulose is assembled, but these polysaccharides are synthesized within the Golgi by the activity of glycan synthases. Taking this into account, the synthesis of the final cell-wall polysaccharide product is astonishingly complex. Not only do they require the coordinated synthesis of a stem region of variable length, and a large catalytic domain that faces the luminal side (30). Proteolysis in the stem regions can generate soluble forms that are found in fluids such as milk and serum. It should also be noted that a crossover exists between the two groups of these GTs, which means that they can act on both lipids and proteins. Protein glycosylation events in the ER-Golgi are divided into two groups that consist of either N- or O-linked glycosylation.

Synthesis of N-linked glycoproteins begins by the precursory glycan first being assembled by GTs on dolichol-phosphatidate. In mammals, successive addition of monosaccharides from UDP-GlcNAc, GDP-Man, dolichol-phosphate-Man, and dolichol-phosphate-Glc results in a 14-sugar chain product [GlcNAc2Man9Glc3] (31). The next step involves the activity of a membrane-associated enzyme complex, which is termed oligosaccharide transferase (GST). This complex catalyzes the en bloc transfer of the assembled sugar chain from the dolichol-phosphatidate donor to an asparagine residue of a nascent protein. GST specifically modifies surface-exposed asparagines within the consensus sequence Asn-Xaa-Ser or Asn-Xaa-Tyr. After the en bloc transfer, the N-glycans are trimmed by ER glucosidases and ER-Golgi mannosidases and then converted to more complex structures by different GTs that use various donor sugars. By this process, it is possible for N-linked glycans to exhibit heterogeneity even on the same polypeptide. These glycans are called glycoforms, and well over 500 different structures have been chemically characterized (32). Among these glycans are the ABO blood group antigens. In the case of the O blood group, the structures terminate in 1,2-Gal. The A and B alleles are characterized (32). Among these glycans are the ABO blood group antigens. In the case of the O blood group, the structures terminate in 1,2-Gal. The A and B alleles are characterized (32).

The synthesis of glycolipids also occurs by successive GT reactions associated with the ER-Golgi. Two common classes of glycolipids synthesized by this method are the glycosphingolipids and the glycoalkyl and glycoalfosphoryls (GPI). Within the glycosphingolipid class, Glc-ceramide synthesis is initiated by the addition of β-Glc from a UDP donor by glucosylceramide synthase (35). For GPIs, the first sugar attached to the lipid is GlcN-ac that is desacylated to GlcN (36). These initial steps occur on the cytoplasmic face of the ER followed by flipping to the luminal face. The GPIs are then conjugated to proteins through N-linkages, whereas glycosphingolipids are modified by the addition of Gal, GalNAc, and NeuAc by Golgi GTs. This method of glycosylation and flipping across a membrane has been noted in plants during galactolipid synthesis on the chloroplast's thylakoid membrane. The GTs that form these galactolipids are particularly interesting because these galactolipids are the most abundant class of lipids in the biosphere (37).
activity of over 50 GTs, but also the action takes place in two different cellular locations.

Given the above processes, it is evident that the synthetic processes of glycans and glycoconjugates are often associated with a membrane in some fashion. Consequently, this process helps to organize or sequester the correct GTs for a particular function. Some important glycosyltransferase reactions have also been noted in other areas of the cell. For example, the energy storage enzyme glycogen synthase and a superfamily of CA2y family 1 glycosyltransferases, which is involved in the biotransformation of drugs and xenobiotics via glucuronosylation in mammals and glucosylation of small molecular weight acceptors in plants (44), are known to be located in the cytoplasm. The protein-modifying O-GlcNAc transferase is particularly interesting in that it uses UDP-GlcNAc to glucosylate serine or threonine residues on specific proteins. Glycosylation of these target proteins is balanced by the action of a soluble N-acetylglucosaminidase and seems to be reciprocal with protein phosphorylation. This dynamic balance in glycosylation levels is predicted to be involved in numerous processes, which include glucose metabolism, chaperone folding, nuclear pore protein translocation, and transcription factor regulation (43). The importance of O-GlcNAc transferase is clearly apparent not only because of its numerous activities in the cell, but also because of its discovery within the nucleus and cytoplasm of all metazoans.

Prokaryotic glycosyltransferases

Glycans in prokaryotic cells are integral to several cellular roles. Peptidoglycan is a determinant of cell shape and helps esubacteria withstand the pressures of the external environment. Lipopolysaccharide (LPS), lipooligosaccharides (LOS), capsules, and slime layers serve as attachment sites/receptors for cellular interactions, provide protection from environmental factors and are involved in immune system modulations/invasion. In addition to these cell wall polysaccharides, the list of other glycosylated natural products, which include antibiotics and other xenobiotics, continues to grow. In fact, the only GTs that prokaryotes were believed to lack were those for the synthesis of glycoproteins. However, this group has recently been re-visited given increasing demonstrations of the structure, function, and biosynthesis of glycoproteins in prokaryotes (45-47). The diversity of prokaryotic carbohydrate structures, and the corresponding GTs that synthesize them, is overwhelming. Moreover, striking similarities seem to exist between the glycosylation events of the prokaryotes and the eukaryotes. For example, both kingdoms use nucleoside-donor sugars to assemble oligosaccharide chains, trimming reactions are present in both, and they both use lipid-bound intermediates. Studies with S-layer glycoproteins have also demonstrated that prokaryotes are O-glycosylated at serine/threonine residues and N-glycosylated at asparagine residues (45).

Despite their similarities, glycosylation events in prokaryotes and eukaryotes have critical differences. The absence of intracellular organelles means that polymerization of prokaryotic oligo/polysaccharides takes place entirely in the cytoplasm or in controlled extracellular reactions. It is important to note that the extracellular glycosyltransferase reactions must be coordinated tightly given the lack of a membrane to prevent the loss of substrates/products combined with the lack of nucleotide donors to fuel the reaction process. The following description outlines the processes by which prokaryotic GTs synthesize carbohydrates destined for either protein glycosylation or macromolecular structures. These polysaccharides are typically synthesized by similar mechanisms. However, it should be noted that in bacteria, N-glycosylation occurs independently of protein translocation (47). Biosynthesis begins in the cytoplasm on a lipid-linked carrier. Undecaprenyl is used in the case of peptidoglycan and LPS, whereas a C24-polyisoprenyl is used for protein glycosylation (45-48). Sugars are added to these carriers by successive transfers of either monomeric sugars from nucleotide precursors or by en bloc transfer of assembled oligosaccharides from other lipid-monophosphate carriers. The GTs involved in this process for LPS with mammalian-like structures which help the bacterium mimic host cell polysaccharides for the purpose of evasion/suppression of the host immune response (7, 8). When the oligo/polysaccharide precursors are assembled, they either remain in the cytoplasm or are transported across the membrane(s). However, the mechanism of transfer across the membrane(s) is not understood clearly. At least for capsular and lipopolysaccharides from E. coli and Salmonella enterica, it seems to involve the coordinated process of several membrane spanning and/or ABC transporter proteins (48, 49). Electron microscopy and X-ray crystallography studies have allowed the visualization of the core Wza-Wzc complex for capsular polysaccharide export in E. coli (49). When on the outer-surface of these cells, the oligo/polysaccharides are released into the external environment or transferred to their target proteins/macromolecular structures by the action of external glycosyltransferases. The recently solved three-dimensional structure of the peptidoglycan biosynthesis domain of the penicillin-binding protein (18, 26) of viruses. First, some bacteriophages can glycosylate their own DNA to avoid host restriction endonucleases. This function became particularly apparent when the first structurally solved GT, T4 bacteriophage-encoded α-glucosyltransferase (22) was crystallized more recently in the presence of a short DNA segment (50). An α-glucosyltransferase with the same fold and similar function has since been described in the same T4 virus (51). Second, viral GTs can modify host structures that result in evasion of the immune response or promoting the transmissibility of the released viruses. Modification of host structures in bacteria has been reported to aid in serotype conversion and immunity to infection by other viruses. This phenomenon has been studied extensively with respect to LPS modification in Salmonella and prokaryotes.
Shigella bacterial species (reviewed by Reference 52). Third, viral GTs can alter host metabolisms to promote the release of increased numbers of progeny viruses. This theory is perhaps best exemplified by the modification of ecdysteroid by a baculovirus-encoded glucosyltransferase. Ecdysteroid is an insect hormone involved in the development of Lepidoptera sp. from the larval to the molting or pupating stages. Inactivation of glucosyltransferase through the addition of glucose by virally-encoded UDP-glucosyltransferase, EGT, prevents this development. It allows the virus to monopolize insect resources and facilitates easier spread of the virus when the larva disintegrate (reviewed by Reference 52). Last, some viruses possess GTs that assemble unique virus-encoded products that aid in immune evasion or infection of surrounding hosts. Typically, viral glycoproteins are produced by "hijacking" the host glycosylation pathways associated with the ER-Golgi apparatus. However, the synthesis of the major capsid protein of the Chlorella virus has been found to differ from this paradigm. Paramaecium bursaria chlorella virus-1 encodes at least five putative GTs, of which one of the gene products, A64-9, has been characterized structurally by X-ray crystallography (53). Four of the five GTs are predicted to be localized in the cytoplasm, whereas the last GT is membrane associated. The coordinated activity of these GTs is proposed to assemble the major capsid protein, Vp54, which is independent of the host ER-Golgi apparatus.

Virus-encoded GTs may actually be more common than believed previously because the list of known viral GTs is far from being complete. Viruses are constantly coevolving with their hosts and many can extract genes, which include those for glycosyltransferases, from the host genome for incorporation into their own. If they can confer a selective advantage to their hosts and many can extract genes, which include those for glycosyltransferases, from the host genome for incorporation into their own. They can confer a selective advantage to the virus, they could be incorporated permanently into their genomes, which lead to new virally encoded GT variants.

**Future Directions**

Although many important GTs in glycobiology have been discussed, a need exists for even more exploration of these enzymes, because specific roles are often not defined. Furthermore, it is increasingly apparent that many GTs are related directly to numerous acquired and inherited diseases. For example, the potent B cell toxin from Clostridium difficile is a GT that inactivates Rho-GTP through glucosylation (54). This GT leads to diarrhea, inflammation, and damage to colonic mucosa. In inherited diseases, aberrant glycosylation of O-linked mannose and GlcNAc on dystroglycan can lead to Walker-Warburg syndrome or muscle-eye-brain disease, respectively (55). Sequence variations in the coding genes of xylosyltransferases, XT-I and XT-II, have also been demonstrated to be responsible for altered proteoglycan metabolism. These variations have thus been identified as risk factors for diabetic nephropathy, osteoarthritis, or pseudohyphae elastidum (56). Although it is certainly an abbreviated list, these findings point to the important role of these GTs in disease modifiers in several different pathologies. Therefore, from a disease standpoint, it is prudent that we continue to outline the roles for existing and unknown GTs within a biological context. Then, the potential to cure these diseases, by genetic or drug administered therapy, can be assessed properly.

GTs are also promising for the generation of pharmacologically relevant glycoconjugates. Indeed, the discovery of the catalytic reversibility of AveB from Streptomyces avermitilis proved to be useful because it led to the production of a variety of avermectin variants that may have commercial use in controlling nematodes, insects, and arachnids (4). These results are particularly intriguing given the recently solved three-dimensional structures of two other antibiotic synthesis enzymes: the vancomycin and clindamycin GTs (16, 24) (Fig. 4d). The characterization of new GTs also holds medical relevance as a way to greatly expand existing candidates for vaccine development, because many of these are known to be natural glycoconjugates synthesized by GTs.

Characterization of GTs also has industrial implications. Glycosylation of biopharmaceuticals such as antibodies can have profound effects on their half-life, stability, and activity, which makes GTs attractive tools for engineering proteins that are ideal for prolonged or specialized industrial processes as exemplified in the production of human glycoproteins in yeast (57). Furthermore, the incorporation of several GTs in a single reaction mixture can lead to the synthesis of many and often complex polysaccharides as opposed to lengthy and laborious chemical methods that are less environmentally benign. This mechanism includes the use of metabolically engineered bacteria for large-scale synthesis of complex glycans (58, 59). Ultimately, all processes would be enhanced greatly if the precise catalytic mechanisms of GTs were understood better. To accomplish this task, new enzyme inhibitors and/or substrate analogs are needed. Elucidation of these mechanisms will certainly not only lend insight into the structure-function relationship of these enzymes, but also provide an increased knowledge of what makes GTs uniquely suited to perform certain functions.

**References**


8. Chiu CPC, Watts AG, Laronn LL, Gilbert M, Lim D, Wakarchuk WW, Withers SG, Strynadka NCJ. Structural analysis of the
Glycosyltransferases, Chemistry of


Further Reading

See Also
Glycan Biosynthesis in Mammals
Glycan, Synthesis, Key Reactions of,
Glycolipids, Synthesis of
Post-Translational Modifications to Regulate Protein Function

Hening Lin, Jintang Du, and Hong Jiang. Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York
doi: 10.1002/9780470048672.wecb467

Protein post-translational modifications (PTM) are very important to regulate protein function and to control numerous important biological processes. Here a brief review of commonly found enzyme-catalyzed PTM is given. These PTM include modifications that occur on protein side chains and those that involve protein backbones. The introduction of different PTM is followed by a summary of the molecular basis for the regulation of protein function by PTM. The focus is then given to a few major PTM that play important roles in eukaryotes, such as phosphorylation, methylation, acetylation, glycosylation, ubiquitylation, and proteolysis. For each modification, a description will be given about the residues modified, the enzymatic reaction mechanisms, the major known biological functions, and its relevance to human diseases. At the end, we discuss challenges in identifying new pathways regulated by known PTM and discovering new PTM.

Introduction

The central dogma of molecular biology, DNA is transcribed to mRNA which is then translated to proteins, implies the importance of proteins. After all, it is the proteins that carry out most of the biological functions of a cell. Thus controlling transcription and translation are very important, as they ultimately control what proteins are synthesized in cells and thus control the properties of cells. However, one should not overlook what happens to proteins after they are synthesized. Many chemical modifications can occur to proteins after translation. Collectively, these modifications are called post-translational modifications (PTM). PTM are very important in regulating protein function, which is reflected by the large number of genes devoted to catalyzing PTM. For example, in the human genome (with less than 30,000 genes total), more than 500 kinases catalyze protein phosphorylation (1), and more than 500 proteases catalyze the hydrolytic cleavage of proteins (2). Deregulation in PTM is the cause of various human diseases, as will be explained later in specific PTM sections. Here, a brief review is given on different types of PTM and on how PTM regulate protein function. Some basic principles will be highlighted so that readers who are unfamiliar with PTM can have a quick but comprehensive understanding of PTM. The recent book on PTM by Professor Walsh from Harvard Medical School provides a more complete description of PTM (3). Where appropriate, references on specific PTM will also be given in different sections for additional information. The abbreviations used are cataloged in Table 1 to help readers who are not familiar with the biological language.

Types of post-translational modifications

PTM can be enzyme-catalyzed and thus controlled carefully, or they can be nonenzymatic with less control. For example, protein glycation during hyperglycemia is a nonenzymatic PTM that accounts for some symptoms of diabetes (4). Protein nitrosylation on Cys residues is another nonenzymatic PTM that can affect protein function (5). Coordination by metal ions can also be considered as a PTM. For many proteins, metal binding is crucial for maintaining the correct structure or the enzymatic activity (6). Here, the focus will be given to enzyme-catalyzed PTM. Figures 1 and 2 show many commonly found enzyme-catalyzed PTM.
Post-Translational Modifications to Regulate Protein Function

Figure 1. Major enzyme-catalyzed PTM that modify protein side chains.
Figure 2. A few PTM that involve protein backbone.
### Table 1: List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABL</td>
<td>A tyrosine kinase encoded from abl (Abelson) gene, the fusion protein ABL-BCR is involved in inhibition of apoptosis in chronic myelogenous leukemia cells</td>
</tr>
<tr>
<td>AC</td>
<td>Adenylate cyclase, converts ATP to cyclic AMP</td>
</tr>
<tr>
<td>AcLys</td>
<td>Acetyllysine</td>
</tr>
<tr>
<td>ACP</td>
<td>Acyl carrier protein, found in fatty acid synthases and polyketide synthases, functions to carry the elongating fatty acyl chain</td>
</tr>
<tr>
<td>ADAM</td>
<td>A disintegrin and metalloprotease, a family of proteases that hydrolyze off extracellular portions of transmembrane proteins</td>
</tr>
<tr>
<td>ADP</td>
<td>Adenosine 5’-diphosphate</td>
</tr>
<tr>
<td>Apaf-1</td>
<td>Apoptotic protease activation factor-1, a cytosolic protein involved in cell death or apoptosis, interacts with cytochrome c to activate caspase 9</td>
</tr>
<tr>
<td>AT</td>
<td>Acyltransferase, found in fatty acid synthases and polyketide synthases, adds a malonyl group to the holo form of the ACP domain</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine 5’-triphosphate</td>
</tr>
<tr>
<td>Bcl2</td>
<td>Named from B-cell lymphoma 2, an antiapoptotic protein</td>
</tr>
<tr>
<td>BCR</td>
<td>A protein encoded from breakpoint cluster region gene, has serine/threonine kinase activity. Fusion with abl protein causes leukemia</td>
</tr>
<tr>
<td>cAMP</td>
<td>3’-5’-cyclic adenosine monophosphate</td>
</tr>
<tr>
<td>CARD</td>
<td>Caspase recruitment domain, mediates the formation of larger protein complexes via direct interactions between individual cards, involved in the regulation of caspase activation and apoptosis</td>
</tr>
<tr>
<td>CARM1</td>
<td>Coactivator-associated arginine(R) methyltransferase 1, methylates Arg17 and Arg26 residues on Histone H3</td>
</tr>
<tr>
<td>Cbl-b</td>
<td>Ubiquitously expressed homolog of Cbl, a mammalian protein involved in cell signaling and protein ubiquitination, named after Casitas B-lineage Lymphoma</td>
</tr>
<tr>
<td>CBP</td>
<td>CREB binding protein, a transcriptional co-activating protein</td>
</tr>
<tr>
<td>CD2</td>
<td>Cellular differentiation marker 2, a cell adhesion protein found on the surface of T cells and natural killer cells</td>
</tr>
<tr>
<td>CDK</td>
<td>Cell-division kinases, serine/threonine kinases, activated by association with cyclins and involved in regulation of the cell cycle, transcription and mRNA processing</td>
</tr>
<tr>
<td>CHD1</td>
<td>Chromodomain helicase DNA-binding protein 1, interacts with methylated Lys4 on Histone H3</td>
</tr>
<tr>
<td>CLOCK</td>
<td>A protein named from circadian locomotor output cycles kaput gene, regulating circadian rhythm</td>
</tr>
<tr>
<td>CML</td>
<td>Chronic myelogenous leukemia, a form of leukemia characterized by the increased and unregulated growth of predominantly myeloid cells in the bone marrow and the accumulation of these cells in the blood</td>
</tr>
<tr>
<td>CREB</td>
<td>Camp response element binding proteins, as transcription factors, bind to certain sequences called camp response elements (CRE) in DNA and thereby increase or decrease the transcription of certain genes</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyto c</td>
<td>Cytochrome c, a small heme protein associated with the inner membrane of the mitochondria and released in response to pro-apoptotic stimuli</td>
</tr>
<tr>
<td>DED</td>
<td>Death effector domain, a protein interaction domain found in inactive procaspases and proteins that regulate caspase activation in the apoptosis cascade</td>
</tr>
<tr>
<td>DH</td>
<td>Dehydratase, found in fatty acid synthases and polyketide synthases, dehydrates the ω-H of acyl thioester</td>
</tr>
<tr>
<td>DNAse</td>
<td>Deoxyribonuclease, catalyzes the hydrolytic cleavage of phosphodiester linkages in the DNA backbone</td>
</tr>
<tr>
<td>ER</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>ER</td>
<td>Enoylreductase, found in fatty acid synthases and polyketide synthases, reduces the enoyl of enoyl thioester to the saturated thioester</td>
</tr>
<tr>
<td>ERK</td>
<td>Extracellular signal-regulated kinase, activates many transcription factors and some downstream protein kinases, involved in functions including the regulation of mitosis, meiosis, and postmitotic functions in differentiated cells</td>
</tr>
<tr>
<td>Factor IX</td>
<td>One of the serine proteases of the coagulation system</td>
</tr>
<tr>
<td>FAD</td>
<td>Flavin adenine dinucleotide</td>
</tr>
<tr>
<td>FADD</td>
<td>Fas-associated protein with death domain, connects the Fas-receptor and other death receptors to caspase-8 through its death domain to form the death inducing signaling complex during apoptosis</td>
</tr>
<tr>
<td>FHA domain</td>
<td>Forkhead-associated domain, a phosphospecific protein-protein interaction motif involved in checkpoint control of the cell cycle</td>
</tr>
<tr>
<td>GcoS</td>
<td>A yeast transcriptional adaptor that has histone acetyltransferase activity</td>
</tr>
<tr>
<td>GDP</td>
<td>Guanosine-5’-diphosphate</td>
</tr>
<tr>
<td>GFP</td>
<td>Green fluorescent protein</td>
</tr>
<tr>
<td>GPCR</td>
<td>G protein-coupled receptor, a transmembrane receptor that senses molecules outside the cell and activates inside signal transduction pathways and cellular responses</td>
</tr>
<tr>
<td>GPl</td>
<td>Glycosylphosphatidylinositol, a glycolipid that can be attached to the C-terminus of a protein during post-translational modification</td>
</tr>
<tr>
<td>GPK</td>
<td>Glycogen phosphorylase kinase, a serine/threonine-specific protein kinase which activates glycogen phosphorylase by phosphorylation</td>
</tr>
<tr>
<td>Grb2</td>
<td>Growth factor receptor-bound protein 2, an adaptor protein involved in signal transduction/cell communication</td>
</tr>
<tr>
<td>I-2P</td>
<td>Guanosine-5’-bisphosphate</td>
</tr>
<tr>
<td>HDAcS</td>
<td>Histone deacetylases, remove acetyl groups from ε-N-acetyl-lysine residues on histones</td>
</tr>
<tr>
<td>hDOT1L</td>
<td>Human DOT1-like protein, methylates histone H3 at Lys79. (DOT1: Yeast disruptor of telomeric silencing-1)</td>
</tr>
<tr>
<td>HECT</td>
<td>Homologous to E6-AP C terminus, mediates E2 binding and ubiquitination</td>
</tr>
<tr>
<td>HIF</td>
<td>Hypoxia inducible factor, a transcription factor that responds to changes in available oxygen in the cellular environment, specifically to decreases in oxygen or hypoxia</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>hnRNPs</td>
<td>Heterogenous nuclear ribonucleoproteins, which forms complexes with pre-mRNA and mRNA and shuttles between the nucleus and the cytoplasm</td>
</tr>
<tr>
<td>HP1</td>
<td>Heterochromatin protein 1, binds to heterochromatin and interacts with numerous partner proteins to organize the higher-order structure of heterochromatin</td>
</tr>
<tr>
<td>IgG</td>
<td>Immunoglobulin G, one antibody isotype</td>
</tr>
<tr>
<td>IKK</td>
<td>Inhibitor of NF-κb kinase, which phosphorylates inhibitor of NF-κb to promote its degradation to release NF-κb dimers to translocate to the nucleus and activate transcription of target genes</td>
</tr>
<tr>
<td>IP7</td>
<td>Inositol pyrophosphate, a proposed physiological phosphate donor</td>
</tr>
<tr>
<td>JHDM</td>
<td>JmjC domain-containing histone demethylase</td>
</tr>
<tr>
<td>JmjC</td>
<td>Jumonji domain-containing, a novel demethylase signature motif</td>
</tr>
<tr>
<td>KR</td>
<td>Ketoreductase, found in fatty acid synthases and polyketide synthases, reduces the β-ketoacyl thioester</td>
</tr>
<tr>
<td>KS</td>
<td>Ketosynthase, found in fatty acid synthases and polyketide synthases, carries out C-C bond-forming chain elongation step</td>
</tr>
<tr>
<td>LSD1</td>
<td>Lysine-specific demethylase 1, demethylates histone H3 at lysine 9</td>
</tr>
<tr>
<td>MAP</td>
<td>Mitogen-activated protein kinase, serine/threonine-specific protein kinases that respond to extracellular stimuli (mitogens) and regulate various cellular activities, such as gene expression, mitosis, differentiation, and cell survival/apoptosis</td>
</tr>
<tr>
<td>MAPK</td>
<td>Mitogen-activated protein kinase, serine/threonine-specific protein kinases that respond to extracellular stimuli (mitogens) and regulate various cellular activities, such as gene expression, mitosis, differentiation, and cell survival/apoptosis</td>
</tr>
<tr>
<td>MLys</td>
<td>ε-N-monomethyllysine</td>
</tr>
<tr>
<td>MeLys</td>
<td>ε-N-dimethyllysine</td>
</tr>
<tr>
<td>Me2Lys</td>
<td>ε-N-trimethyllysine</td>
</tr>
<tr>
<td>MAPK/ERK</td>
<td>MAPK/ERK kinase, activates a MAP kinase or ERK through phosphorylation</td>
</tr>
<tr>
<td>MIO</td>
<td>4-methyl-5-imidazole-5-one</td>
</tr>
<tr>
<td>MOZ</td>
<td>Monocytic leukemia zinc finger protein, a histone acetyltransferase implicated in leukemogenic and other tumorigenic processes, regulates expression of genes required for proliferation and repopulation of potential of stem cells in the hematopoietic compartment</td>
</tr>
<tr>
<td>NAD</td>
<td>Nicotinamide adenine dinucleotide</td>
</tr>
<tr>
<td>NGFp1</td>
<td>Nerve growth factor p1, a secreted protein which induces the differentiation and survival of particular target neurons, belonging to neurotrophins protein family</td>
</tr>
<tr>
<td>OStase</td>
<td>Oligosaccharyltransferase</td>
</tr>
<tr>
<td>PARs</td>
<td>Protein Arg deiminases, hydrolyzes the guanidino side chain of Arg residues to citrulline residues in proteins</td>
</tr>
<tr>
<td>PARP-1</td>
<td>Poly(ADP-ribose) polymerase-1, catalyzes the transfer of poly ADP-ribose to substrate proteins by using NAD as substrate, involved in cellular response to DNA damage and DNA metabolism</td>
</tr>
<tr>
<td>PKA</td>
<td>Protein kinase A, a family of kinases whose activity are dependent on the level of cyclic AMP, involved in the regulation of glycogen, sugar, and lipid metabolism</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRMT</td>
<td>Protein Arg(R) methyltransferase, catalyzes the transfer of methyl group from S-adenosylmethionine to the guanidino nitrogen atoms of arginine residues</td>
</tr>
<tr>
<td>pSer</td>
<td>Phosphoserine</td>
</tr>
<tr>
<td>PTB</td>
<td>Phosphotyrosine binding</td>
</tr>
<tr>
<td>pThr</td>
<td>Phosphothreonine</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modifications</td>
</tr>
<tr>
<td>pTyr</td>
<td>Phosphotyrosine</td>
</tr>
<tr>
<td>RAIDD</td>
<td>RIP-associated ICH-1/CED-3 homologous protein with a death domain, functions as an adaptor in recruiting the death protease ICH-1 to the TNFR-1 signaling complex (ICH: Ice and ced-3 homolog; TNFR: tumor necrosis factor receptor)</td>
</tr>
<tr>
<td>RING</td>
<td>Really interesting new gene. Ring proteins are components of ubiquitin E3 enzyme complexes.</td>
</tr>
<tr>
<td>RIP</td>
<td>Receptor-interacting protein</td>
</tr>
<tr>
<td>SAH</td>
<td>S-adenosylhomocysteine</td>
</tr>
<tr>
<td>SAHA</td>
<td>Vorinostat, suberoylanilide hydroxamic acid, brand name Zolinza, a class of agents known as histone deacetylase inhibitors, as a drug for the treatment of cutaneous T cell lymphoma (a type of skin cancer)</td>
</tr>
<tr>
<td>SAM</td>
<td>S-adenosylmethionine</td>
</tr>
<tr>
<td>SCF</td>
<td>Skp1-Cullin-F Box, a multi-protein complex catalyzing the ubiquitylation of proteins destined for proteasomal degradation</td>
</tr>
<tr>
<td>SET</td>
<td>Suppressor of variegation-Enhancer of zeste-Trithorax. SET domains have methyltransferase activity.</td>
</tr>
<tr>
<td>Set8</td>
<td>A novel human SET domain-containing protein, which specifically methylates H4 at Lys20</td>
</tr>
<tr>
<td>Set9</td>
<td>A novel human SET domain-containing protein, which specifically methylates H3 at Lys4</td>
</tr>
<tr>
<td>SH2</td>
<td>Src homology 2, a phosphotyrosine-recognition protein domain of about 100 amino acid residues first identified as a conserved sequence region among the oncoproteins Src and Fps</td>
</tr>
<tr>
<td>SMAD</td>
<td>Proteins homologs of both the drosophila protein, mothers against decapentaplegic (MAD) and the C. Elegans protein SMA, as signal-activated transcription factors regulated by the TGF-β superfamily</td>
</tr>
<tr>
<td>Smyd</td>
<td>Proteins containing SET and MYND domain. MYND encoded mynd (myosin) gene, which have histone methyltransferase activity</td>
</tr>
<tr>
<td>snRNPs</td>
<td>Small nuclear ribonucleoproteins, combining with pre-mRNA and various proteins to form spliceosomes to removes introns from pre-mRNA segment</td>
</tr>
<tr>
<td>Sos</td>
<td>Son of sevenless, a guanine nucleotide exchange factor that activates Ras</td>
</tr>
<tr>
<td>STAT</td>
<td>Signal transducers and activators of transcription, proteins which are involved in the development and function of the immune system</td>
</tr>
<tr>
<td>SUMO</td>
<td>Small ubiquitin-like modifier, a family of small proteins that are covalently attached to and detached from other proteins in cells to modify their functions</td>
</tr>
</tbody>
</table>
Post-Translational Modifications to Regulate Protein Function

Table 1 (Continued)

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAF10</td>
<td>TATA box-binding protein-associated factor 10, a component of the general transcription factor complex TFIIID and the TATA box-binding protein (TBP)-free TAF-containing complex</td>
</tr>
<tr>
<td>TIF2</td>
<td>Transcription intermediary factor 2, a transcriptional coregulatory protein which contains several nuclear receptor interacting domains and an intrinsic histone acetyltransferase activity</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor, a protein that binds to specific region of DNA by DNA binding domains and mediates the transcription from DNA to RNA</td>
</tr>
<tr>
<td>TGFβ1</td>
<td>Transforming growth factor β1, a secreted protein that performs many cellular functions, including the control of cell growth, cell proliferation, cell differentiation and apoptosis, belonging to the transforming growth factor beta superfamily of cytokines</td>
</tr>
<tr>
<td>TGN</td>
<td>Trans Golgi network, a part of the golgi apparatus in cells</td>
</tr>
<tr>
<td>TOPA</td>
<td>2,4,5-trihydroxyphenylalanine</td>
</tr>
<tr>
<td>TRADD</td>
<td>Tumor necrosis factor receptor-associated protein with death domain, an adapter protein that recruits other proteins to the cytoplasmic TNF (tumor necrosis factor) receptor complex, involved in apoptosis</td>
</tr>
<tr>
<td>UAP</td>
<td>Ubiquitin activating protein</td>
</tr>
<tr>
<td>UBA</td>
<td>Ubiquitin binding associated domain, one class of ubiquitin binding domains</td>
</tr>
<tr>
<td>UBD</td>
<td>Ubiquitin binding domain, which binds mono- or poly-ubiquitin</td>
</tr>
<tr>
<td>UBP</td>
<td>Ubiquitin-specific protease, hydrolyzes both linear and branched Ub modifications</td>
</tr>
<tr>
<td>UDP</td>
<td>Uridine diphosphate</td>
</tr>
<tr>
<td>UMP</td>
<td>Uridine monophosphate</td>
</tr>
</tbody>
</table>

As can be observed from Fig. 1, most PTM happen to protein side chains. Typically, the side chains involved are nucleophilic, such as Cys (palmitoylation, isoprenylation, disulfide bond formation, ADP-ribosylation), Lys (acetylation, methylation, ubiquitinylation), Arg (methylation, ACP-ribosylation), Asp/Glu (methylated, polyADP-ribosylation), Ser/Thr (phosphorylation, O-glycosylation), and Tyr (phosphorylation). Weaker nucleophiles are also used, such as the side chain amide nitrogen in Asn (in N-glycosylation), the C-2 position of Trp (in C-glycosylation), and the C-2 position of His (in diphthamide). In amider mediated reactions catalyzed by transglutaminases and polyglutamylation/polyglycylation reactions that happen to Glu residues, the ε-NH$_2$ from Lys or α-NH$_2$ from Glu/Gly acts as the nucleophile, whereas the side chain of Glu or Glu serves as the electrophile. In addition, several amino acid side chains can be oxidized, such as Pro, Lys, Asn, Tyr, Trp, and Cys, to give oxidized amino acids.

A few PTM reactions also involve changes in protein backbone. These reactions include the hydrolytic cleavage of the peptide backbone by proteases, the anchoring of proteins to glycosylphosphatidylinositol (GPI) or cholesterol, and the C-terminal amide formation by oxidative cleavage of glycine residues. Some PTM involve changes in both the side chain and the main chain, such as the formation of 4-methylidene-5-imiazole-5-one (MIO) prosthetic group in deaminases and aminomutases, the formation of the fluorophore in GFP (green fluorescent protein), and the formation of pyruvamide in decarboxylases (Fig. 2).

Molecular basis for the regulation of protein function by PTM

As with all other chemical species, protein structure determines protein function. PTM can regulate protein function because they can change protein structure. The structure change introduced by PTM can be local and small. For example, methylation of Lys residues makes the side chain more hydrophobic without changing protein backbone conformation significantly (at least based on crystal structures in which methylated and unmethylated histone peptides are bound by another protein [2]), whereas phosphorylation can change the backbone conformation within a limited region of a protein by charge-pairing with nearby Arg residues or by interacting with main chain NH and helical
Post-Translational Modifications to Regulate Protein Function

Changing protein structure to turn on/off catalytic activity of enzymes

The best-known PTM that is widely used to regulate enzymatic activity is phosphorylation. Phosphorylation regulates the activity of many enzymes by different mechanisms. For example, glycogen phosphorylase is activated allosterically by phosphorylation at Ser14, whereas *Escherichia coli* isocitrate dehydrogenase is inhibited by phosphorylation because of the block of substrate access to the active site (9). The most interesting and very important catalytic activity regulated by phosphorylation is protein kinase activity. Most protein kinases are activated by phosphorylation of Thr/Tyr residue(s) in the activation segment. The structural changes induced by phosphorylation, which are illustrated in Fig. 3 with ERK (extracellular signal-regulated kinase), convert the inactive kinases to active kinases (8). The regulation of protein kinase activity by phosphorylation bears enormous biological significance because protein phosphorylation is important in signal transduction, and the control of downstream kinase activity via phosphorylation by upstream kinase is one major method to propagate signals to downstream partners, as will be elaborated later.

Proteolysis is another way to control enzymatic activity, although unlike phosphorylation, the change in activity is irreversible. Many proteases are synthesized as inactive precursors (zymogens) that have to be cleaved by proteolysis to become active. These precursors include proteases that are secreted into digestive tracts or lysosomes, the catalytic active subunit(s) in the eukaryotic 20S proteasome that are activated by self-cleavage (10), and the effector caspases involved in apoptosis that are activated by initiator caspases-mediated cleavage (11).

Changing protein structure to create or to mask recognition motifs

Many PTM exert their biological functions by creating recognition motifs to recruit binding partners (12) or by masking recognition motifs to disrupt existing interactions. Phosphorylated Ser/Thr residues can be recognized by proteins that contain 14-3-3 domains, FHA (forkhead-associated) domains, SMAD (proteins homologs of both the drosophila protein, mothers against decapentaplegic (MAD) and the *Caenorhabditis elegans* protein Sma) domains, and several other domains (13). Phosphorylated Tyr residues can recruit proteins that contain SH2 (Src homology 2) domains and PTB (phosphotyrosine binding) domains (14). Acetyl Lys residues can be recognized by proteins with bromodomains (15, 16), and methylated Lys residue can be recognized by proteins with chromodomains and Tudor domains (17). The ubiquitin and ubiquitin-like protein tags can also be recognized by various protein domains that mediate the biological function of modification with these protein tags (18, 19). The structures of a few domains dedicated to recognition of post-translationally modified residues are shown in Fig. 4. Typically, domains that recognize post-translationally modified residues have specificities in that they recognize not only the modified residue, but also the local structure in which the residue resides. The specific recognition of PTM in different contexts is the key to understand many biological consequences of PTM, as will be explained in more detail in particular PTM sections later.

In addition to creating recognition motifs to recruit proteins, a few PTM can also increase interaction with other species, such as the lipid bilayer of different cellular membranes. These modifications include the formation of GPI-anchored proteins (20), protein myristoylation on the α-amino group of the N-terminal dipole (8). In contrast, some PTM can alter protein overall structure more dramatically, such as the proteolytic cleavage of proteins into smaller fragments, or the addition of protein tags like ubiquitin. These structure changes, small or big, are the basis for the biological functions of different PTM and typically lead to one or more of the consequences described below.
Post-Translational Modifications to Regulate Protein Function

Figure 4 Structures of a few dedicated domains that recognize post-translationally modified residues. (a) SH2 domain of v-Src in complex with pTyr peptide (pTyr-Val-Pro-Met-Leu). Residues Arg12, Arg32, Ser34, Thr36, and Lys60 from the SH2 domain interact with pTyr (figure made using PDB 1SHA); (b) Bromodomain of yeast histone acetyltransferase Gcn5 in complex with AcLys peptide (histone H4 residues 15–29, AK(Ac)RHRKILRNSIQGI). Bromodomain residues Pro351, Gln354, Tyr364, Met372, Val399, and Asn407 interact with AcLys (with some of the interaction is mediated by water molecules, figure made using PDB 1E6I); (c) Chromodomain of HP1 in the complex with histone H3 Me3Lys9 (figure made using PDB 1KNE). Chromodomain residues Tyr 24, Trp 45, Tyr 48 and Gln 52 bind MeLys; (d) UBA domain of Cbl-b in complex with ubiquitin (figure made using PDB 200B). UBA domain residues Asp933, Ala937, Met940, Phe946, and Lys950 interact with ubiquitin residues Leu8, Ile44, Ala46, Gly47, Gln49, His68, and Val70. UBA: ubiquitin binding associated.

Gly (21), protein C-terminal prenylation on Cys residues (22), and protein palmitoylation on Cys residues that are close to membrane surface (23). These lipid modifications occur to many signaling proteins, which include G protein-coupled receptors and small G proteins, and they play important roles in signal transduction and membrane trafficking (24).

Adding functional groups to allow catalysis

Typically, proteins are formed with the most common 20 amino acids, which only offer a limited number of choices of functional groups for catalyzing different reactions. The limit in the number of functional groups is complemented by the use of various cosymes or cofactors, many of which are attached covalently to the corresponding enzymes. One class of PTM with this function is the addition of “swinging arm” prosthetic groups (biotin, phosphopantetheine, and lipoic acid) to proteins (25). Biotin is used as a carrier of CO2 in carboxylation reactions, and the disulfide bond in lipoic group is used as an electron carrier and acyl carrier in 2-keto acid dehydrogenases. The phosphopantetheine group provides a thiolate as the carrier of acyl chains and is used in fatty acids synthases, polyketide synthases, and nonribosomal peptide synthetases (26). Although a thiolate side chain can also be provided by Cys, the longer phosphopantetheine can shuttle the acyl chains to different catalytic domains, which allows multiple reactions to occur in sequence on the acyl chains (Fig. 5). This “swinging arm” catalysis, which is also enabled by bioinlilation and lipoylation, cannot be achieved by natural proteinogenic amino acids with shorter side chains.

Another type of PTM provides new functional groups for enzyme catalysis by oxidation of side chain. These include TOPA (2,4,5-trihydroxyphenylalanine) quinone in amine oxidases (27), tryptophan tryptophanyl quinone in methylenecarboxylases, and formylglycine in sulfatases. (28) Main chain modifications can also generate prosthetic groups for enzyme catalysis, such as the MIO group in His/Phe ammonia-lyase (29, 30) and Tyr aminomutases, (31) and the pyruvyl group in decarboxylases (Fig. 6) (32). The formation of these cofactors by PTM extends the catalytic power of enzymes greatly, which enables them to catalyze chemistry that is difficult with just the side chains of the 20 amino acids commonly found in proteins.

Locking proteins into the correct structures or increasing protein stability

The major type of PTM that has this function is protein disulfide bond formation (33). Disulfide bonds are more stable thermodynamically than the reduced thiolis in an oxidizing environment. In eukaryotes, proteins that undergo the secretory pathway start to form disulfide bonds once they are translocated into the endoplasmic reticulum (ER) lumen, which is an oxidizing environment. These disulfide bonds help to stabilize
Post-Translational Modifications to Regulate Protein Function

Figure 5  Fatty acid biosynthesis catalyzed by fatty acid synthases. The growing acyl chain is tethered to the phosphopantetheinylated ACP domain, which enables it to undergo cycles of condensation, ketone reduction, dehydration, and enol reduction catalyzed by different domains: AT, acyltransferase; ACP, acyl-carrier protein; KS, ketosynthase; KR, ketoreductase; DH, dehydratase; ER, enoylreductase.

Figure 6  Post-translationally generated cofactors provide functional groups to allow catalysis. The mechanisms of TOPA quinone in amine oxidases, MIO in deaminase, and pyruvamide in decarboxylase are shown.
the desired protein structure by locking the protein in a certain conformation, and perhaps to assist protein folding too. Many secreted proteins later undergo proteolysis in the Golgi to give smaller fragments (see the proteolysis section below). In this case, disulfide bonds also serve to link the fragments covalently to maintain a certain structure. One textbook example is insulin, which is produced as a single peptide chain that later undergoes several proteolysis steps, and the mature insulin consists of two chains connected via two disulfide bonds (Fig. 7) (34). The light and heavy chains of antibodies are connected by disulfide bonds. Another PTM that can increase protein stability is glycosylation. For example, erythropoietin N-glycosylation has been found to increase its in vivo lifetime (35), which is probably because of the blocking of tissue proteases action by carbohydrate modifications.

Exploration of major PTM

In this section, a few major PTM will be explored in more details. For each PTM discussed, a brief introduction on the PTM reaction and the enzymes catalyzing the reaction will be given. A few biological processes that involve the PTM will be explained to demonstrate the important function of the PTM in biology.

Phosphorylation

Protein phosphorylation typically occurs on Ser, Thr, and Tyr residues (Fig. 1), although His and Asp residues can also be phosphorylated as in bacteria two-component signal transduction systems. The universal phosphate donor is adenine triphosphate (ATP, Fig. 8), and the reaction is catalyzed by more than 500 kinases in humans. Many kinases are Ser/Thr specific, some are Tyr specific, whereas some have dual specificity. It was reported that inositol pyrophosphate (IP7) can also serve as phosphate donor in protein phosphorylation (36). However, the reaction is not enzyme catalyzed and the physiologic relevance is not proven yet.

The large number of protein kinases in the human genome reflects that this PTM is widely occurring and regulates numerous biological processes. The most well understood function is signal transduction, because phosphorylation of proteins can turn on/off catalytic activity or create recognition motif to recruit other protein partners, thus allowing signal to propagate. In accord with its role in signal transduction, protein phosphorylation is reversible so that the signaling process can be terminated as needed. The removal of the phosphate group is catalyzed by phosphatases (Fig. 8).

Two signaling processes will be discussed here to illustrate how protein phosphorylation can play a critical role in cell signaling. A more detailed description of these two signaling processes can be found in the Molecular Cell Biology textbook by Lodish et al. (34). The first one, which is shown in Fig. 9, involves protein kinase A (PKA), which can be activated by cyclic AMP (cAMP) (37). PKA at resting state exists as an inactive tetramer that consists of two copies of a regulatory subunit and two copies of the catalytic subunit. Hormones that signal through G-protein coupled receptors can activate the trimeric G protein, which in turn can activate an effector enzyme, adenylate cyclase (38). Adenylate cyclase catalyzes the formation of cAMP from ATP (39), which results in the increase in cAMP concentration. Binding of cAMP to the regulatory subunits of PKA dissociate the inactive tetramer, which releases the catalytic subunit of PKA. The catalytic subunit can then be activated by phosphorylation at the activation loop. Activated PKA can phosphorylate many different substrates and produce both short-term and long-term effects. Short-term effects come from the change of the catalytic activities of substrate proteins on phosphorylation by PKA. The substrates of PKA include proteins involved in glycogen synthesis and degradation, such as glycogen phosphorylase kinase and glycogen synthase...
(a) Catalytic mechanisms of protein kinases; (b) Catalytic mechanism of bimetallic pSer/pThr or dual specificity protein phosphatases; (c) Catalytic mechanism of pTyr phosphatases.

Figure 8 Kinase-catalyzed phosphorylation and phosphatase-catalyzed dephosphorylation reactions. (a) Catalytic mechanisms of protein kinases; (b) Catalytic mechanism of bimetallic pSer/pThr or dual specificity protein phosphatases; (c) Catalytic mechanism of pTyr phosphatases.

(40) Phosphorylation of these proteins by PKA leads to activation of glycogen degradation and inhibition of glycogen synthesis. Long-term effects come from the changes in gene transcription. PKA can affect transcription by phosphorylating CREB (cAMP response element binding proteins) and other transcription factors (41). On phosphorylation, CREB can bind to specific regions of the chromosomal DNA, and it can recruit the basal transcription machinery via CBP/CREB binding protein/P300 to activate the transcription of certain genes.

The second example of cell-signaling process that involves protein phosphorylation is receptor tyrosine kinase signaling (Fig. 10) (42). Receptor tyrosine kinases are transmembrane proteins with an extracellular ligand-binding domain and an intracellular tyrosine kinase domain. Ligand binding to the extracellular domain triggers receptor dimerization and/or activation, so that the intracellular catalytic domains from two receptor protein molecules can phosphorylate each other at the activation segment. This transphosphorylation activates the catalytic domain so that it can phosphorylate other Tyr residues in the receptor and other substrate proteins. These phosphorylated Tyr residues then recruit protein-binding partners that contain SH2 or PTB domains that recognize specific phosphorylated Tyr residues. One of the proteins recruited is Grb2 (growth factor receptor-bound protein 2), which contains an SH2 domain. Grb2 in turn recruits Sos (son of sevenless), which is a guanine nucleotide exchange factor for the G protein Ras. Sos catalyzes the exchange of Ras-bound GDP (guanosine-5’-diphosphate) for GTP (guanosine-5’-triphosphate), which converts Ras to the activated form. Activated Ras can bind to and activate Raf, which is the most upstream kinase in the MAP kinase (Mitogen-activated protein kinase) cascade (43). By phosphorylation of MEK (MAPK/ERK kinase, a dual specificity MAP kinase kinase) on the activation segment, Raf activates MEK, which in turn...
Post-Translational Modifications to Regulate Protein Function

Figure 9 The signaling process that involves G protein-coupled receptors (GPCR) and PKA. (1) Binding of hormone produces conformational change in the GPCR; (2) GPCR binds to Gs protein; (3) GDP bound to Gs is replaced by GTP and the β and γ subunits of Gs dissociate from the α subunit; (4) Gs subunit binds to adenylate cyclase (AC), which activates the synthesis of cAMP (4a), the hormone tends to dissociate, and hydrolysis of GTP to GDP causes Gs α to dissociate from adenylate cyclase and binds to Gβγ, which regenerates a conformation of Gs that can be activated by an GPCR hormone complex (4b); (5) dissociation of regulatory subunits (R) from PKA as cAMP concentration increases; (6) subsequent activation of the catalytic subunits (C) by phosphorylation in the activation loop generates the fully active kinase; (7) activated PKA can phosphorylate glycogen phosphorylase kinase (GPK) and other enzymes, which leads to activation of glycogen degradation and inhibition of glycogen synthesis; and (8) PKA can affect transcription by phosphorylating the transcription factor CREB.

phosphorylates and activates ERK. Activated ERK can phosphorylate many transcription factors, which leads to changes in gene transcription and ultimately cell division/differentiation.

The two examples mentioned above illustrate basic principles how protein phosphorylation serves specific biological purposes. Although different kinases might be involved in diverse pathways, the molecular mechanism for the regulation of protein function by phosphorylation is similar: By changing protein structure, phosphorylation can turn on/off the catalytic activity of a protein, or create/mask recognition motif for binding by other molecules.

The 500 or so protein kinases in the human genome regulate numerous biological processes. Consequently, deregulation of protein phosphorylation can lead to various diseases, among which cancer is the most prominent one. Accordingly, kinase inhibitors are being sought for treating various cancers. One best understood example is chronic myeloid leukemia, which is caused by chromosomal abnormality that fuses a kinase ABL (encoded from Abelson gene) with another protein BCR (encoded from breakpoint cluster region gene) (44). The BCR-ABL fusion protein was shown to be sufficient to cause chronic myeloid leukemia in mice. Imatinib mesylate (Gleevec; Novartis Pharmaceuticals, East Hanover, N.J) is a clinically used BCR-ABL inhibitor to treat CML (chronic myelogenous leukemia). The receptor tyrosine kinase and MAP kinase-signaling pathway mentioned above are key pathways that regulate cell proliferation and differentiation; frequently, tumor cells have mutations in proteins involved in this pathway (45). This pathway has thus been studied intensively for the search of cancer drugs. Other kinases, such as cell-division kinases (CDKs), have also been targeted for therapeutics (46). In addition, because phosphatases reverse the effects of kinases, mutations in phosphatases have been indicated in human diseases such as cancer, diabetes, and neurologic disorders (47).
Post-Translational Modifications to Regulate Protein Function

Figure 10  Receptor tyrosine kinase signaling process and the activation of MAP Kinase. (1) Binding of hormone to the receptor causes activation of the kinase activity of the receptor, which leads to phosphorylation of Tyr residues; (2) pTyr residues recruit GRB2, which in turn recruit Sos; (3) Sos promotes exchange of GTP for GDP in Ras, which leads to the active Ras-GTP complex. Then, Sos dissociates from the active Ras; (4) active Ras binds to and activate the kinase Raf; (4a) hormone can dissociate from the receptor; (4b) activated Raf phosphorylates and activates MEK; (6) activated MEK phosphorylates and activates of MAP kinase; (7) activated MAP kinase can phosphorylate transcription factors (TF); and (8) phosphorylated translation factors then bind to DNA and lead to changes in gene transcription and ultimately cell division/differentiation.

Acetylation

Acetylation of Lys residues is a very well known PTM because of histone acetylation, which is involved in transcriptional regulation of genes. The acetyl group comes from Acetyl-CoA, and typically, the acetyl acceptor is Lys residues (Fig. 11). Histone acetylation correlates with transcription activation, and accordingly, histone acetyltransferases (HATs) are normally multidomain proteins associated with transcription activator/coactivator complexes (48). The correlation of histone acetylation with transcription activation can be explained by the relaxation of the chromatin structure on histone acetylation and the recruitment of other proteins via acetyl Lys. In eukaryotic cells, chromosomal DNA wrap around core histone octamers consisted of two copies each of histone H2A, H2B, H3 and H4 (49). The complex formed between the histone octamer and the DNA associated with it is called a nucleosome. Nucleosomes can pack into a more condensed structure. Evidence suggests that the tight packing suppresses transcription, whereas transcription activation correlates with relaxed chromatin structure. The N-terminal tails of the histones have many Lys and Arg residues, among other residues, that can be modified post-translationally. No detailed structure information is available to explain how histone tail modification affects nucleosome packing. However, intuitively, masking the positive charges on histones by Lys acetylation can decrease the interaction with negatively charged DNA, which loosens the chromatin structure (50). In addition, acetylated Lys residues can be recognized by proteins that contain bromodomains (Fig. 4) (16, 51), which serve to recruit other proteins (including chromatin remodeling complexes) that help to activate the transcription of the gene.

Histone acetylation not only affects transcription, but also affects other processes that involve DNA, such as nucleosome assembly, heterochromation formation, and DNA repair (52). The acetylation/deacetylation of different Lys residues can have different biological effects. For example, histone H4 Lys5, 8, and 12 acetylation are involved in nucleosome assembly, heterochromation formation, and DNA repair (52). The acetylation/deacetylation of different Lys residues can have different biological effects. For example, histone H4 Lys5, 8, and 12 acetylation is involved in nucleosome assembly, heterochromation formation, and DNA repair (52).

Proteins other than histones can also be modified by Lys acetylation. Many transcription factors, cytoskeleton proteins, metabolic enzymes, and signaling proteins are acetylated (50).
Post-Translational Modifications to Regulate Protein Function

Figure 11  (a) Lys acetylation catalyzed by acetyltransferases; (b) mechanism of Zn-dependent HDACs-catalyzed deacetylation; (c) mechanism of sirtuins-catalyzed deacetylation.

Transcription factors are known to be substrates of HATs, whereas the enzymes responsible for the acetylation of non-nuclear proteins in many cases are not well known (55). The number of proteins that are regulated by acetylation will continue to increase as methods to detect protein acetylation improve. Acetylation of nonnuclear proteins can change protein–protein interaction, regulate enzymatic activity, and increase protein stability by suppressing ubiquitinylation (55).

Lys acetylation can be reversed by the action of deacetylases. Many deacetylases are Zn-dependent enzymes that use Zn\(^{2+}\) in the active site to activate water molecules to hydrolyze the amide bond (Fig. 11) (56). Recently, another type of deacetylases that are nicotinamide adenine dinucleotide (NAD)-dependent, also known as sirtuins, have been identified (57, 58). Their unique ability to couple NAD degradation to Lys deacetylation (Fig. 11) suggests that this type of enzyme can sense the metabolic state (for example, NAD concentration) of the cell and use that information to regulate the acetylation state and thus the function of the substrate proteins. In addition to Lys side chain acetylation, protein N-terminal can also be acetylated (59). In eukaryotic cells, the first residue Met in most proteins is cleaved by N-terminal methionine peptidase. The newly released N-terminal amino group is then acetylated. This modification can happen co-translationally before the mature peptide chain is released from the ribosome. The function of this modification in most cases is still not understood, although deletion of the genes involved in this modification has clear phenotypes (59).

Because of the involvement of protein Lys acetylation in regulation of transcription, protein–protein interaction, enzymatic activity, and protein stability, the deregulation of protein acetylation has been associated with many diseases, such as cancer and neurodegeneration (60, 61). Frequently, mutations in
histone acetyltransferases are found in cancer (60). Chromosomal abnormalities that generate fusions of acetyltransferases are known to lead to acute myeloid leukemia. These abnormalities include the fusions of M-02 (monocytic leukemia zinc finger protein) with CBF (CREB binding protein) or p300, and fusion of M-02 with the transcription factor TIF2 (transcription intermediary factor 2) (61). M-02, CBF, p300, and TIF2 all contain histone acetyltransferase domains. Presumably, the generation of these aberrant fusion proteins disrupts normal gene transcription, which leads to leukemia. Depletion of histone deacetylases is also suggested to be associated with cancer (61).

A histone deacetylase inhibitor, SAHA (Vorinostat, Merck & Co., Inc., Whitehouse Station, N.J.), was approved by Food and Drug Administration recently for treatment of cutaneous T-cell lymphoma (62).

**Methylation**

Although methylation can happen to several different residues (3, 63), most attention has been given to protein Lys/Arg methylation because the methylation of Lys/Arg in histones controls gene transcription. For Lys and Arg methylation, multiple methyl groups can be added to the same Lys or Arg residue (Fig. 12). The methyl group comes from S-adenosyl methionine (SAM), which is a versatile small molecule that is used in many enzymatic transformations (64). Almost all Lys methyltransferases belong to the SET (suppressor of variegation-Enhancer of zeste-Trithorax) family of methyltransferases, whereas the protein Arg methyltransferases belong to a different class (65–67). Both histone Lys/Arg methylation and acetylation are associated with transcription regulation. In contrast to histone acetylation, which usually correlates with transcription activation, histone methylation can lead either to transcription activation or to suppression (17, 68). The effect of histone methylation, which is based on current understanding, is mediated by proteins that are recruited by methylated Lys or Arg residues. Tudor domains and chromodomains are known to recognize methylated Lys/Arg residues via both charge interaction and cation interaction (69–73). The methylated Lys/Arg residue is more hydrophobic and sterically bulkier than free Lys/Arg, and it can be differentiated by the domains that recognize methylated Lys/Arg residues (69, 74). Sequences that surround the methylate Lys residues are also read by the chromo domains and Tudor domains (69–71).

This finding explains why different Lys residues could recruit different proteins on methylation and thus have different biological effects. For example, H3K4 methylation activates transcription by recruiting chromodomains and helicase-DNA-binding protein 1 (CHD1) specifically in yeast whereas H3K9 methylation represses transcription by recruiting heterochromatin protein 1 (HP1) (75–77). Nonhistone proteins are known to be methylated on Lys residues, which include transcription factors, such as p53 (78–80), TATA box-binding protein-associated factor (TBF1) (81), and translation factors (63). The p53 protein is methylated by different methyltransferases (Seth, 78; Smyd2 (79), and Setd8 (80)) on different Lys residues (Lys372, 370, and 382, respectively). These different methylation events either activate or repress p53 activity. Arg methylation has been found frequently in nonhistone proteins. For example, PRMT1 has been reported to methylate the transcription factor STAT1 (signal transducers and activators of transcription) (82), PRMT4 and CARM1 (coactivator-associated arginine methyltransferase-1) can methylate CBP/p300 (83), and heterogenous nuclear ribonucleoproteins (hnRNPs) and small nuclear ribonucleoproteins (snRNPs) that are involved in pre-mRNA splicing are also Arg methylated (67). The biological functions of these Lys/Arg methylations in most cases can also be explained by the effect of methylation to block or create interaction with other proteins or nucleic acids.

Compared with acetylation, methylation is more stable. For this reason, it was thought that methylation could be a permanent epigenetic mark. The recent discovery of two types of Lys demethylases suggests that methylation is also a reversible PTM. The first Lys demethylase discovered is LSD1 (lysine-specific demethylase 1), which is a FAD (flavin adenine dinucleotide)-dependent enzyme similar to amine oxidases (Fig. 12) (84). It is believed that LSD1 uses two-electron oxidation mechanism and thus cannot demethylate tri-methylated Lys residues (85). The second type of Lys demethylase, which contains the jmjC (jumonji domain-containing) domain, is a nonheme Fe(II)-dependent enzyme that is capable of doing one-electron oxidation, and thus, it can demethylate trimethylated Lys residues (86). The effect of Arg methylation was proposed to be reversed by protein Arg deminase 4 (PAD4), which generate citrulline via demethylamination (87, 88). However, later studies indicate that PAD4 as well as other PAD enzymes do not catalyze demethylamination with appreciable rates in vitro (88–91). A recent report showed that Arg methylation can be truly reversed by jmjC domain containing demethylases, which suggests that PADs are probably not required for Arg demethylation (92). Thus, both Lys and Arg methylation are reversible modifications.

Similar to Lys acetylation, abnormality in Lys methylation has been considered a contributing factor to cancer (93, 94). Decrease in H3 Lys9 and H4 Lys20 methylation is found in cancer cells. Both H3 Lys9 and H4 Lys20 methylation are associated with heterochromatin formation. Presumably, the decrease in the methylation leads to defects in heterochromatin formation, which in turn lead to chromosomal instability and tumor formation (93). Histone methyltransferase fusion proteins generated from chromosomal translocation are found frequently in leukemia and are thought to contribute to the development of leukemia. For example, the H3 Lys79 methyltransferase HDMT1 (human DOT1-like protein) fusion found in mixed lineage leukemia is sufficient to cause leukemic transformation (95). The close association of methylation and cancer suggests that protein methyltransferases and demethylases can be potential therapeutic targets.

**Glycosylation**

In eukaryotic cells, glycosylation has many membrane and secreted proteins (i.e., proteins that transit through the ER and the Golgi secretory pathway). Glycosylation can occur either on Asn residues (N-glycosylation, Fig. 13), Ser/Thr and post-translationally hydroxylated Lys and Pro residues (O-glycosylation, Fig. 14), or Trp residues (C-glycosylation, Fig. 14). N-glycosylation is a complicated process and involves...
Post-Translational Modifications to Regulate Protein Function

(a) Lys/Arg N-Methylation

(b) LSD1 demethylation (FAD-dependent): two-electron oxidation mechanism

(c) JHDM demethylation (Fe-dependent): one-electron oxidation mechanism

Figure 12  (a) Lys/Arg N-methylation; (b) mechanism of FAD-dependent LSD1-catalyzed Lys demethylation; (c) mechanism of Fe-dependent JHDM (JmjC domain-containing histone demethylase) -catalyzed demethylation.

three stages: 1) the formation of donor substrate with 14 sugar units (Glc3Man9GlcNAc2-PP-dolichol), which occurs in both the cytosolic and the luminal faces of ER (96); 2) the transfer of the tetradecasaccharyl group to the Asn residues found in the consensus sequence Asn-X-Ser/Thr, which occurs in the ER (97); and 3) the hydrolytic removal of the terminal sugar residues on the tetradecasaccharide, the addition of more sugar units (Fig. 13) (98), and the sulfation and phosphorylation of the carbohydrate moieties in the ER and Golgi (99).
Post-Translational Modifications to Regulate Protein Function

Figure 13  Protein N-glycosylation. (1) The formation of the donor substrate with 14 sugar units (Glc3Man9GlcNAc2-PP-dolichol), (2) the reaction scheme that shows the transfer of the tetradecasaccharyl group to the Asn residues found in the consensus sequence Asn-X-Ser/Thr in proteins; (3) hydrolytic removal of the terminal sugar residues on the tetradecasaccharide and addition of more sugar units in the ER and Golgi. OSTase, oligosaccharyltransferase.

Trimming steps can generate different sets of N-linked carbohydrates, such as the high-mannose type glycans, the complex type glycans, and the hybrid type glycans (Fig. 13) (99). Each stage is achieved by the function of multiple proteins. For example, up to nine proteins are required for the transfer of the tetradecasaccharyl group in yeast (100).

Different from N-glycosylation, O-glycosylation starts with the addition of a single sugar residue, which can be followed by the addition of more sugars (101). Similar to N-glycosylation, most O-glycosylation also occurs to proteins that transit through ER and Golgi. However, the addition of a single GlcNAc residue to Ser/Thr is a type of O-glycosylation that occurs to cytosolic proteins (102). This cytosolic O-glycosylation has drawn much attention recently because it can regulate the activity of the substrate proteins, especially because it can compete with protein phosphorylation for the same Ser/Thr on substrate proteins (103).

C-glycosylation is the addition of a single mannosyl group to the indole C-2 position of Trp residues of membrane and secreted proteins (104). The Trp residue that is C-mannosylated reside in a consensus Trp-X-X-Trp sequence, and the first Trp is C-mannosylated. About a dozen proteins in humans are C-mannosylated. The enzyme that catalyzes the modification has not been cloned yet, and currently, the function of this modification is not clear.

The large number of enzymes involved in protein glycosylation and the fact that this complicated N-glycosylation pathway is conserved throughout eukaryotic species suggest that glycosylation has important functions. Deficiency in protein glycosylation causes several diseases in humans, such as lysosomal storage diseases (105), congenital disorders of glycosylation, and leukocytes adhesion deficiency II (106). In addition, changes in glycosylation patterns are associated with cancer and inflammation (107). Protein glycosylation can serve several different biological purposes. One purpose is to help proteins that transit through the secretory pathway to fold correctly. Particularly, the removal of the glucose residue by glucosidase II and the reglucosylation in the ER have been well known to help secreted proteins to fold and make sure only correctly folded proteins are secreted (Fig. 13) (108).
Post-Translational Modifications to Regulate Protein Function

O-fucosyltransferase I that modifies Notch protein was reported to have chaperon activity that helps Notch folding and secretion, and this chaperon activity is independent of its catalytic activity (109). Glycosylation is also important for sorting secreted proteins. For example, the phosphorylation of Man on N-glycan (Fig. 16) creates a recognition signal for sorting lysosomal proteins to lysosome. Glycosylation is also believed to increase the protein stability, as has been shown for erythropoietin mentioned earlier. Glycosylation is also proposed to affect ligand receptor interaction and thus regulates cell–cell signaling. However, a detailed molecular understanding about the effect of glycosylation on ligand receptor interaction is hard to obtain in most cases. In two well-studied cases, human CD2 (cellular differentiation marker 2) and IgG (immunoglobulin G), N-glycosylation is found to affect the interaction with their ligands or receptors. Structural data show that the carbohydrate portion does not contact the binding partner directly. Instead, glycosylation affects the binding by changing the conformation of the glycosylated proteins (110–112).

Ubiquitylation
Ubiquitin is an abundant small protein (76 amino acids) found in all eukaryotes. It can be conjugated to many proteins covalently and regulates important biological processes. The addition of ubiquitin to substrate proteins goes through an E1–E2–E3 enzymatic cascade (Fig. 17) (113). E1, which is also called ubiquitin-activating protein (UAP), uses ATP to adenylate the C-terminal Gly of ubiquitin and then captures the activated ubiquitin with a Cys residue in the active site. Most eukaryotic species only have one E1 enzyme responsible for activating all the ubiquitin molecules needed. The ubiquitin-E1 conjugate then is recognized by several dozens of E2 enzymes, which capture ubiquitin from E1 via a transthiolation reaction. The ubiquitin-conjugated E2 enzymes are then recognized by many different E3 enzymes, which recruit the substrate proteins and transfer ubiquitin from E2 to Lys residues of the substrate proteins, either directly or indirectly (Fig. 17). Two major families of E3 enzymes exist: the RING (really interesting new gene) E3s and HECT (homologous to E6AP C terminus) E3s. The Pfam database lists more than 400 RING proteins and 70 HECT proteins. Many E3s form complexes with other proteins. One well-understood E3 complex is the SKF (Skp1-Cullin-F Box)
RING E3, for which a crystal structure was reported (114). In humans, multiple Cullins and multiple F Box proteins exist (115). Considering the different combinations, the number of possible E3 complexes can be much more than the number of E3 enzymes (3). E3s decide which substrate proteins get ubiquitylated, thus the large number of E3s and E3 complexes reflects the diverse substrate proteins that must be recognized.

Ubiquitin itself has 7 Lys residues (Lys6, 11, 27, 29, 33, 48, and 63) that can be used for ubiquitin attachment, which lead to polyubiquitylation of substrate proteins. Polyubiquitin chain assembled via different Lys residues have different biological functions (116), as will be explained later. Which Lys residue is used in the polyubiquitin chain is controlled by the specific E3 involved. E3 presumably also controls the length of the polyubiquitin chain, although the detailed chain assembly mechanism is still not clear (117). Ubiquitylation can be reversed by the action of ubiquitin-specific proteases (UBPs). About 60 UBPs exist in the human genome, which presumably recognize different types of ubiquitin modifications at various cellular locations (118).
The biological function of ubiquitylation was recognized originally as targeting proteins to the proteasome for degradation. The importance of this function can be illustrated by many examples. In cell division, progression through the cell cycle is driven by cell division kinases, the activities of which are controlled by a group of proteins called cyclins. Different cyclins function only at certain stages of the cell cycle. Then, they must be degraded, which requires polyubiquitylation by specific E3 enzymes (119). A loss in the ubiquitylation and degradation of cyclins is associated with cancer. Misfolded proteins must be degraded by the ubiquitin and proteasome system. A aggregation of misfolded proteins is known to cause neurodegeneration, such as Parkinson’s disease (116). Ubiquitylation and proteasome degradation of proteins are also important for other biological processes, such as hypoxia and circadian clock.

Ubiquitylation is required for the degradation of hypoxia inducible factor (HIF) on hydroxylation at high oxygen levels (120). Maintaining the circadian clock requires the ubiquitylation and degradation of proteins that inhibit the CLOCK (a protein named from circadian locomotor output cycles kaput gene) transcription factor (121).

It is becoming clear that the biological function of ubiquitylation is not limited to proteasome degradation. Other functions have been discovered, such as promoting membrane protein endocytosis, targeting membrane protein to lysosome for degradation, and regulating cytoplasm/nuclear shuttling (116, 122). It is now generally believed that polyubiquitylation via Lys48 of ubiquitin is a signal for proteasomal degradation, and this action requires minimally 4 ubiquitin units in the chain (123). In contrast, monoubiquitylation, multiple
monoubiquitylation on different Lys residues of substrate proteins, and polyubiquitylation via Lys 63 of ubiquitin typically signal proteasome-independent pathways (116). How can so many different functions be achieved? The diverse sets of ubiquitin binding domains (UBDs) provide the molecular explanation to this question (19). Presumably, different UBDs recognize different types of ubiquitin modifications (monoubiquitylation vs. polyubiquitylation, and Lys48-linked vs. Lys63-linked polyubiquitylation, for example), and thus they mediate different functional consequences of ubiquitylation. UBD on yeast proteins Rad23, Rpd10, and Dsk2 recognize the Lys48-linked polyubiquitin chain and deliver the modified substrate proteins to the 26 S proteasome (124). The UBD on the vacuolar proteins recognize monoubiquitylation or Lys63-linked polyubiquitin chain on membrane proteins, which mediate their sorting into lysosome or vacuole. Binding of the Lys63-linked polyubiquitin chain on inhibitor of NF-κB kinase (IKK) by other proteins has been proposed to activate IKK and thus turn on NF-kB signaling (116). The recognition of ubiquitin by UBDs can also explain some “unusual” functions of protein ubiquitylation. For example, Lys48-linked polyubiquitylation of a yeast transcription factor Met4p does not signal for proteasome degradation, but instead it inactivates the transcription factor. It inactivates the proteasomal pathway (125).

In generalization of the function of ubiquitylation, we can say that ubiquitin is an “information-rich protein tag” that can be read by different proteins that contain UBD domains (3), and the exact consequence of ubiquitylation is determined by how the tag is recognized. Besides ubiquitin, eukaryotic cells also have about a dozen known ubiquitin-like protein tags, with SUMO being the best studied one. In addition, many proteins have built-in ubiquitin-like domains. The logic that underlies the biological functions of these ubiquitin-like proteins/domains will likely be the same as what is learned from ubiquitin (3).

**Proteolysis**

Hydrolytic cleavage of proteins by proteases is an irreversible PTM. The large number (more than 500) of proteases in the human genome indicates that proteolysis occurs often. Proteases can be classified into four types based on catalytic mechanisms (Fig. 18): Ser/Thr proteases, Cys proteases, Asp proteases, and metalloproteases.

At first glance, proteolysis may seem to be an uncontrolled destruction process like the digestion of food proteins in the gut. In fact, proteolysis in cells is under tight regulation. Even proteases secreted to the digestive tract must be controlled to avoid self-destruction. Typically, proteases are made in the inactive forms (zymogens) that can be activated by proteolysis. Inside eukaryotic cells, two major locations exist for proteolytic degradation of unwanted proteins: the 26 S proteosome and the lysosome (126, 127). Access to the two degradation organelles is controlled tightly. The lysosome is an acidic membrane organelle that contains many proteases and is responsible for degradation of endocytosed membrane proteins, such as activated receptor tyrosine kinases and G protein-coupled receptors that are ubiquitylated and sorted to the lysosome (described in...
Figure 18

Catalytic mechanisms of different proteases:

- Ser/Thr proteases
- Cys proteases
- Asp proteases
- Metalloproteases (Zn-dependent)
The eukaryotic 26 S proteasome. Subunit compositions of the 19 S regulatory particle of *Saccharomyces cerevisiae* is shown on the left. The α and β rings of the 20 S proteasome, each of which consists of seven different subunits, are included to indicate how the base 19 S complex is linked to the core 20 S protease complex. The crystal structure of the 20 S degradation chamber is shown in both side and top views (figure made using PDB 1RYP).

In autophagy, the lysosome is responsible for degrading cellular organelles and some cytosolic protein complexes (126). The 26 S proteasome (Fig. 19) has a 20 S degradation chamber that consists of four rings αβαβαβαβαβ. In eukaryotes, each α ring has seven different α subunits, and each β ring has seven different β subunits. Three β subunits are catalytically active Thr proteases that are responsible for the degradation of subproteins. By forming this chamber, the active sites of the proteases are buried inside the chamber to avoid proteolysis of proteins that should not be digested. Access to the degradation chamber is controlled by the 19 S regulatory complex that caps both ends of the degradation chamber. The regulatory complex contains subunits that recognize polyubiquitylated substrates, subunits that recycle the ubiquitin tag, and subunits that use ATP hydrolysis to unfold and translocate the protein into the degradation chamber. Degradation of the unwanted proteins by the 26 S proteasome or lysosome is a timely fashion is very important. For example, cyclins that activate cell division kinases have to be polyubiquitylated and degraded by the proteasome at specific times to drive cell cycle progression (119). Degradation of activated membrane receptors in the lysosome is important to avoid over stimulation (130, 131). Misfolded proteins must be degraded by the proteasome or lysosome (in autophagy). Failure to do so is thought to contribute to neurodegeneration disorders such as Parkinson’s disease and Alzheimer’s disease (126).

In addition to the “destructive” proteolysis processes in the proteasome and lysosome, many “constructive” proteolysis processes occur in cells. In both prokaryotes and eukaryotes, secreted proteins contain a signal peptide at the N-terminus that directs them to the secretory pathway. This signal peptide must be cleaved later by signal peptidases (typically serine proteases) so that the protein can transit further in the secretory pathway (132). Many secreted proteins, which include insulin, TGF-β1, transforming growth factor β1, nerve growth factor β1, albumin, Factor IX, insulin receptor, and Notch, also contain a propeptide that is cleaved by proprotein convertases in the Golgi (133). Selective proteolysis also occurs at the cell membrane in signal transduction processes. Notch protein, on binding to its ligand Delta/aggregated membrane proteins on neighboring cells), is cleaved by one of the ADAM (a disintegrase and metalloprotease) proteins at a site close to the transmembrane region. This cleavage activates Notch for regulated intramembrane proteolysis, which cuts within the membrane-spanning region of Notch and releases the intracellular domain of Notch from the cytoplasm membrane. Then, the intracellular domain translocates into the nucleus where it acts as a transcription factor to turn on genes required for development (Fig. 20) (134).

Regulated intramembrane proteolysis is catalyzed by the membrane protein complex called presenilin that contains Asp protease subunits. Presenilin is also responsible for cleavage of the amyloid-β precursor proteins in Alzheimer’s disease. This proteolysis-triggered proteolysis signaling occurs often. Similar signaling pathways are present also in bacteria. For example, the release of the transcription factor σE is achieved via the sequential cleavage of the membrane protein RseA by DegS (a Ser protease) and YaeL (a Zn protease) (135).

Similar to the MAP kinase cascades for protein phosphorylation, protease cascades exist, in which downstream proteases are activated by the action of upstream proteases (3). One of the most famous cascades is the caspase cascade that leads to apoptosis (Fig. 22) (111, 136). Caspases are Cys proteases that cleave the amide bond specifically after an Asp residue. Two types of caspases exist; initiator caspases (Caspase 2, 8, 9, 10) and effector caspases (Caspase 3, 6, 7). Both initiator and effector caspases are activated by upstream caspases and cleave other downstream caspases to amplify the apoptotic response.
caspases are produced in zymogen forms. Initiator caspases use their N-terminal DED (death effector domain) and CARD (caspase recruitment domain) domains to interact with other proteins to receive apoptosis signals. The signals cause the dimerization of the initiator caspases and activate them so that they can cleave themselves and the effector caspases after specific Asp residues. Cleavage by the initiator caspases activates the effector caspases, which then cleave their substrate proteins to carry out cell apoptosis. The substrate proteins of effector caspases include the inhibitor of caspases-activated DNAse (deoxyribonuclease), Bcl2 (named from B-cell lymphoma 2, an antiapoptotic protein), and PARP-1 (poly(ADP-ribose) polymerase-1, an enzyme catalyzing protein poly(ADP-ribosyl)ation and required for DNA repair). Cleavage of the inhibitor of DNAse by effector caspases activates its catalytic activity, resulting in the fragmentation of chromosomal DNA, which is a hallmark of apoptosis. The caspases cascade and apoptosis is very important for the development and homeostasis of metazoans. Decreased ability of cells to undergo apoptosis will lead to cancer, whereas too much apoptosis can lead to autoimmune diseases (137).

Identifying new pathways regulated by known PTM and discovering new PTM

The brief description above on a few major PTM demonstrates clearly that PTM can regulate many important biological processes. So far, a fairly good understanding of many aspects of PTM has been obtained. What remaining challenges must be addressed?

One direction is to figure out the molecular details of many of the biological processes that are regulated by PTM. Structural biology and biochemistry is needed to answer questions like what structural changes are induced by a particular PTM and how the structure changes lead to changes in activity or recognition by binding partners. Much progress has been made in this direction but still more remains to be figured out. For example, in protein ubiquitylation, no structural details about E1 exist, it is not clear how the polyubiquitin chain is made (117), and it is not clear how specificities of different ubiquitin binding domains are achieved (19).

Another direction is to identify the proteomes that are modified by a specific PTM. Advancement in protein identification by mass spectrometry (MS) has greatly facilitated studies in this direction and many efforts have been invested. Generally, an affinity purification method is used to enrich proteins that are modified by a specific PTM, and then these proteins are identified by MS. For example, phosphotyrosine-specific antibodies have been used to enrich proteins that are modified on Tyr residues, and metal affinity columns have been used to isolate all phosphopeptides (138). These isolated phosphopeptides can then be identified by MS. A His6 tag has been fused to the N-terminus of ubiquitin and used to isolate ubiquitylated proteins that are then identified by MS (139). GlcNAc with an azide group attached has been used to label proteins that are O-GlcNAc modified, and then a biotin tag is conjugated to the modified protein via Staudinger ligation. O-GlcNAc modified proteins can be pulled out using streptavidin beads and identified using MS. Using this method, close to 200 O-GlcNAc modified proteins were identified (140). A clever method to detect protein S-acylation has been reported recently (141). These proteomic studies have provided much information. However, to understand the function of a PTM in cell physiology completely, it is desirable to know which enzyme is responsible for the modification of a particular substrate protein. With the availability of bioinformatics tools and completed genome sequences, it is now relatively straightforward to identify all the enzymes in a genome that share similar biochemical function. For example, we now know that the human genome
Post-Translational Modifications to Regulate Protein Function

contains more than 500 protein kinases, more than 500 pro-
teases, and ∼400 ubiquitin E3s. But without knowing what
substrate proteins they modify, it will be very difficult (if not im-
possible) to understand their biological functions on a molecular
level. Currently, no efficient and reliable method exists yet to
identify the substrate proteins for an enzyme. A straightforward
method is to make a library of short peptides and try to identify
consensus sequences that are recognized by an enzyme (142,
143). The disadvantage is that the structure of a short peptide
may be different from the structure of the same sequence present
in a folded protein. Thus, the reliability of this method must be
validated by other methods. Shokat and coworkers (144) have
used a clever approach to identify kinase substrates (Fig. 22).
This approach uses a bulky ATP analog that can be used only
by a kinase mutant as a cosubstrate. By incubating 32P-labeled
ATP analog and the kinase mutant with cell extract, the substrate
proteins of the specific kinase can be labeled. Identification of
the substrate proteins may be difficult though because the ra-
diolabeled substrate proteins cannot be enriched/purified easily
for identification by MS. It is not clear whether this method can
be applied easily to other PTM enzymes.

Parallel to the efforts of identifying substrate proteins for a
particular enzyme, the activity-based small molecule probes pio-
nereed by Cravatt and coworkers can facilitate the identification
of the biological functions of an enzyme that catalyzes protein
post-translational modifications (145). The major advantage of
this type of probes is that potentially they can detect enzymes
that are in the active states, and thus can provide snapshots of
enzymes that are in the active states at different development
stages or different types of cells. Among enzymes that catalyze
PTM, so far probes have been developed for studying proteases
(145, 146), kinases (147), pTyr phosphatases (148), and protein
Arg deiminases (149).

Perhaps a more challenging question is how can we discover
new PTM reactions. In principle, there are analytic tools that
can be used to research this topic. One such tool is top-down
FT-MS, which determines the molecular weight of the whole
protein with high accuracy. By comparing the obtained tan-
dem MS (MS/MS) result with the expected MS/MS result,
post-translational modifications can be identified (150). Crys-
tallography can also discover new PTM, if a protein expressed
in the proper host can be crystallized. Some rare modifications
or protein side chains were discovered this way (151). How-
ever, the success of using these methods would require that
Post-Translational Modifications to Regulate Protein Function

Shokat’s (144) “bump and hole” method to identify substrates for kinases.

A significant portion of the protein population is modified and the modification is stable. This condition cannot be met by all PTM. Thus, discovering new PTM poses a great challenge to chemical biologists. Undoubtedly, new PTM reactions are waiting to be discovered and the identification of these new PTM, together with the identification of new pathways that are regulated by known PTM, will advance our understanding about the molecular logic of living systems.

References

Post-Translational Modifications to Regulate Protein Function

Post-Translational Modifications to Regulate Protein Function


Molecular Chaperones

R. John Ellis, University of Warwick, Coventry, United Kingdom

doi: 10.1002/9780470048672.wecb068

The generic tendency of proteins to aggregate into nonfunctional, and sometimes cytotoxic, structures poses a universal problem for all types of cells. This problem is exacerbated by the high total concentrations of macromolecules found within most intracellular compartments, but it is solved by the actions of certain proteins that function as molecular chaperones. Different chaperones act by distinct mechanisms on both the folding of polypeptide chains and their subsequent assembly into oligomeric structures. Many chaperones, but not all, are also stress (or heat shock) proteins because the need for a chaperone function increases under stress conditions that cause proteins to unfold.

Proteins have evolved over billions of years to function inside highly complex, intracellular environments, but their properties commonly are studied after purification and after exposure to much simpler and very different conditions in the test tube. Christian Anfinsen received the Nobel Prize for chemistry in 1972 for discovering that some purified denatured proteins will refold into their biologically active conformations when the concentration of the denaturing agent is lowered (1). This classic type of refolding experiment has been repeated many times, and it is clear that most proteins are capable of refolding correctly in dilute media in the absence of either other macromolecules or an added energy source. The same conclusion applies to the assembly of large multisubunit structures such as viruses and ribosomes. These observations were codified in the statement that proteins are capable of spontaneous self-assembly, the term "assembly" is used here to describe both the folding of individual polypeptide chains and their association into oligomers. The principle of protein self-assembly states that all the information for a protein chain to reach its correct conformation is encoded in its aminoacyl sequence and is an important corollary of the central dogma of molecular biology. If self-assembly did not occur, molecular biology would be in deep trouble because it would have to be postulated that protein folding and assembly requires direction by an external agent. Generally, it was assumed that proteins also fold and assemble spontaneously inside the cell, but observations made in the late 1970s and early 1980s challenged this assumption. Some proteins were found to bind to preexisting native proteins before folding or after folding but before assembling. Such binding prevents the proteins from aggregating into nonfunctional structures, which is a problem aggravated by the macromolecular crowding found in most intracellular compartments (2, 3). The term "molecular chaperone" (often abbreviated to "chaperone") was proposed to describe these helper proteins (4, 5), and the concept advanced that all cells require a chaperone function to prevent some proteins from misbehaving (6). Thus the original view that proteins fold and assemble inside the cell by a process of spontaneous self-assembly has been replaced by the concept that many proteins use an assisted self-assembly process. Note that the concept of protein self-assembly itself is unaffected by this change of paradigm—proteins contain all the information required for their correct folding and assembly. Their problem is to avoid aggregation.

Biologic Background

Protein aggregation is defined as the association of two or more polypeptide chains to form nonfunctional structures. Some aggregates are cytotoxic and cause neurodegenerative disease. Protein aggregation is at least a second-order process, so it is very sensitive to the concentration of interacting chains. Aggregation is also a highly specific process, so that only identical or very similar chains aggregate with one another. Because the free energy difference between folded and unfolded states is small, most folded proteins are stable only marginally and are easily unfolded into aggregation-prone states. All cells require a chaperone function to combat aggregation because two universal features of the intracellular environment, where proteins fold and assemble, aggravate both of these properties of aggregation. These features are the synthesis of proteins on polyribosomes and macromolecular crowding.

Polyribosomes

The linear DNA base sequence information encoded in genes is copied into a linear sequence of ribonucleotides in messenger RNA (mRNA) by the process of transcription. Ribosomes bind one at a time near to the 5' end of each mRNA molecule and translate the ribonucleotide base sequence into an aminoacyl sequence according to the genetic code. Each mRNA is bound to more than one ribosome at a time, forming a polyribosome, often abbreviated to polysome. Typically the rate of
polypeptide chain synthesis is between 5 and 20 amino acid residues added per second, but the rate of protein folding is much faster. Thus, incomplete chains may fold into nonfunctional conformations before the entire chain is made. The fact that polypeptide chains are made by polynucleotides ensures that partly folded identical chains are within touching distance when they are synthesized, which results in the possibility that chains may aggregate with one another. The advantage of simultaneously making several polypeptide chains from one molecule of mRNA has been earned at the risk of both misfolding and aggregation. Some chaperones reduce this risk by binding to incomplete and newly synthesized polypeptide chains before they complete their folding.

Macromolecular crowding

This term is used to highlight that the total concentration of macromolecules inside cells is very high, which is in the range 80–400 g/L. Such high total concentrations occur in all intracellular compartments where proteins fold and assemble. Generally, it is not appreciated that such a degree of crowding stimulates macromolecular association reactions, by levels that can be two to three orders-of-magnitude greater than those in uncrowded media (2, 3). Such association reactions include protein aggregation, and during some, but not all, protein refolding experiments, a fraction of the chains aggregate with one another, even though such experiments are almost always done in uncrowded media. Refolding experiments show the concentration and temperature-dependence of protein aggregation, so protein chemists solve this problem by reducing the concentration of chains and lowering the temperature. Cells cannot do this but employ molecular chaperones to combat the enhanced tendency to aggregate, which is caused by macromolecular crowding.

Origin of the molecular chaperone concept

The term “molecular chaperone” first appeared in 1978 to describe the properties of a nuclear protein called nucleoplasmin found in amphibian eggs (4). This protein assists the assembly of nucleosomes from histones and DNA that occurs when these eggs divide rapidly after fertilization. Nucleosomes are held together by the electrostatic bonds that are disrupted by high concentrations of salt. If the isolated histones and DNA are then incubated together at physiologic concentrations of salt, a dramatic failure of self-assembly exists—an insoluble aggregate forms instead of nucleosomes. Eggs solve this problem by breaking the concentration of chains and lowering the temperature. Cells cannot do this but employ molecular chaperones to combat the enhanced tendency to aggregate, which is caused by macromolecular crowding.

Definitions

The molecular chaperone function currently is defined as the ability of all cells to prevent and/or reverse incorrect interactions that may result when potentiallyinteractive surfaces are exposed to the environment (5). These surfaces occur on growing and newly released polypeptide chains, on mature proteins unfolded by environmental stresses, and on folded proteins in near-native and native conformations. The same concept applies to other macromolecules that can undergo incorrect interactions, especially RNA. Incorrect interactions are defined as those that result in products that do not carry out the functions for which they have been selected in evolution.

Molecular chaperones currently are defined as a large and diverse group of proteins that share the functional property of assisting the noncovalent folding/unfolding and the association/dissociation of other macromolecular structures but are not components of these structures when these are performing their normal biologic functions (7). This definition differs from the original one (6) in that includes unfolding and disassembly processes. This change was necessary because of the discovery that some chaperones are required for processes such as the remodeling of chromatin during fertilization, the disassembly of clathrin coats, and the dissociation of insoluble aggregates. Some molecular chaperones are also stress or heat shock proteins (hsps), because the requirement for chaperone function increases under conditions such as heat shock and other environmental stresses that cause some proteins to unfold and aggregate.

The above definition of molecular chaperone is entirely functional and contains no constraints on the mechanisms by which different chaperones may act. The term “noncovalent” is used to exclude those proteins that carry out posttranslational covalent modifications. Protein disulfide isomerase may seem to be an exception, but it is both a covalent modification enzyme and a molecular chaperone. It is helpful to think of a molecular chaperone as a function rather than as a molecule. Thus, no reason exists why a chaperone function should not be a property of the same molecule that has other functions. Other examples include peptidyl-prolyl isomerase, which possesses both enzymatic and chaperone activities in different regions of the molecule, and the alpha-crystallins, which combine two essential functions in the same molecule in the lens of the eye—contributing to the transparency and the refractive index required for vision as well as the ability to promote protein folding.
as to the chaperone function that combats the loss of transparency as the protein chains aggregate with increasing age. The proteasome particle has a chaperone-like activity involved in unfolding proteins prior to their proteolysis. A recent proposal is that some chaperones also possess extracellular cell-cell signaling functions, based on many experiments that show that some coevolved applied chaperones elicit responses from cultured cells (10). It remains an open question as to how many other effects chaperones may have.

Classes of Chaperone

It is convenient to divide chaperones into two broad classes: those that assist the folding of newly synthesized or stress-denatured proteins and those that assist the assembly of oligomeric structures. Some of these chaperones are highly specific for their protein substrates, whereas others are not. Many proteins have been termed chaperones, usually on the basis of their in vitro properties, but no comprehensive list is available.

Chaperones involved in protein folding—small chaperones

Some chaperones assist the folding of both nascent chains bound to ribosomes, newly synthesized chains released from ribosomes (i.e., in both cotranslational and posttranslational modes), as well as mature proteins unfolded by environmental stresses and some membrane proteins (11–12–13). The chaperones working in these cotranslational and posttranslational modes are distinct, and they can be termed small and large chaperones, respectively, because this is a case where size is important for function. Small chaperones are less than 200 kDa in size and include trigger factor, nascent chain-associated complex, prefoldin, the hsp70 and hsp40 families, and their associated cochaperones. Cochaperones are defined as proteins that bind to chaperones to modulate their activity; they may or may not also be chaperones in their own right. Large chaperones are more than 800 kDa in size and include DnaK, which allows the latter to bind another molecule of ATP, and GrpE, a nucleotide exchange factor that is a cochaperone but not a chaperone. DnaK binds to DnaJ through its J domain and increases the rate of ATP hydrolysis, which facilitates peptide binding. DnaJ, like all the hsp40 proteins, acts as a chaperone in its own right because it also binds to hydrophobic peptides. Thus, DnaK and DnaJ cooperate to bind each other to nascent chains; all hsp70 chaperones are thought to cooperate with hsp40 chaperones. The role of GrpE is to stimulate release of ADP from DnaK, which allows the latter to bind another molecule of ATP and to release the peptide. In the eukaryotic cytosol, the role of GrpE is fulfilled by an unrelated cochaperone called Bag-1. Some Archaea lack hsp70 proteins, but it is speculated that their role in protein folding may be replaced by that of an unrelated chaperone called prefoldin.

Some chaperones are highly specific for their protein substrates, whereas others are not. Many proteins have been termed chaperones, usually on the basis of their in vitro properties, but no comprehensive list is available.

Molecular Chaperones

because it is associated with the ribosomal large subunit at the tunnel from which the chains emerge (14). A cell of Escherichia coli contains about 20,000 copies of this chaperone, which is enough to bind to all nascent chains. Trigger factor shows peptidyl-prolyl isomerase activity and contains a hydrophobic groove that binds transiently to regions of the nascent chain enriched in aromatic residues. It binds to nascent chains as short as 57 residues and dissociates in an ATP-independent manner after the chain is released from the ribosome; this binding does not require prolyl residues in the nascent chain. The isomerase activity may provide a means to keep nascent chains that contain prolyl residues in a flexible state. The eukaryotic cytosol lacks the trigger factor, but its function may be replaced by that of a heterodimeric complex of 33-kDa and 22-kDa subunits, termed the nascent chain-associated complex. Like the trigger factor, this complex binds transiently to short nascent chains, but unlike the trigger factor, it does not possess peptidyl-prolyl isomerase activity.

Cells lacking the trigger factor show no phenotype; but this is because its function can be replaced by that of the other major small chaperone, hsp70 (15). The hsp70 family has many 70-kDa proteins distributed between the cytoplasm of bacteria and some Archaea, the cytosol of Eukarya, and eukaryotic organelles such as the endoplasmic reticulum, mitochondria, and chloroplasts. Some members are also stress proteins. Unlike the trigger factor, most hsp70 members do not bind to ribosomes but do bind to short regions of hydrophobic residues exposed on nascent and newly synthesized chains. Such regions occur statistically about every 40 residues and are recognized by a peptide-binding cleft in hsp70.

Most information is available about the hsp70 member in E. coli, termed DnaK. Like all hsp70 chaperones, DnaK contains an ATPase site and occupation of this site by ATP promotes rapid but reversible peptide binding. ATP hydrolysis tightens the binding through conformational changes in DnaK. The cycling of ATP between these states is regulated by a 41-kD cochaperone—one of the hsp40 family, termed DnaJ in E. coli, and GrpE, a nucleotide exchange factor that is a cochaperone but not a chaperone. DnaJ binds to DnaK through its J domain and increases the rate of ATP hydrolysis, which facilitates peptide binding. DnaJ, like all the hsp40 proteins, acts as a chaperone in its own right because it also binds to hydrophobic peptides. Thus, DnaK and DnaJ cooperate to bind each other to nascent chains; all hsp70 chaperones are thought to cooperate with hsp40 chaperones. The role of GrpE is to stimulate release of ADP from DnaK, which allows the latter to bind another molecule of ATP and to release the peptide. In the eukaryotic cytosol, the role of GrpE is fulfilled by an unrelated cochaperone called Bag-1. Some Archaea lack hsp70 proteins, but it is speculated that their role in protein folding may be replaced by that of an unrelated chaperone called prefoldin.

Enough DnaK is in a cell of E. coli for one molecule to bind to each nascent chain. DnaK binds to longer chains than the trigger factor, and so it probably binds after the trigger factor. When the gene for the trigger factor is deleted, the fraction of nascent and newly synthesized chains that bind to DnaK increases from about 15% to about 40%. However, removal of the genes for both the trigger factor and the DnaK in the
Molecular Chaperones

**Table 1 Chaperones that assist protein folding**

<table>
<thead>
<tr>
<th>Chaperones that assist protein folding</th>
<th>Eukaryotes</th>
<th>Prokaryotes</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hsp100</strong></td>
<td>hsp104, 78</td>
<td>ClpA/B/X</td>
<td>Disassembly of oligomers and aggregates</td>
</tr>
<tr>
<td><strong>hsp90</strong></td>
<td>hsp82, hsp83, Grp94</td>
<td>HtpG</td>
<td>Regulation of assembly of steroid receptors and signal transduction proteins</td>
</tr>
<tr>
<td><strong>hsp70</strong></td>
<td>hsc70, Ssa1-4, Ssa2-2, Bip, Grp75</td>
<td>DnaK, Hsc66 Absent from many Archaea</td>
<td>Prevention of aggregation of unfolded protein chains</td>
</tr>
<tr>
<td><strong>Chaperonins</strong></td>
<td>hsp60, TRiC, CCT, TCP-1, Rubisco subunit binding protein</td>
<td>GroEL, GroES</td>
<td>Sequestering partly folded chains inside central cage to allow completion of folding in absence of other folding chains</td>
</tr>
<tr>
<td><strong>hsp40</strong></td>
<td>Ydj1, Sis1, Sec63p, auxilin, zuotin, Hdj2</td>
<td>DnaJ</td>
<td>Stimulation of ATPase activity of hsp70</td>
</tr>
<tr>
<td><strong>Prefoldin</strong></td>
<td>GimC</td>
<td>Absent from Bacteria Present in Archaea</td>
<td>Prevention of aggregation of unfolded protein chains</td>
</tr>
<tr>
<td><strong>Trigger factor</strong></td>
<td>Absent from Eukarya</td>
<td>Present</td>
<td>Binding to nascent chains as they emerge from ribosomes</td>
</tr>
<tr>
<td><strong>Calnexin, calreticulin</strong></td>
<td>Present</td>
<td>Absent from prokaryotes</td>
<td>Binding to partly folded glycoproteins; located in ER membrane and lumen, respectively</td>
</tr>
<tr>
<td><strong>Nascent chain-associated complex (NAC)</strong></td>
<td>Present</td>
<td>Absent from prokaryotes</td>
<td>Binding to nascent chains as they emerge from ribosomes</td>
</tr>
<tr>
<td><strong>PapD</strong></td>
<td>Absent from Eukarya</td>
<td>Present in some</td>
<td>Prevention of aggregation of subunits of pili</td>
</tr>
<tr>
<td><strong>Membrane chaperones</strong></td>
<td>Shl3p Gs2p Phodlp</td>
<td>Chp3p</td>
<td>Prevent aggregation of some integral polytopic membrane proteins</td>
</tr>
</tbody>
</table>

same cell causes the aggregation of many newly synthesized chains and is lethal to the cell (15). This observation suggests that the redundancy of important control systems is as good a design principle for cells as it is for passenger planes. An exception to this principle is the GroEL chaperone (see the discussion in the next section), whose gene is essential for the survival of E. coli. Small chaperones function essentially by reducing the time that potentially interactive surfaces on neighboring chains are exposed by cycling these chains on and off until they have folded; they do not seem to change the conformation of the chains. The other major class of chaperones involved in protein folding, however, functions by a much more sophisticated mechanism enabled by its large size.

**Large chaperones involved in protein folding—the chaperonins**

Most information is available for GroEL and GroES, the chaperonin and cochaperonin found in E. coli, but the general principles of their mechanism (Fig. 2) are suspected to apply also to the thermosome found in Archaea and to the TCP-1 complex (also called the TRiC or CCT complex) found in the eukaryotic cytosol. GroEL (800 kDa) consists of two heptameric rings of identical 57-kDa ATPase subunits stacked back to back, which contain a cage in each ring (17). The term “cage” is used because the walls surrounding each central cavity contain gaps, perhaps to allow water to enter and to exit. Each subunit contains three domains. The equatorial domain contains the nucleotide binding site and is connected by a flexible intermediate domain with the apical domain. The latter presents several hydrophobic side chains at the top of the ring oriented toward the cavity of the cage, which is an arrangement that permits either a partly folded polypeptide chain or a molecule of GroES to bind but prevents binding to another GroEL oligomer. GroES is a single heptameric ring of 10-kDa subunits that cycle on and off either end of the GroEL in a manner regulated by the ATPase activity of GroEL. A single ATPase unit is essential for the survival of E. coli. The reaction cycle starts with a GroEL–GroES complex containing ADP bound to the cis ring (Fig. 2, step 1).
Molecular Chaperones

Figure 1. Models for the chaperone-assisted folding of newly synthesized polypeptides in the cytosol. (a) Bacteria. TF, trigger factor; N, native protein. Most nascent chains probably interact with TF, and most small proteins (about 65–80% of total chain types) may fold rapidly during synthesis without additional chaperone assistance. Longer chains (10–20% of total chain types) interact subsequently with DnaK and DnaJ and fold after one or several cycles of ATP-dependent binding and release. About 10–15% of total chains fold within the chaperonin GroEL/GroES system. GroEL does not bind to nascent chains and is likely to receive its substrates after their release from DnaK. (b) Archaea. PFD, prefoldin; NAC, nascent-chain associated complex. Only some archaean species contain DnaK/DnaJ. The existence of a ribosome-bound NAC homologue and the binding of prefoldin to nascent chains have not been shown. (c) Eukarya. Like TF, NAC probably interacts with many nascent chains. Most smaller chains may fold without additional chaperone assistance. About 15–20% of chains reach their native states after assistance by hsp70 and hsp40, and a specific fraction of these are then transferred to hsp90. About 10% of chains are passed to the TRiC system in a reaction involving PFD. Reprinted from Reference (12) with permission from the American Association for the Advancement of Science.
rate of interactions that lead to compaction of the folding chain. Current estimates suggest that the fraction of newly synthesized polypeptide chains that bind in vivo to either hsp70 proteins or the chaperonins is in the range 10–20%. Whether most newly synthesized chains bind to other, as yet undiscovered, chaperones or fold unassisted because their sequences have evolved to avoid aggregation is unknown. Nor is it understood what determines that only a few defined polypeptides bind to GroEL in the intact cell. Some mycoplasmas lack genes for GroEL and GroES, perhaps because the limited number of proteins they contain do not include ones prone to aggregation.

Chaperones assisting oligomeric assembly

The nuclear compartment contains high concentrations of negatively charged nucleic acids bound to positively charged proteins that together form chromatin and various ribonucleoprotein particles. Besides nucleoplasm, several other chaperones assist the assembly and disassembly of these structures by preventing incorrect ionic interactions between them, whereas cytosolic assembly chaperones assist the assembly of ribosomes, proteasomes, and proteins such as hemoglobin (7, 20, 21). The continuing emphasis on the role of chaperones in protein folding may explain the relative dearth of information about assembly chaperones.

Common misconceptions

As with any new field, misconceptions abound. The literature contains statements of the form “chaperones fold proteins,” which suggests to the uninformed that chaperones possess steric information essential for protein folding; such grammatical forms should be avoided. Molecular chaperones and stress proteins are sometimes used as though they were interchangeable categories, but although many proteins share both chaperone and stress functions, this is not the case. A common error is to use the term “chaperonin” as synonymous with the term “chaperone,” but it should be noted that the chaperonins are just one particular family of chaperone defined by sequence similarity, i.e., the family that contains GroEL, hsp60, and TCP-1. The occasional use of the nonsense term “molecular chaperonin” in some respectable journals suggests that some people use these terms casually and without reference either to their meaning or to their history. It should be obvious that the word “molecular” is used to qualify “chaperone” because in common usage, “chaperone” refers to a person. So the term “molecular chaperonin” is as nonsensical as the term “molecular immunoglobulin.”

Another common misconception is that molecular chaperones are necessarily promiscuous, i.e., that each assists the assembly of many different types of polypeptide chains. This assumption is true for the hsp70, hsp40, and GroE chaperonin families but not for hsp90, PapD, hsp47, Lsm, Syc, ExbB, Pim, PrsA, and prosequences, which are specific for their substrates. Similarly, it is not a universal property of chaperones that they hydrolyze ATP: hsp100, hsp90, and hsp70, and the chaperonins hydrolyze ATP, but they trigger factor, hsp40, prefoldin, calnexin, protein disulfide isomerase, and PapD do not.

The term “chemical chaperone” has been proposed to describe small molecules such as glycerol, dimethylsulfoxide, and trimethylamine N-oxide that act as protein-stabilizing agents. This terminology is unfortunate and confuses students, because proteins are also chemicals. This term should be replaced by the term “kosmotrope” that physical chemists use to describe small molecules that stabilise proteins.

References


Further Reading


See Also

Protein Folding, Chemical Biology of Diseases
Protein Misfolding
Proteins, Chemistry and Chemical Reactivity of Proteins, Structure, Function and Stability; Self-Organization and Self-Assembly in Biology:
Protein Misfolding, Amyloid Formation

Andisheh Abedini, Joslin Diabetes Center, Harvard Medical School, Boston, Massachusetts

Daniel P Raleigh, Department of Chemistry and Institute for Chemical Biology and Drug Discovery, State University of New York at Stony Brook, Stony Brook, New York
doi: 10.1002/9780470048672.wecb484

Amyloids are β-sheet rich fibrillar protein structures that result from the self-assembly of polypeptides and proteins. A wide range of proteins are known to form amyloid fibrils in vivo and an even larger number do so in vitro. More than 20 different human diseases involve amyloid formation, and the amyloid fibril or intermediates populated during its assembly are cytotoxic. All amyloid fibrils share common structural features despite exhibiting considerable variation in primary sequence. These features include a crossed β-sheet organization in which the individual β-strands are arranged perpendicular to the fiber axis such that the hydrogen bonds are oriented parallel to the fibril axis, resistance to proteolysis, and the ability to bind to certain dyes. Normally, the aggregation of proteins into amyloid is a pathological event; however, evidence indicates that naturally occurring amyloids may also play a beneficial biological role in vivo.

A wide range of human diseases result from the inability of specific proteins to fold into their correct biologically active three-dimensional structures or result from the failure of proteins to remain in their properly folded states. These conditions are referred to broadly as protein misfolding diseases, and they result from a variety of causes. In some cases, the efficiency of folding may be compromised by a range of posttranslational events that lead to insufficient production of active proteins; however, many protein misfolding diseases are caused by the transformation of normally soluble proteins or polypeptides into ordered aggregates. The latter diseases are referred to commonly as amyloidoses. They represent a large group of diseases characterized by the deposition of insoluble ordered protein deposits that are known as amyloid fibrils or amyloid plaques (Table 1). The term "amyloid" was used first by Rudolph Virchow in 1854 when he used the word to describe a macroscopic tissue abnormality that exhibited a positive iodine staining reaction. For reviews of the history of amyloid, see References 1 and 2. Today the term amyloid refers to a specific type of protein quaternary cross-β structure that results from the self-assembly of peptides, polypeptides, and proteins into ordered aggregates. The first recognition of amyloid as something more than just an amorphous deposit was made in 1927, when polarized light microscopy was used to show that Congo Red-stained amyloid from a variety of tissues exhibited positive birefringence (3). These early observations of Congo Red birefringence promoted electron microscopic (EM) investigations of human amyloid. The first EM studies revealed a common fibrillar ultrastructure among the various amyloid deposits investigated (4). Since then, our understanding of the general architecture of amyloid fibrils has increased dramatically with the advancement and refinement of experimental techniques (5–12). Research into amyloid has exploded in recent years into a large and dynamic field that continues to grow and to develop. Because space limitations prevent a comprehensive discussion of this fascinating area of chemical biology, we will avoid in-depth discussions of specific amyloid diseases, or of the computational and theoretical studies that are emerging. The development of inhibitors of amyloid is an active area of research, but space limitations prevent a detailed discussion. We do note, however, that several amyloid inhibitors are in various stages of clinic trials, and their development bears watching.

In this article, we provide a brief overview of the fundamental aspects of amyloids by focusing on the common features of amyloid fibrils, which include their mechanisms of formation and their cytotoxicity. From the vast and ever-growing amyloid literature, we cite several key papers and some interesting historical references, and we provide citations to several recent review articles.
Protein Misfolding, Amyloid Formation

A Diverse Range of Proteins Form Amyloid

Amyloids are formed when normally soluble proteins or polypeptides aggregate and deposit as amyloid plaques in the tissues of affected individuals. In vivo, amyloid deposits consist of one major protein and, usually, a set of common minor components that are derived largely from building blocks of the basement membrane, which includes proteoglycans (10, 13, 14). Minor protein components of in vivo amyloid deposits include serum amyloid P component, collagen, and apolipoprotein E. In vitro, highly purified proteins can self-assemble into amyloid fibrils in the absence of these other components. A wide range of proteins are known to form amyloid in vivo, and even more proteins can form amyloid in vitro (Table 1). Indeed, it has been proposed that nearly all proteins can form amyloid under appropriate conditions, and the cross-β structure has been hypothesized to represent a default free energy minimum for the conformation of a polypeptide chain (15). Thus, it has been suggested that amyloid was an important early fold in pre-biotic evolution and perhaps has existed for as long as proteins (16).

Table 1: Prevalent pathological and functional amyloid and amyloid-like structures, and their major protein components

<table>
<thead>
<tr>
<th>Disease or amyloidosis</th>
<th>Aggregating protein</th>
<th>Amyloidosis type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amyloidotic polyneuropathy; familial amyloid cardiopathy; senile systemic amyloidosis</td>
<td>Transthyretin</td>
<td>Systemic</td>
</tr>
<tr>
<td>Finnish hereditary amyloidosis</td>
<td>Fragments of gelsolin mutants</td>
<td>Systemic</td>
</tr>
<tr>
<td>Huntington’s disease</td>
<td>Human huntingtin with expanded polyglutamine repeats</td>
<td>Local</td>
</tr>
<tr>
<td>Tuberculosis and Rheumatoid arthritis</td>
<td>Serum amyloid A</td>
<td>Systemic</td>
</tr>
<tr>
<td>Pulmonary alveolar proteinosis</td>
<td>Surfactant protein C (SP-C)</td>
<td>Local</td>
</tr>
<tr>
<td>Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy</td>
<td>Notch 3</td>
<td>Systemic</td>
</tr>
<tr>
<td>(CADASIL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cystic fibrosis, AA (secondary) amyloidosis</td>
<td>Amyloid A protein</td>
<td>Systemic</td>
</tr>
<tr>
<td>Serpinopathies</td>
<td>Serpins</td>
<td>Systemic</td>
</tr>
<tr>
<td>Aortic medial amyloidosis</td>
<td>Medin (lactadherin)</td>
<td>Local</td>
</tr>
<tr>
<td>Atrial amyloidosis</td>
<td>Atrial natriuretic factor</td>
<td>Systemic</td>
</tr>
<tr>
<td>Intracytoplasmic neurofibrillary tangles; Tauopathies</td>
<td>Tau protein</td>
<td>Local</td>
</tr>
<tr>
<td>Alzheimer’s disease; inclusion-body myositis; Down’s syndrome; retinal ganglion cell degeneration in glaucoma; Central β-amyloid angiopathy</td>
<td>Amyloid β peptide 40 and 42</td>
<td>Local</td>
</tr>
<tr>
<td>Hereditary cerebral haemorrhage with amyloidosis</td>
<td>Mutants of amyloid β peptide</td>
<td>Local</td>
</tr>
<tr>
<td>Familial British dementia</td>
<td>Aβ42</td>
<td>Local</td>
</tr>
<tr>
<td>Familial Danish dementia</td>
<td>A40</td>
<td>Local</td>
</tr>
<tr>
<td>Type II diabetes, pancreatic islet amyloidosis</td>
<td>Amylin, also known as IAPP</td>
<td>Local</td>
</tr>
<tr>
<td>Parkinson’s disease and other synucleinopathies</td>
<td>α-Synuclein</td>
<td>Local</td>
</tr>
<tr>
<td>Familial amyotrophic lateral sclerosis</td>
<td>Superoxide dismutase (SOD1); TDP-43</td>
<td>Local</td>
</tr>
<tr>
<td>Creutzfeldt-Jakob disease; bovine spongiform encephalopathy (mad cow disease); Gerstmann-Sträussler’s syndrome</td>
<td>Prion protein</td>
<td>Local and systemic</td>
</tr>
<tr>
<td>Injection-localized amyloidosis</td>
<td>Insulin</td>
<td>Local</td>
</tr>
<tr>
<td>Familial amyloidosis</td>
<td>Variants of fibrinogen α-chain</td>
<td>Local</td>
</tr>
<tr>
<td>Fibrinogen amyloidosis</td>
<td>Mutants of lysozyme</td>
<td>Systemic</td>
</tr>
<tr>
<td>Lysozyme amyloidosis</td>
<td>Apolipoprotein A1</td>
<td>Local</td>
</tr>
<tr>
<td>Restrictive amyloid heart; ApoAI amyloidosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ApoAII amyloidosis</td>
<td>Apolipoprotein A1</td>
<td>Local</td>
</tr>
<tr>
<td>ApoAIV amyloidosis</td>
<td>N-terminal fragment of apolipoprotein AIV</td>
<td>Local</td>
</tr>
<tr>
<td>Pulmonary alveolar proteinosis</td>
<td>Lung surfactant protein C</td>
<td>Nonpathologic</td>
</tr>
<tr>
<td>Glucagon amyloid-like fibrils</td>
<td>Glucagon</td>
<td></td>
</tr>
<tr>
<td>Cutaneous lichen amyloidosis</td>
<td>Keratins</td>
<td>Systemic</td>
</tr>
</tbody>
</table>
Despite the vast differences in the amino acid sequence of amyloid-forming polypeptides, all amyloid fibrils share common structural features and ligand binding properties. These properties include a crossed β-sheet organization in which the individual β-strands are arranged perpendicular to the fiber axis such that the hydrogen bonds are oriented parallel to the fibril axis, resistance to proteolysis, and the ability to bind to the dyes Congo Red and thioflavin-T (5–12, 17, 18). Typically, amyloid fibrils are 5 to 10 nm wide, unbranched, and variable in length. Transmission electron microscopy (TEM) and atomic force microscopy (AFM) studies have shown that amyloid fibrils are usually made up of protofilaments, each of which are typically 2–5 nm in diameter. Normally, the protofilaments twist together into rope-like structures with 2–6 protofilaments per fibril or self associate in a lateral fashion to generate long ribbons (7, 10, 11). The submicroscopic packing of the protofilaments varies according to differences in the primary amino acid sequence of the polypeptide and the environmental conditions in which amyloid formation is triggered. Because of the noncrystalline and insoluble nature of amyloids, it has been difficult to obtain detailed molecular structures of amyloidogenic proteins in their fibril form using conventional biophysical techniques such as X-ray crystallography and solution-state nuclear magnetic resonance (NMR) spectroscopy. Solid state NMR methods have been applied to the study of amyloid fibrils and have proven to be enormously useful, most notably in studies of the Aβ peptide (8, 19, 20). Solid-state NMR studies of synthetic amyloid fibrils have demonstrated that fibrils formed from peptides can be composed of parallel or antiparallel β-sheets, depends on the length and amphility of the peptide (20). Exciting recent work has led to high-resolution X-ray structures of several 7-mer polypeptides derived from amyloidogenic proteins (9, 12). These new structures reveal an extremely well-packed interface between pairs of β-sheets and the expected cross-β structure. It is not yet clear that these exceptionally well-packed structures will be representative of amyloid fibrils formed by larger polypeptides and proteins. One theory is that large globular proteins, like lysozyme (21), β2 microglobulin (22), and cystatins (23, 24) undergo three-dimensional domain swapping during amyloid fibril formation. Studies of amyloid-like fibrils of ribonucle-ase A with a Q10 hinge-loop expansion demonstrate domain swapping and functional native-like domains. This study indicates that the native-like conformation of the primary sequence can be maintained in the non-cross-β regions of the protein chain (25). It is, however, not clear whether the subunits of other domain-swapped amyloid-like fibrils might also consist of native-like polypeptide structures. How the non–cross-β regions of the protein chain are accommodated in the amyloid fibril is still unknown.
Protein Misfolding, Amyloid Formation

The highly ordered structures of amyloids allow the specific binding of histological dyes like Congo Red and thioflavin-T (3, 17, 18). Dye-binding studies have played a major role in amyloid research and form the basis of simple convenient assays; however, they are not without their pitfalls. For many years, Congo Red birefringence has been the standard test for the presence of amyloid (3, 17). Recent studies argue that Congo Red also binds to native proteins and lacks secondary structure specificity, which indicates that the dye is not specific for amyloid (26). Congo Red can also affect the rate of amyloid formation, either enhancing or inhibiting fibril formation, which emphasizes even more that it should be used with caution as a diagnostic tool for studying amyloid fibrils in vitro (27). Thioflavin-T is also used extensively for characterizing the presence of amyloid fibrils and their rate of formation (18). This dye, which can be detected by fluorescence, is a better alternative for in vitro fibril detection than Congo Red. Studies by confocal microscopy that used polarized light have indicated that thioflavin-T binds to amyloid with the long axis of the dye parallel to that of the fibril axis (i.e., perpendicular to the direction of the individual strands). The dye likely binds by inserting itself into the grooves formed between sidechains at positions i and i + 2 on the surface of the cross-β structure (28). For a flat β-sheet, approximately four strands are sufficient to generate a groove long enough to accommodate one molecule of thioflavin-T. It is thought that the steric constraints imposed by these surface grooves account for the relative specificity of the dye for the cross-β structure of amyloid fibrils.

Medically, amyloid diseases can be divided into three broad classes: neurodegenerative, systemic, and local amyloidosis (10). In neurodegenerative diseases, amyloids are deposited in the brain. Important examples in this class include the spongiiform encephalopathies, Huntington’s, Alzheimer’s, and Parkinson’s disease. In systemic amyloidoses, aggregation occurs in multiple organs and tissues. A subset of systemic amyloidoses includes senile systemic amyloidosis; lysozyme amyloidosis; A.L amyloidosis, which is caused by the aggregation of immunoglobulin light chains or fragments thereof; familial transthyretin-associated amyloidosis, which develops from de- position of wild-type or one of more than 50 mutated forms of transthyretin (TTR); and diseases of chronic inflammation, in which an N-terminal fragment of the acute phase protein serum amyloid A forms amyloid deposits. In the non-neurological, lo- calized amyloidoses, amyloid deposition occurs in one target organ, which is usually proximal to the production site of the amyloidogenic peptide. Common examples of this third class include medullary thyroid carcinoma, which is associated with amyloid deposition of prealbumin; familial amyloidosis, which is caused by a familial nanocrine factor; amyloid formation by the γ-crystallins associated with cataracts; type 2 diabetes, which is characterized by deposition of β-sheet amyloid polypeptide (1A PP also known as amylin); as well as many others (Table 1). Some diseases are sporadic whereas others are hereditary. The associ- ation between amyloid fibril formation and disease pathogenesis is common for all amyloidoses, and the cytotoxic properties of amyloidogenic peptides are well documented (29, 30). The ex- act mechanism of amyloid-induced cell death is, however, still not completely clear.

The proteins and polypeptides that form amyloid can be di- vided into two broad structural classes: namely those that adopt a well-defined tertiary fold in their normally soluble state and those that are flexible and intrinsically disordered in their un- aggregated state. Normally, proteins that adopt a well-defined globular fold in their unaggregated state require a partial unfolding event to become aggregation competent. Particu- larly well-characterized examples include TTR, β2 microglob- ulin (responsible for dialysis-related amyloid), and lysozyme (21, 22, 31–35). Molecules that bind and stabilize the native states are potential therapeutic agents for amyloid diseases that develop from folded proteins, because they reduce the ten- dency to populate the amyloidogenic precursor states (31). In some cases, natively unfolded polypeptides also seem to re- quire a partial fold to form amyloid. Studies on tau (associ- ated with neurofibrillary tangles) and Aβ (which forms amyloid deposits in Alzheimer’s disease) suggest that amyloid forma- tion in these systems is preceded by formation of a helical intermediate (36, 37).

Common Features of Amyloid Formation

An increasing number of studies on a variety of amyloidogenic peptides and proteins argue that underlying commonalities exist in mechanisms of amyloid formation at least in vitro, which are independent of the details of the polypeptide sequence. This finding is true for both of the structural classes of amyloids defined above. The exact mechanism of amyloid formation has not been determined fully; however, extensive experimental ev- idence indicates that amyloid formation proceeds by a variation of the so-called nucleation-dependent polymerization pathway (38, 39). The kinetics of amyloid formation are complex and typically exhibit a lengthy lag phase during which little or no amyloid is formed, which is followed by a much more rapid growth phase. Oligomeric nuclei are formed during the slow lag phase. Once a critical assembly of precursors form an ac- tive seed, a second, more rapid phase of fibril polymerization occurs, which leads to the classic amyloid morphology. Amy-loid formation can be accelerated substantially by the addition of preformed seeds (40). Often, the lag phase can be abolished by seeding a solution of unaggregated peptide with a small amount of preformed fibrils (Fig 1). Seeding is generally, although not completely, specific. Fibrils often preferentially, but not always, seed reactions that contain the same polypeptides from which they were formed. This finding indicates that primary structure similarity is important for binding interactions between seeds and polypeptide monomers. Although the general ultrastructure of all amyloids seem similar, clear morphological differences exist between amyloid fibrils formed from different proteins. These variations in morphology are caused by differences in the molecular packing and the organization of fibril subunits, as influenced by the polypeptide primary sequence. Other external environmental factors such as solution pH, ionic strength, tem- perature, and agitation have also been shown to affect fibrillar morphologies of amyloids produced in vitro.
There is a rich experimental and theoretical literature exists on protein assembly and aggregation, and various kinetic models have been used to rationalize the time course of amyloid formation (38-41). Perhaps the classic example of a protein assembly reaction is that of actin, which is described by a nucleation-dependent mechanism. The nucleation-dependent model comprises three steps: 1) association of monomers into oligomers, 2) conversion of the oligomers to a nucleus, and 3) growth of the final fibril by addition of monomers to the end. The nucleation-dependent polymerization model cannot explain all amyloid formation reactions. The existence of the characteristic lag phase is not well predicted. Double nucleation schemes, which were developed initially by Mozzarelli et al. (42) during their pioneering studies of hemoglobin S polymerization, have been invoked to explain the existence of a lag time. Double nucleation models include primary nucleation steps similar to the nucleation-dependent polymerization model, but they also invoke a second nucleation step that is dependent on the presence of fibrils (i.e., the fibrils can act as secondary nucleation sites). The second step is unimportant before fibrils are formed but becomes increasingly important as the amount of fibrils increases. This mechanism was used early on to model amyloid formation kinetics; however, the dependence of the lag phase on protein concentration in amyloid formation is not always well predicted by this model. In particular, the model predicts a lag time that scales much more strongly with monomer concentration than is sometimes observed. Variants of the original double nucleation scheme have been developed that can account for the weakened concentration dependence. Irrespective of the details, the view that amyloid formation proceeds by a cooperative, nucleation-polymerization mechanism is widely accepted (10, 40).

The structure shared by amyloid fibrils and the common themes in its assembly provide a unifying mechanism of cell toxicity: thus, a better understanding of the amyloid assembly process at the molecular level should give invaluable insight into the identification and the development of effective therapeutic innovations for a wide variety of human amyloid diseases. Amyloid fibrils or their precursors are toxic and are directly responsible for cell death. The majority view in the field is that partially structured oligomers that appear as intermediates in the aggregation process may be the toxic entities, rather than the mature fibrils (43-49). In this view, amyloid may represent a relatively benign state that sequesters the protein, and if true, may suggest that several compounds that have been developed to inhibit the assembly of amyloid fibrils could actually be harmful if they lead to the build up of toxic prefibrillar intermediates. It is important to realize, however, that good evidence suggests both fibrils and prefibrillar species can be toxic, at least in some cases. For example, the pathological effects in the systemic amyloidoses are caused by disruption of organ function by the formation of large amyloid deposits. In addition, the results of several recent studies are very difficult to reconcile with the “only oligomers are toxic” model. For example, the Tottori and English Familial Alzheimer disease mutations are associated with aggressive early onset of Alzheimer’s disease, yet their effect is to increase the rate of fibril formation without increasing meta-stable intermediates. Thus, a detailed understanding of the mechanism of amyloid fibril formation and the identification of the toxic species and their mode of action are subjects of major importance and are currently the focus of considerable research.

Until very recently, the aggregation of proteins into ordered amyloid fibrils was thought to be a pathological event that leads to cytotoxicity and highly debilitating diseases; however, amyloid need not always be deleterious. Amyloid-like structures may have applications in bio-nanotechnology, which includes roles as templates for the assembly of novel structures, as materials, and as cell-supporting matrices (51, 52). In addition, evidence now suggests that naturally occurring amyloids can play beneficial biological roles in vivo. Several functional amyloid fibrils derived from a diverse range of single and multi-cellular organisms have been identified and described recently (49). A list of functional amyloid and amyloid-like structures is provided in Table 1.

A particularly well studied example of functional amyloid is provided by Curli assembly (53). Curli amyloids are assembled by bacteria such as Escherichia coli and Salmonella. Once assembled on the extracellular surface, Curli amyloid fibers function as natural cell adhesion molecules that link together bacterial cells into robust cellular networks of biofilms. Other examples of functional amyloids include the silk fibers observed commonly in spider webs; the Chorion proteins of egg shells; Factor XII, which is an activator of the hemostatic system; and other naturally produced adhesives and materials (54).

In humans, the formation of Pmel17 amyloid has been shown to be important for melanin formation (55). Melanin is a natural pigment for skin coloration, UV protection, and chemical detoxification, and it is synthesized in melanosomes. Pmel17 amyloid fibrils have been shown to protect melanosomes against pigment melanin toxicity by decreasing the diffusion of toxic melanin precursor molecules out of the cell. The formation and the degradation of functional amyloids seem to be highly regulated, which is hardly surprising given the toxicity associated with nonfunctional amyloids (54, 55).

Figure 1: Schematic representation of the progress of fibril formation. The solid line represents an unseeded reaction. A distinct lag phase is observed followed by a rapid growth phase. Seeding (dashed line) can bypass the lag phase.
A amyloid structures can also act as nonchromosomal genetic elements, which lead to non-Mendelian inheritance. This finding can have drastic and disastrous effects in the prion diseases. However, several amyloid forming prions have been suggested to be potentially functional, which include the yeast proteins Het-s and p-prion. The Het-s prion, which is found in 80% of wild isolates, carries heterokaryon incompatibility to prevent infection by incapacitating fungal viruses (56, 57). The p-prion, which has been shown to be necessary for meiosis and for cell survival in the stationary phase (58), has also been proposed to be potentially beneficial for fungal cells. Other prion proteins, like Sup35p, Ure2p and Rnq1p, which have been demonstrated to form amyloid formers in the cytoplasm in a non-Mendelian epigenetic fashion, have been argued to provide evolutionary advantages for host cells (59-61). This hypothesis is intriguing and will no doubt be the focus of much attention.

The study of the assembly, degradation, and regulation of functional amyloids may well provide important clues for controlling aberrant amyloid formation, and a greater understanding of the diverse physiological applications of the amyloid fold may open potential new avenues for the treatment of amyloid diseases. The discovery of functional amyloids has interesting implications, which may open potential new avenues for the treatment of amyloid diseases. The discovery of functional amyloids has interesting implications. Whether this issue is critical remains to be determined.

References


Further Reading


Protein Misfolding, Amyloid Formation

See Also

Prion Diseases: Chemical Biology of
Protein Misfolding: Amyloid Formation
Proteins: Computational Analysis of Structure, Function and Stability
Protein Folding: Chemical Biology of Diseases Related to
Chaperones, Molecular
Intramembrane Proteolysis

Michael S. Wolfe, Center for Neurologic Diseases, Harvard Medical School and Brigham and Women’s Hospital, Boston, Massachusetts
doi: 10.1002/9780470048672.wecb258

Evolutionarily conserved membrane-embedded enzymes somehow use water to hydrolyze peptide bonds that reside within the lipid bilayer. These proteases are stitched into the membrane and bear little or no sequence similarity to other known proteases with one exception: All known intramembrane proteases apparently have converged on catalytic mechanisms that resemble those of proteases that reside in the aqueous milieu. The identities of these residues suggest that the intramembrane proteases use catalytic functionalities similar to those found in their soluble counterparts. However, in contrast to classical proteases, residues essential for catalysis lie within predicted hydrophobic transmembrane domains. Current research is focused on elucidating the range of biologic roles of these proteases, identifying their substrates, and understanding their structure and mechanism of action. Such knowledge is important because the intramembrane proteases discovered thus far play critical regulatory roles in biology and contribute to human disease.

Proteases catalyze the hydrolysis of the amide bonds that link amino acids together in peptides and proteins, and this process requires the concerted effort of key residues within the active site of the enzyme. These hydrolytic enzymes are classified into four general types based on their catalytic residues and mechanism of action: 1) serine/threonine proteases, 2) cysteine proteases, 3) aspartyl proteases, and 4) metalloproteases. Each of these four main protease categories contains hundreds of known examples and has representatives in all forms of life (1). Until recently, all the identified proteases had been water-soluble enzymes: Either the entire enzyme normally is found in an aqueous environment or a membrane anchor holds down an otherwise aqueous protease. In recent years, however, new proteases have been discovered that apparently are embedded within the hydrophobic environment of the lipid bilayer and yet somehow carry out hydrolysis on the transmembrane region of their substrates in the generally water-excluding environment of the membrane. Another unusual feature of this process are the substrates, which typically are folded into an α-helix, a conformation that makes the backbone amide bonds inaccessible to nucleophilic attack because of steric hindrance by the amino acid side chains. These intramembrane-cleaving proteases (I-CLiPs) therefore must create an environment for water and the hydrophilic residues needed for catalysis and must bend or unwind their substrates to make the amide bonds susceptible to hydrolysis. Supporting these mechanistic notions is the observation that these newly discovered I-CLiPs apparently are variations on familiar themes in protease biochemistry: Despite the novelty of being membrane-embedded and cleaving transmembrane domains, the residues essential for catalysis by these I-CLiPs are virtually the same as those found in aqueous proteases.

The S2P Family of Metalloproteases

The first discovery of an I-CLiP arose from studies on the regulation of sterol and fatty acid metabolism. Sterol regulatory element binding proteins (SREBPs) are transcription factors that promote the expression of genes involved in the synthesis of cholesterol and fatty acids (3). Coordinated gene expression is controlled through negative feedback inhibition by cholesterol to ensure that lipids and sterols are produced only when needed. SREBPs are synthesized as a precursor protein that contains three distinct domains: a domain exposed to the cytosol that binds DNA and activates transcription, two transmembrane regions, and a regulatory domain involved in the feedback control by cholesterol (Fig. 1). When cholesterol levels are high, the SREBP precursor is kept in the endoplasmic reticulum (ER) by a multipass membrane protein called SCAP (SREBP cleavage-activating protein) (4). Reduced cholesterol levels result in the dissociation of Insig from SCAP, which allows SCAP to shepherd SREBP to the Golgi apparatus. Proteolysis of SREBP in the Golgi results in the release of the transcription factor and its translocation to the nucleus.

Proteolytic release of SREBPs occurs in two steps (Fig. 1). First, the luminal loop between the two transmembrane regions is cleaved by the membrane-tethered Site-1 protease (S1P) (6).
The release of the transcription factor requires subsequent cleavage by the Site-2 protease (S2P), which performs a hydrolysis of an amide bond predicted to lie three residues within the transmembrane domain (7). The requirement for a prior proteolytic event is a common theme with I-CLiPs. Complementation identified S2P as a multipass membrane protein that contains a conserved HEXXH sequence characteristic of zinc metalloproteases (8). The two histidines and the glutamate are required for S2P activity, consistent with known metalloprotease biochemistry in which the two histidines coordinate with zinc and the zinc in turn activates the glutamate for interaction with the catalytic water. Additional analysis led to the discovery of a conserved aspartate located ∼300 residues from the HEXXH sequence that is likewise critical for S2P activity and thought to be a third residue involved in zinc coordination (9). The involvement of zinc in S2P activity has not been demonstrated, and a cell-free assay for S2P activity has not been reported yet; therefore, S2P has not been shown directly to act as a protease (however, see below about bacterial S2Ps). Nevertheless, extensive genetic analysis has not uncovered any other proteins required for S2P cleavage of SREBP. Similar to SREBP, sequential processing by S1P and S2P of the otherwise membrane-associated transcription factor ATf6 are essential steps in the ER stress response (10).

More support for the proteolytic function of S2P comes from the discovery of a family of related proteins in bacteria (11). These prokaryotic proteins play an essential role in the proteolysis of otherwise membrane-bound transcription factors needed for sporulation. These factors control gene expression in the mother cell after engulfment of the forespore. The cleavage of pro-αk and the release of the transcription factor requires the multipass membrane protein SpoIVFB in Bacillus subtilis, and this protein likewise contains the HEXXH motif and a second conserved region with an aspartate, both of which are essential for proteolysis. Another bacterial S2P family member, YaeL (also called RseP) in Escherichia coli, similarly requires HEXXH and a conserved aspartate to play a role in coordinating cell growth and cell division, through intramembrane proteolysis of RseA, a factor critical for responding to extracytoplasmic stress (12). Interestingly, the membrane orientations of the substrates SREBP and αk are opposite to each other, which correlates with that of their respective enzymes, S2P and SpoIVFB, which similarly are thought to have opposite orientations (11). This opposition implies that the catalytic region must align with the peptide substrate with proper relative directionality. Although SpoIVFB and YaeL are both S2P-like enzymes that cleave transmembrane proteins during sporulation, the regulation of this key intramembrane proteolytic event for these two I-CLiPs is quite different. For cleavage of RseA by YaeL/RseP, the regulation is similar to that in SREBP cleavage by S2P: Intramembrane proteolysis requires a prior cleavage event outside the membrane by another protease called DegS (13). In contrast, SpoIVFB apparently does not require a prior proteolysis and regulation occurs more directly at the level of SpoIVFB. Two membrane proteins, Boa and SpoIiFA, serve to inhibit SpoIVFB activity, and this inhibition is released by the proteolysis of SpoIiFA by other proteases (14, 15) (Fig. 2).

As mentioned above, the α-helical conformation of the transmembrane substrate renders the amide bonds inaccessible to attack by a catalytic residue or water, which requires some bending or unwinding of the helix before proteolysis can occur. The SREBP substrate contains a conserved asparagine-proline (NP) sequence within its transmembrane region that is critical for proteolytic processing by S2P (16). These two residues have the lowest propensity to form α-helices, which suggests that the NP-containing SREBP transmembrane region may be metastable. After S1P cleavage and dissociation of the other transmembrane region, the NP sequence may facilitate unwinding of the residues immediately upstream, including the leucine-cysteine bond that gets cleaved. Unwinding may result in a protrusion of this bond to the membrane surface and access by the active site residues of S2P.

So far, one member of the S2P family has been purified to homogeneity with preservation of proteolytic activity (17). The purification of this protease, the E. coli YaeL/RseP, should allow a more rigorous and direct determination of substrate sequence requirements. Moreover, the ability to purify an I-CLiP and reconstitute activity in an appropriate detergent and usually in the presence of added lipids is an important step toward studying these unusual membrane-embedded proteases in terms

Figure 1  S2P contains conserved HEXXH and LDG motifs found in metalloproteases. SREBP first is cleaved by S1P in the luminal loop. The regulatory domain (Reg) interacts with the cholesterol-sensing SCAP to ensure that S1P proteolysis only occurs when cholesterol levels are low. Subsequent intramembrane proteolysis releases this transcription factor for the expression of genes essential to cholesterol and fatty acid synthesis.
Intramembrane Proteolysis

When the mother cell engulfs the forespore, a signaling pathway that involves the transcription factor $\sigma^G$ is initiated in the forespore that triggers the synthesis of the IVB serine protease. This protease degrades SpoIVFA, which along with BofA serves to inhibit SpoIVFB. With the inhibition of the S2P-like protease released, SpoIVFB cleaves pro-$\sigma^G$, which allows this transcription factor to signal in the mother cell for more factors needed for spore maturation.

of structural biology. Indeed, one such protease has been seduced to crystallize, which provided the first detailed structures of an I-CLiP (see the section below on the Rhomboid serine proteases).

$\gamma$-Secretases: Presenilin-Containing Aspartyl Protease Complexes

A key step in the pathogenesis of Alzheimer’s disease is APP proteolysis that results in the formation of the amyloid-$\beta$ peptide (Aβ), the principle protein component of the characteristic cerebral plaques of the disease (18). The N-terminus of Aβ is produced from the amyloid $\beta$-protein precursor (APP) by the action of $\beta$-secretase, which leads to membrane shedding of the large luminal/extracellular APP domain (Fig. 3a). The 99-residue remnant (C99) then is cleaved in the middle of its transmembrane region by $\gamma$-secretase, which releases Aβ and is cleaved again near the inner leaflet at the $\epsilon$ site to release the APP intracellular domain (AICD). As described below, chemical probes played important roles in the characterization, identification, purification, and mechanistic understanding of the I-CLiP that now is known as the $\gamma$-secretase complex.

Two contemporaneous observations provided critical clues for the identification of the elusive $\gamma$-secretase, a subject of intense interest as a potential therapeutic target. First, the knock-out of presenilin genes eliminated the $\gamma$-secretase cleavage of APP (19). Second, the types of compounds that could inhibit $\gamma$-secretase contained moieties typically found in aspartyl protease inhibitors (20). These findings led to the identification of two conserved transmembrane aspartates in the multipass presenilin that are critical for $\gamma$-secretase cleavage of APP (Fig. 3a), which suggests that presenilins might be the responsible aspartyl proteases (21). Presenilin is cut into two pieces, an N-terminal fragment (NTF) and a C-terminal fragment (CTF), the formation of which results in the liberation of AICD.

Figure 3  Presenilin, the $\gamma$-secretase complex, and the proteolysis of APP to Aβ. (a) Presenilin is processed into two pieces, an N-terminal fragment (NTF, dark portion) and a C-terminal fragment (CTF, light portion) that remain associated. Each fragment donates one aspartate that is essential for $\gamma$-secretase activity. APP is cleaved first in the extracellular domain by $\beta$-secretase, and the remnant is cleaved twice within the membrane by $\gamma$-secretase to produce the Aβ peptide of Alzheimer’s disease (secreted) and the intracellular domain (AICD, freed into the cytosol). (b) Presenilin interacts with three other membrane proteins, nicastrin, Aph-1, and Pen-2, to form active $\gamma$-secretase.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Intramembrane Proteolysis

**Figure 4** Chemical tools for the study of γ-secretase. Transition-state analog inhibitors include hydroxyl-containing moieties that interact with the catalytic aspartates of aspartyl proteases. Helical peptide mimics the APP transmembrane domain and interact with the substrate docking site on the protease. These potent inhibitors were converted into affinity labeling reagents that contain a chemically reactive bromoacetyl or photoreactive benzophenone for covalent attachment to the protein target and a biotin moiety to allow isolation and detection of the labeled protein. Both types of chemical probes interacted with the two presenilin subunits but at distinct locations, which suggests that both the active site and the docking site of γ-secretase lie at the interface between these subunits.
Unlike the other intramembrane proteases, the enzyme does not recognize by process so many different substrates. The broad substrate specificity to be unique among intramembrane proteases in its ability to cleave so many different substrates and its apparently poor substrate recognition (see below). In the interacting with ligands such as the protease Delta and Jagged triggers a shedding of the ectodomain by membrane-anchored metalloprotease-mediated cleavage at S2. The remnant then is cleaved at least twice, at the S3 and S4 sites, to release the Notch counterpart of Aβ (Aβ) and the intracellular domain (NICD). The latter translocates to the nucleus where it interacts with transcription factors to influence gene expression relevant to cell differentiation.

Figure 5  
Proteolytic processing and signaling of the Notch receptor. In the ER, Notch is cleaved at S1 by a furin-like protease to produce a stable heterodimeric receptor that is trafficked to the cell surface. Interaction with ligands such as the proteins Delta and Jagged triggers a shedding of the ectodomain by membrane-anchored metalloprotease-mediated cleavage at S2. The remnant then is cleaved at least twice, at the S3 and S4 sites, to release the Notch counterpart of Aβ (Aβ) and the intracellular domain (NICD). The latter translocates to the nucleus where it interacts with transcription factors to influence gene expression relevant to cell differentiation.

---

**Figure 4**  
Intramembrane Proteolysis

---

**Figure 3**  
Proteolytic processing and signaling of the Notch receptor. In the ER, Notch is cleaved at S1 by a furin-like protease to produce a stable heterodimeric receptor that is trafficked to the cell surface. Interaction with ligands such as the proteins Delta and Jagged triggers a shedding of the ectodomain by membrane-anchored metalloprotease-mediated cleavage at S2. The remnant then is cleaved at least twice, at the S3 and S4 sites, to release the Notch counterpart of Aβ (Aβ) and the intracellular domain (NICD). The latter translocates to the nucleus where it interacts with transcription factors to influence gene expression relevant to cell differentiation.
Intramembrane Proteolysis

DSI

NTF

CTF

TSA

S

Figure 6. Model for how inhibitors and substrates interact with presenilin. Helical peptides are docking site inhibitors (DSIs) and interact on the outside of the presenilin molecule at the NTF/CTF heterodimeric interface. Transition-state analog inhibitors (TSAs) interact on the inside of the presenilin molecule where the active site resides. The active site, which contains water and two aspartates, is thought to be sequestered away from the hydrophobic environment of the lipid bilayer. These findings have implications for how substrate interacts with the enzyme. The transmembrane domain of the substrate (S) interacts with the docking site and passes either in whole or in part into the active site for proteolysis.

Up until recently, all the action seemed to be taking place on presenilin. However, an elegant study has demonstrated that nicastrin also plays a critical role in substrate recognition (54). The ectodomain of nicastrin bears sequence resemblance to aminopeptidases, although certain catalytic residues are not conserved. Nevertheless, nicastrin recognizes the N-terminus of γ-secretase substrates derived from APP and Notch, and mutation of the aminopeptidase domain prevents this interaction. One conserved glutamate is especially important because this residue forms an ion pair with the amino terminus of the substrate. The sequence of the substrate N-terminus apparently is not critical for the interaction, but a free amino group is. Indeed, the simple formylation of the substrate N-terminus is enough to prevent effective substrate interaction and proteolytic processing. Thus, nicastrin can be thought of as a kind of gatekeeper for the γ-secretase complex: Type I membrane proteins that have not shed their ectodomains cannot interact properly with nicastrin and do not gain access to the active site.

Although γ-secretase has in many ways been an attractive target for Alzheimer therapeutics (see inhibitor in clinical trials, Fig. 7), interference with Notch processing and signaling may lead to toxicities that preclude the clinical use of inhibitors of this protease. Long-term treatment with γ-secretase inhibitors causes severe gastrointestinal toxicity and interferes with the maturation of B- and T-lymphocytes in mice, effects that are indeed because of the inhibition of Notch processing and signaling (55, 56). However, compounds that can modulate the enzyme to alter or block Aβ production with little or no effect on Notch would bypass this potential roadblock to therapeutics. Recent studies suggest that the protease complex contains allosteric binding sites that can alter substrate selectivity and the sites of substrate proteolysis. Certain nonsteroidal anti-inflammatory drugs (NSAIDs; e.g., ibuprofen, indomethacin, and sulindac sulfide) can reduce the production of the highly aggregation-prone Aβ peptides.

Figure 7. γ-secretase inhibitor (LY-450,139) and NSAID-like modulator (R-flurbiprofen) in clinical trials for Alzheimer’s disease.
A (A2) peptide and increase a 38-residue form of Aβ, a pharmacologically independent of inhibition of cyclooxygenase (57). The alteration of the proteolytic cleavage site is observed with isolated or purified γ-secretase (58-60), which indicates that the compounds can interact directly with the protease complex to exert these effects. Enzyme kinetic studies and displacement experiments suggest that the selective NSAIDs can be noncompetitive with respect to APP substrate (60) and to a transition-state analog inhibitor, which suggests an interaction with a site distinct from the active site (61). The site of cleavage within the Notch transmembrane domain is affected similarly, but this subtle change does not inhibit the release of the intracellular domain and thus does not affect Notch signaling (62).

For this reason, these agents may be safer as Alzheimer thera
pagnics than inhibitors that block the active site or the docking site. Indeed, one compound, R-flurbiprofen (Fig. 7), has advanced recently to Phase III clinical trials. Surprisingly, the site of proteolytic cleavage by the presenilin homologue SPP also can be modulated by the same NSAIDs that affect γ-secretase. Because SPP apparently does not require other protein cofactors, these findings suggest that presenilin is the site of NSAID binding within the γ-secretase complex and that SPP and presenilin share a conserved drug binding site for the allosteric modulation of substrate cleavage sites (63).

Another type of allosteric modulator is compounds that seem like kinase inhibitors and interact with a nucleotide binding site on the γ-secretase complex. The discovery that ATP can increase Aβ production in membrane preparations prompted the testing of a variety of compounds that interact with ATP binding sites on other proteins (64). In this focused screen, the α2-kinase inhibitor Gleevec emerged as a selective inhibitor of Aβ production in cells without affecting the proteolysis of Notch. In light of these findings, ATP and other nucleotides were tested for effects on purified γ-secretase preparations and found to increase selectively the proteolytic processing of a purified recombinant Aβ-based substrate without affecting the proteolysis of a Notch counterpart (65). Furthermore, certain compounds known to interact with ATP binding sites were found to inhibit selectively the APP processing via a γ-secretase in purified protease preparations. These and other results suggest that the γ-secretase complex contains a nucleotide binding site and that this site allows the allosteric regulation of γ-secretase processing of APP with respect to Notch. Whether this regulation is physiologically important is unclear, but the pharmacologic relevance is profound and may lead to new therapeutic candidates for Alzheimer’s disease.

The purification of the γ-secretase complex (59) has allowed the first glimpse into its structure. Electron microscopy and single particle analysis reveals that the complex has a globular structure that at low resolution (∼15 Å) appears rather amorphous (66). A similar structure, elucidated in a similar manner but of poorer resolution (∼45 Å), also has been reported (67). Nevertheless, two important features can be gleaned. The first is a rather large interior cavity of ∼20 Å diameter that is presumed where the active site resides, a characteristic reminiscent of the proteasome. The second is the presence of two small openings that may be the site of entry for water. Other structural features have been revealed by cysteine mutagenesis with the cross-linking of chemical probes (68, 69). The generation of a cysteine-less version of presenilin (that retains the ability to assemble with other complex members, to undergo endo-proteolysis to NTF and CTF, and to process APP) allowed incorporation of single cysteine residues at various sites near the key aspartates. Disulfide formation with thiol-containing reagents then provided information about the relative accessibility of these sites from the aqueous milieu, which allowed the construction of a model in which water can funnel down to where the aspartates reside. More detailed information likely will require a crystal structure of presenilin or a presenilin homologue (see below).

**SPP Aspartyl Proteases**

The concept of presenilin as the catalytic component for γ-secretase was strengthened considerably when signal peptide protease (SPP) was found to be a similar intramembrane aspartyl protease. SPP clears remnant signal peptides from the membrane after their production by signal peptidase (Fig. 8). However, this process apparently also plays a role in immune surveillance, in which signal peptides from the major histocompatibility complex (MHC) type I are cleaved by SPP and the peptide products are presented onto the cell surface as an indication to natural killer cells whether MHC synthesis is proceeding normally (70). In addition, SPP is exploited by the hepatitis C virus for the maturation of its core protein, which suggests that this protease may be a suitable target for antiviral therapy (71).

SPP was identified by affinity labeling with a peptidomimetic inhibitor, and the protein sequence displayed intriguing parallels with presenilin (Fig. 8) (72). SPP contains two conserved aspartates, each predicted to lie in the middle of a transmembrane domain, and the aspartate-containing sequences resemble those found in presenilins. The predicted topology of SPP also resembles that of presenilins, placing the key aspartates in the same relative position to each other in the membrane. As with SPP compared with its bacterial relatives, the orientation of the aspartate-containing transmembrane domains of SPP apparently is opposite that of presenilins, again in correlation with the orientation of SPP substrates, which is opposite that of γ-secretase substrates. Interestingly, before the identification of SPP, a computational search for presenilin-like proteins netted an entire family of so-called presenilin homologs (PSHs) (73); however, it is not yet clear if all these proteins have catalytic activity. Two homologs, SPP-like proteases SPLL2a and SPLL2b, recently have been found to cleave tumor necrosis factor a (74), although the biologic role of this proteolysis is unknown.

SPP seems to be less complicated than γ-secretase. The expression of human SPP in yeast reconstituted the protease activity, which suggests that the protein has activity on its own and does not require other mammalian protein cofactors (72). Moreover, unlike presenilins, SPP is not processed into two pieces. Thus, SPP may be a more tractable enzyme for understanding this type of intramembrane aspartyl protease and may shed light on γ-secretase structure and function. Indeed, the catalytic sites of the two proteases seem remarkably similar; their activities...
Intramembrane Proteolysis

Figure 8  Comparison of signal peptide peptidase (SPP) with presenilin and the γ-secretase complex. Signal peptides are removed from membrane proteins via signal peptidase (SP), and these peptides are released from the membrane by SPP-mediated intramembrane proteolysis. SPP, like presenilin, contains two aspartates that are essential for protease activity, but the conserved aspartate-containing motifs are in the opposite orientation compared with their presenilin counterparts. Consistent with the flipped orientation of SPP vis-à-vis presenilin, the substrates of these two proteases also run in the opposite direction. Unlike presenilin, SPP apparently does not require other protein cofactors or cleavage into two subunits for proteolytic activity. Are inhibited by some of the same active site-directed peptidomimetics (75, 76) and helical peptides (77), and activity can be modulated by the same NSAIDs that affect γ-secretase (77). SPP forms a homodimer very rapidly in cells, and this dimer is stable enough to allow isolation and analysis (78). Moreover, this dimer can be labeled specifically by a transition-state analog inhibitor, which suggests that the dimer is catalytically active. The functional importance of this dimer, however, is unclear; dimerization may not be necessary for proteolytic activity. In terms of substrate recognition, however, SPP does display an important difference with γ-secretase: the apparent requirement for helix-breaking residues that should facilitate the ability of the enzyme to access the site of hydrolysis (79).

Rhomboid Serine Proteases

The study of a conserved growth factor signaling pathway also led to intramembrane proteolysis. Epidermal growth factor (EGF) receptor ligands are synthesized as single-pass membrane proteins, but signaling requires the proteolytic release and secretion of the ligand for interaction with its cognate receptor. In vertebrates, this release and secretion is accomplished by membrane-tethered metalloproteases. Genetic analysis in Drosophila, however, identified two essential players, dubbed Star and Rhomboid-1, in the proteolysis of an EGF ortholog Spitz. No other components apparently are required. Full-length Spitz remains in the ER until it is ushered by Star to the Golgi apparatus where it encounters Rhomboid-1 (80). Rhomboid-mediated proteolysis in the Golgi then is followed by secretion for intercellular communication. But how does Rhomboid allow cleavage of Spitz?

Mutational analysis of conserved non-glycine residues revealed a tantalizing requirement for a serine, a histidine, and an asparagine, which together might serve as a catalytic triad typically found in serine proteases (81) (although subsequent studies supported Ser-His dyad (82)) (Fig 9). These three residues were predicted to reside about the same depth within the membrane and thus have the potential to interact with each other. Consistent with this idea, the cleavage site of Spitz was estimated to be at an equivalent depth in the transmembrane region, and Spitz...
Rhomboids contain a conserved serine and histidine, which comprise a putative catalytic dyad of a serine protease. Rhomboid-1 cleaves within the transmembrane region of the Drosophila EGF-like growth factor Spitz. Cleavage was sensitive only to serine protease inhibitors. Moreover, a careful analysis of concentration dependence revealed that expression of catalytic amounts of Rhomboid-1 still allowed Spitz proteolysis. Taken together, Rhomboid-1 apparently is a novel intramembrane serine protease.

What determines Rhomboid substrate specificity, and how is this proteolytic event regulated? Most of the Spitz transmembrane region could be swapped with that of a nonsubstrate protein without affecting cleavage by Rhomboid; however, the N-terminal quarter of the transmembrane region was critical for substrate recognition (83). Indeed, incorporation of this substrate motif into Delta allowed this Notch ligand to be processed by Rhomboid. Additional examination of the substrate motif led to the tentative identification of a critical glycine-alanine, which suggests that, as with S2P and SPP, Rhomboid seems to require helix-destabilizing residues within the transmembrane domain of its substrates. Rhomboid activity is distinguished from that of the other I-CLiPs because Rhomboid does not require prior substrate cleavage by another protease. Rhomboid regulation apparently occurs mainly by the translocation of the substrate from the ER to the Golgi (mediated by Star) and the spatial control of Rhomboid transcription.

Like S2P, Rhomboid genes have been conserved throughout evolution. Surprisingly, despite overall low homology with eukaryotic Rhomboids, several bacterial Rhomboids could cleave Drosophila Rhomboid substrates and mutation of the putative catalytic triad residues abolished protease activity, which illustrates the evolutionary conservation of the serine protease function of Rhomboid. The natural substrates for the bacterial Rhomboids are unknown. As for substrates of eukaryotic Rhomboid-1 homologues, two mitochondrial membrane proteins have been identified as substrates for the yeast Rhomboid RBD1 (85–87). The RBD1-mediated release of one of these substrates is essential for the remodeling of the mitochondrial membrane, and the human orthology of RBD1, PARL, could restore the substrate proteolysis and proper growth rates and mitochondrial morphology in a yeast RBD1 mutant (86), which suggests that the role of these Rhomboids in mitochondrial function has been conserved evolutionarily. Indeed, a recent study identified a mitochondrial protein DPA1 as a likely substrate for PARL, the cleavage of this substrate being critical to cristae remodeling and cytochrome c release during apoptosis (88). In Toxoplasma, TgROM5, one of five nonmitochondrial Rhomboids in these parasites, cleaves a cell surface adhesion protein as a key step in cell invasion; similar findings in the related Plasmodium falciparum, the malarial parasite, recently have been reported (89), which suggests that Rhomboids are potential targets for treating infections by these deadly pathogens.

Most recently, the first crystal structures of an I-CLiP have been reported by three different research groups, all on the E. coli Rhomboid GlpG (90–92). These structures show remarkable similarities and important differences that provide insight into how this class of membrane-embedded protease carries out hydrolysis in the lipid bilayer. The structures all reveal that the key serine and histidine implicated as the catalytic dyad indeed are coordinated with each other and lie at a depth within the membrane consistent with where Rhomboids cleave their substrates.
transmembrane substrates (Fig. 10). A cavity is open to the periplasmic space with the catalytic dyad at the bottom of this opening, and this cavity contains multiple water molecules. How substrate enters this cavity is not entirely clear, but the position of the transmembrane domain 5 varies in the different structures and the movement of this domain can provide a space through which substrate may reach the catalytic dyad. Indeed, one reported structure contains a bound lipid in this space (S2) with the phosphate group residing near the Ser–His dyad and a key Asn residue that may contribute to the oxygen hole that stabilizes intermediates and transition-states during serine protease catalysis. These exciting structural findings validate the molecular and biochemical studies on Rhomboids and suggest that such approaches have been providing true mechanistic insight into the workings of other I-CLiPs. Moreover, these structures offer details that inspire specific hypotheses about how Rhomboids handle substrates to hydrolyze transmembrane domains.

Perspective

I-CLiPs are membrane-embedded enzymes that hydrolyze transmembrane substrates, and the residues essential to catalysis reside within the boundaries of the lipid bilayer. These enzymes seem to recapitulate the mechanisms of soluble proteases, and the first crystal structures of an E. coli Rhomboid support this notion, at least for the serine I-CLiPs. All I-CLiPs would be predicted to contain an initial substrate docking site, but to date evidence for such a docking site has been provided only for γ-secretase. The I-CLiPs discovered so far each play critical roles in biology and are regulated closely, but the means of control vary. They all are involved in cell signaling but do so in a variety of ways. Membrane topology seems to dictate the types of substrates that can be cleaved, but this concept remains speculative. Most I-CLiPs seem to require helix-breaking residues near the cleavage sites of their substrates, although γ-secretase may be a notable exception.

Critical remaining issues include the identity of substrates for the I-CLiP family members whose roles are unknown. For instance, although an entire family of PSHs and Rhomboids may be a notable exception.

Acknowledgment

I would like to acknowledge a postdoctoral fellow in my lab, Dr. Raquel Lieberman, for rendering details that inspire specific hypotheses about how Rhomboids handle substrates to hydrolyze transmembrane domains.

References


Intramembrane Proteolysis


See Also

Peptidomimetics
Peptidomimetics: Protease Pathways, Small Molecules to Elucidate Proteolysis and Disease, Chemical Biology of Signal Transduction Across Membranes
Metalloproteinases, Biophysics and Chemistry of

Ariel Solomon* and Tzvia Selzer*, The Weizmann Institute of Science, Israel
Marcos E. Milla, Group Leader, Department of Biochemical Pharmacology, Roche, Palo Alto, California
Irit Sagi, Associate Professor, Department of Structural Biology, The Weizmann Institute of Science, Israel
doi: 10.1002/9780470048672.wecb329

Zinc-dependent metalloproteases constitute a large family of enzymes, part of the protease superfamily. Fundamental to the structural integrity and catalytic activity of metalloproteases is the presence of both zinc and calcium ions in the structure of the protein. This article will focus on matrix metalloproteinases that are involved in extracellular matrix (ECM) catabolism and serve as important enzymes in many aspects of biology, ranging from cell proliferation, differentiation, and proliferation to cancer, tumor metastasis, inflammation, and other pathologic states. Despite their key role in many normal and pathologic processes, the molecular mechanisms by which zinc-dependent proteases hydrolyze their physiologic substrates are only known partly. Recent theoretical analyses have suggested reaction models for which limited and controversial experimental evidence exists. Here we will discuss the importance of quantifying the biophysical properties and the structural dynamic behavior of these enzymes to reveal their underlying molecular mechanisms. Such molecular knowledge holds promise in providing the basis for the novel design of specific antagonists as drug candidates for these important enzymes. In addition, we will discuss the use of real-time spectroscopic tools for studying the reactive metal sites in these enzymes.

Introduction

Metalloproteinases catalyze the hydrolysis of the peptide bond, which is the most stable chemical bond in nature (1, 2). Thermo-dynamic analysis of its hydrolysis reveals that although the free energy is relatively low, ∼2.4 kcal/mol, and the spontaneous reaction is exothermic, the scissile bond cleavage is hindered by a high-activation energy barrier of ∼20 kcal/mol (3). Thus, it may have taken 400 years to achieve the spontaneous bond cleavage in the absence of enzymes. Remarkably, metalloproteinases accelerate peptide bond hydrolysis by a factor of 10^16, which enables this process to be completed within milliseconds (4).

The zinc-dependent metalloproteases comprise a large family of enzymes with a wide variety of biologic roles. For example, important metalloproteases for cell viability include the bacterial metalloendopeptidase thermolysin, the digestive exopeptidases, carboxypeptidase A or B, and the matrix metalloproteinases (MMPs) (Table 1). MMPs are members of the metzincin superfamily of greater than 770 zinc endopeptidases, which includes astacins, serinelysins, adamalysins, leishmanolysins, and snapalysins. Metzincins are characterized by an absolutely conserved methionine residue C-terminal to the third histidine in the consensus sequence HEXXHXXGXXH/D, where the histidine residues chelate a catalytic zinc ion. Because of the extensive experimental and theoretical literature available, we will deliberately limit our discussion to MMPs that act within the extracellular matrix (ECM) milieu. A ample evidence
Table 1. Biologic roles of the zinc-dependent proteinases categorized into four subgroups based on the conserved structural motif of their active site (5)

<table>
<thead>
<tr>
<th>Protein family</th>
<th>Biologic distribution</th>
<th>Main biologic function</th>
<th>Preferred Substrate (cleavage site)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermolysins (Thermostable neutral proteinase)</td>
<td>Bacillus caldolyticus, sp., steatothermophilus, thermoproteolyticus</td>
<td>Extracellular, thermostable, digestive bacterial proteases. Hydrolyze the N-terminal side of hydrophobic residues.</td>
<td>Xaa—Leu &gt; Xaa—Phe</td>
<td>53–56</td>
</tr>
<tr>
<td>Carboxypeptidases</td>
<td>Range from Yeast to Homo sapiens</td>
<td>Degradation of a wide array of proteins. The well-studied pancreatic enzymes (carboxypeptidases A1, A2, and B) are involved in the digestion of food. Several members of the metallocarboxypeptidase gene family (carboxypeptidases D, E, M, and N) are more selective enzymes and are thought to play a role in the processing of intercellular peptide messengers.</td>
<td>Carboxyl terminal site</td>
<td>37, 38, 57, 58</td>
</tr>
<tr>
<td>Angiotensin-converting enzymes (ACEs)</td>
<td>Eukaryotes</td>
<td>ACEs are involved in blood chemistry. They mostly exist at the cell surface as ectoenzymes, where they hydrolyze circulating peptides. Gene targeting studies in mice have established that the tissue-bound form of ACE controls both blood pressure and renal structure and function, although a soluble form of ACE, which is derived from the membrane form through the action of a secretase, is also present in serum and other body fluids.</td>
<td>Nonselective endopeptidase</td>
<td>59, 60</td>
</tr>
<tr>
<td>Protein family</td>
<td>Biologic distribution</td>
<td>Main biologic function</td>
<td>Preferred Substrate (cleavage site)</td>
<td>References</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------</td>
<td>------------------------</td>
<td>------------------------------------</td>
<td>------------</td>
</tr>
</tbody>
</table>

Astacins, bacterial serralysins, matrix metalloproteinases (MMPs), ADAMs (a disintegrating and metalloproteinase), ADAMTS (a disintegrating and metalloproteinase with thrombospondin motifs). MMPs are involved in the catabolism of extracellular matrix (ECM) molecules and the activation of other MMPs; they also promote several cell surfaces and serve as soluble regulators of cell behavior in bioactive molecules. The ADAMs are major ectodomain sheddases; they release a variety of cell surface proteins, including growth factors, cytokines, cell adhesion molecules, and receptors. ADAMs can also cleave and remodel components of the extracellular matrix.
exists on the role of MMPs in normal and pathologic processes, including embryogenesis, wound healing, inflammation, arthritis, connective tissue diseases, inflammation, cardiovascular and autoimmune diseases, and cancer (6-14). The MMP family consists of more than 25 enzymes, with differences in substrate preference, domain structure, and sequence homology. As depicted in Fig. 1, MMPs are multidomain proteins with a signal peptide, a "pro" domain that maintains enzyme latency until it is removed or disrupted, and a catalytic domain that is common to the entire family. Additional domains observed in different MMP structures include fibronectin-type III-like, hemopexin-like, and transmembrane domains (15, 16). MMPs catalyze the hydrolysis of peptide bonds of a large range of biologic substrates, including collagen, gelatin, fibronectin, elastin, growth factors, cytokines, and chemokines. Notably, the broad association between MMP peptide hydrolysis activity and several serious diseases has made MMPs an attractive target for developing novel drugs aimed at inhibiting MMP activity (17).

In MMPs, the hydrolysis of a peptide bond is mediated by a catalytic zinc ion that resides in a structurally conserved catalytic cleft of the enzyme. Zinc is the second-most abundant transition metal in biology, after iron. It plays structural, chemical, and regulatory roles in biologic systems and is an essential ingredient at the active site in many enzymes (18, 19). Zinc plays a role in gene expression, stabilizes the structure of proteins and nucleic acids, preserves the integrity of subcellular organelles, participates in the transport process, and plays important roles in viral and immune phenomena (20). Moreover, zinc exhibits flexible coordination geometry and facilitates fast ligand exchange, and it is a Lewis acid with intermediate polarizability lacking redox activity (5). All the above characteristics give zinc its versatility, which enables it to form different types of chemical and bonding interactions. The families, motifs, enzymology, and protein structures containing one or two zinc ions have been discussed (5, 18, 21).

The structural and chemical knowledge regarding enzymatic peptide cleavage in zinc-dependent metalloproteinases is limited, and experimental evidence is controversial, partly because the zinc atom is spectroscopically silent and hence difficult to study using conventional spectroscopic and analytical tools (20). In addition, most structural and biochemical studies carried out so far were limited to nondynamic structure/function characterization, thus preventing qualitative and quantitative analysis of the evolving intermediates formed during peptide cleavage mediated at the catalytic zinc-binding site. Most of our mechanistic knowledge comes from theoretical calculations derived from crystal structures. Unfortunately, this process results in controversial proposals as to how proteases hydrolyze their peptide substrates. Thus, in the absence of direct experimental tools that would enable us to characterize the reaction in real time, it will be difficult, not to say impossible, to elucidate the underlying reaction mechanisms.

Here we will discuss the progress achieved so far in our mechanistic understanding of peptide hydrolysis reactions by metalloproteinases. We will focus on the need for comparative structure-function analyses of individual metalloproteinases to reveal the reaction mechanisms. Finally, we will describe new frontiers in the biophysics and chemistry of metalloproteinases and the application of such studies to reveal directly reaction mechanisms and their relevance to drug discovery.

**Structures of Available Matrix Metalloproteinases**

To date, the crystal structures of 12 different MMPs have been solved. Full structures were obtained for MMP-1 (2CLT), MMP-2 (1CK7), and MMP-7 (1MMP). As for the rest, only the catalytic domains in the presence of different inhibitors were determined. The hemopexin-like domains of MMP-2 (1RTG), MMP-9 (1TV), and MMP-13 (1PEX) were crystallized, and structures were determined separately. Nuclear magnetic resonance (NMR) structures of the catalytic domains of MMP-1 (1LYK), MMP-2 (1HOV), MMP-3 (1UMS), MMP-12 (1YCM), and MMP-13 (1UEB) have also become available.

The overall structures of all MMP catalytic domains known so far are very similar (Fig. 2). These MMP catalytic domains are shaped like an oblate ellipsoid, with a small active-site cleft, harboring the catalytic zinc ion, notched into the flat ellipsoid surface. The active site cleft is defined by helix Hb, which provides two histidine residues that coordinate to the catalytic zinc ion, and the catalytic Glu in between, all belonging to the zinc-binding consensus sequence HEXXH (GXXH) (5, 22, 23). The active-site helix ends at a Gly residue, where the peptide chain bends, presenting the third zinc-liganding His. The zinc ion also coordinates to a water molecule that...
Metalloproteinases, Biophysics and Chemistry of

**Figure 2** Ribbon structure of the MMP catalytic domain. The catalytic domain of MMP-8 (1ZP5) is superimposed with the catalytic domains of MMP-3 (green) (1HY7), MMP-12 (yellow) (1UTT), MT1-MMP (orange) (1AQ0), and MT3-MMP (pink) (1RM8); only the active site-conserved motif is shown for clarity. The catalytic and structural zinc (center and top) and the two calcium ions are displayed as red and blue spheres, respectively.

is used to hydrolyze the peptide bond of the substrate (see the discussion in the next chapter). The water molecule is also held in place by the side chain of the active site Glu. Another basic feature of the MMP active site is the presence of three substrate-binding subsites. The surface of the protease that can accommodate a single side chain of a substrate residue is called the subsite. Subsites are numbered S1–Sn upward toward the N terminus of the substrate (nonprimed sites) and S1’–Sn’ toward the C terminus (primed sites), beginning from the sites on each side of the scissile bond. These subsites accommodate the side chains of the peptide to be cleaved, and the local structural characteristics and electrostatic environment of the individual subsites effectively determine the specificity of the substrate.

A detailed description of the fine structural differences between different MMPs can be found in recent reviews and publications (12, 15, 24). More relevant to the current article are the insights gained from integrating structural data obtained by different methods, more specifically by comparing X-ray structures (possibly more than one) with the solution structures obtained from NMR, to obtain a structural model that is beyond a single structural datum. Examples can be found in recent studies done by Rush et al. on MMP-1 and MMP-13 (25) and Bertini et al. on MMP-12 (26); both looked into the structure of these different MMPs complexed with various inhibitors by using both NMR and X-ray techniques. The backbone generalized order parameter (S2), a parameter related to the amplitude of fast (picosecond to nanosecond) movements of NH vectors, peak multiplicity, and weak or missing peaks, suggested active site mobility of inhibitor-free MMP-1 when compared with the structure of the inhibited enzyme. The data indicated that a slow conformational change in the active site results in a concerted motion of helix hB and the zinc-ligated histidines. Furthermore, the presence of an inhibitor that binds by chelating zinc effectively removes this motion, while maximizing the interaction of the inhibitor with the enzyme. The mobility of the random coil region in the vicinity of the active site was maintained even in the presence of a bound inhibitor. On the other hand, a comparison of the crystal structures of three MMP-12-inhibitor complexes indicated that the conformational heterogeneity observed is largely independent of the type of inhibitor. These studies have shown that flexibility/conformational heterogeneity in crucial parts of the catalytic domain is the rule rather than an exception in MMPs, and its extent may be underestimated by inspection of one X-ray structure.

**Reaction Mechanisms of Metalloproteinase-Mediated Peptide Hydrolysis**

All proteinases achieve catalysis by providing a favorable electrostatic environment where the chemical reaction occurs. The bottleneck is the formation of the nucleophile, which in turn, attacks the carbonyl of the peptide that is properly located at the binding site and thus initiates bond cleavage (Scheme 1). A II well-characterized proteinases are categorized, based on the nature of the most prominent functional group at the active site, into one of four families: serine, cysteine, aspartic, and metalloproteinase. Cysteine proteinases use a low pKa cysteine as a reactive nucleophile (27); serine proteases use the serine hydroxyl as the catalyst in conjugation with a hydrogen bond network that allows the general base catalysis (28, 29); a catalytically active diad of aspartates facilitates hydrolysis in aspartic proteinases, and metalloproteinases use zinc to promote catalysis.

The mechanism by which metalloproteinases execute catalysis has been of interest for many years. Most studies focused on carboxypeptidase A and thermolysin-like proteases for which extensive structural, chemical, and biochemical data are available. The first peptide hydrolysis mechanisms to be proposed...
were based on the available X-ray structures that have facilitated establishing structural models for the different reaction steps (30). Crystallographic analyses of enzyme-inhibitor complexes were used to reveal how enzymes interact with their substrates. Suitably chosen inhibitors were used to provide structural models for various stages in catalysis, including the Michaelis complex, transition states, and products. A classical example of a complex between an enzyme and a substrate analogue was described by Shomah et al. (31). In this study the X-ray crystal structure of the complex between carboxypeptidase A (CPA) and 5-aminoo-(4-butyrylcarboxy)l-2-benzyl-4-oxo-6-phenyl-hexanoyl (BBP), the ketomethyl enzyme substrate analogue of the peptide substrate N-[(4-butyrylcarboxy)-6-phenylalanine] in peptide hydrolysis (32, 33). A second alternative pathway for peptide cleavage by carboxypeptidase-A proposed by Lipscomb and MocK by performing semiempirical theoretical calculations (43). The proton transfer step to the nitrogen of the peptide, following the nucleophilic attack of the peptide carbonyl group by a hydroxide, was calculated to be rate limiting. It was shown that under kinetic control both reactions are feasible; however, a calculated thermodynamic enthalpy difference of ~20 kcal/mol indicated that the reaction path suggested by Lipscomb is more stable than the other.

Biochemical studies and enzyme modification work, including mutagenesis of active site residues, zinc ion substitution, steady-state kinetics with different types of substrates, O18 exchange studies and others, together with the X-ray data, led to the proposition of two major types of hydrolysis mechanisms. The first mechanism to be proposed soon after the first crystal structure of carboxypeptidase A was published was the direct nucleophilic attack (also called the "Zinc-carbonyl mechanism" or "acyl pathway") on the peptide carbonyl by the conserved active site Glu resulting in the formation of an anhydride intermediate. A cumulative evidencing from O18 exchange studies during the late 1960s and 1970s have ruled out the involvement of an anhydride intermediate in peptide hydrolysis (32, 33). A second type of mechanism proposed is the general-acid-general-base (GAGB) mechanism (the "Zinc-hydroxide mechanism" or the "promoted water pathway") in which a water molecule initially attacks the carbonyl while, or after, losing a proton. This mechanism results in the formation of a gem-diol intermediate (34-36). Alternative paths for the GAGB mechanism were proposed and can be categorized into two groups. One alternative, following Christianson and Lipscomb (37), suggests that the substrate binds directly to the zinc ion while not replacing the water molecule that occupied the fourth coordination site of the zinc in the native structure. This water molecule is activated by the metal ion or by the conserved active site Glu (or both); it loses a proton and attacks the peptide carbonyl. According to this mechanism, the zinc ion has two roles: polarizing the carbonyl group of the substrate and facilitating the deprotonation of the water nucleophile. The second GAGB mechanism, following MocK and Zhang (38), suggests that the substrate carbonyl binds to the zinc ion and is activated by a nucleophilic attack initiated by a water molecule that is deprotonated by the C-terminal carboxylate of the substrate itself, not by the conserved binding site Glu or by the zinc ion. In the latter, the role of zinc is minimized to the polarization of the carbonyl group of the substrate only, and the conserved active site Glu is not given any active role. Mock and Zhang's (38) and MocK and Tsay's (39) proposed mechanism was based on enzyme kinetic experiments that showed significant pH dependencies of the inhibition constants of transition-state analog inhibitors. However, several studies implied the importance of the active site Glu residue in achieving the enzymatic activity. Mutagenesis studies show that all of the active site residues, only the modification of active site Glu residue completely abolishes catalysis (40, 41).

As observed in other systems, the obvious difficulty in elucidating reaction mechanisms based on static structural snapshots subsequently initiated structural-dynamic theoretical studies of metalloproteases. The active site chemistry of zinc-dependent enzymes has been studied using a variety of theoretical approaches. For example, mixed quantum mechanical/molecular calculations and classical molecular dynamics simulations have been employed, especially studies using density functional methods on redox-active metal centers (42).

Because of the availability of detailed structural information, the majority of theoretical studies focused on carboxypeptidase A and thermolysin-like proteases, and only one study investigated the mechanism of peptide hydrolysis by human MMP-3. Recently, Klithain-Vardi et al. compared the two GAGB alternative pathways for peptide cleavage by carboxypeptidase-A proposed by Lipscomb and MocK by performing semiempirical theoretical calculations (43). The proton transfer step to the nitrogen of the peptide, following the nucleophilic attack of the peptide carbonyl group by a hydroxide, was calculated to be rate limiting. It was shown that under kinetic control both reactions are feasible; however, a calculated thermodynamic enthalpy difference of ~20 kcal/mol indicated that the reaction path suggested by Lipscomb is more stable than the other.

Pelmenschikov and Siegbahn investigated the mechanism underlying peptide hydrolysis by human MMP-3 via quantum chemical methods (3) using the crystal structure of the inhibited enzyme with the transition state analogue pipieridin sulfonamide inhibitor as the structural reference. The importance of the weakly bound water molecule as a potent electrophile for the zinc-coordinated substrate oxygen was revealed by reducing the activation barrier by about 5 kcal/mol. Furthermore, the conserved active site Glu residue was confirmed to play a key role by acting as a base during the reaction water deprotonation. Interestingly, the zinc ion was shown to retain pentacoordinated geometry during the reaction, with distorted trigonal bipyramidal coordination, whereas a tetrahedral coordination sphere was suggested for the final product (Fig. 3). Remarkably, the formation of a pentacoordinated zink-protein complex is followed by distinct electronic transitions mediated by water and the conserved Glu residue. Essentially a single-step reaction mechanism has been obtained with an energy barrier of 13.1 kcal/mol. This work provided novel structural-dynamic insights into the reaction mechanism governing peptide hydrolysis. Naturally, such a theoretical model must be confirmed by experimental results, especially when the use of long-lived inhibitor-enzyme complexes (instead of enzyme-substrate complexes) may provide only structural models for different stages.
in catalysis. Such models are obviously not true intermediates; hence, in the absence of supporting experimental data, the obtained results must be interpreted with caution.

After more than 40 years of research on the structure/function of metalloproteases, tremendous advances in our understanding have been achieved. A consensus exists regarding the basics of peptide hydrolysis: 1) A water-formed hydroxyl nucleophilically attacks the peptide backbone carbonyl, 2) a carbonyl oxyanion is formed and coordinates the catalytic zinc ion, 3) the backbone amide is protonated, and 4) the zinc ion reduces the activation energy of this reaction by polarizing the oxygen carbonyl and by coordinating the negative charge of the intermediates evolving throughout the reaction (44). However, our understanding is far from complete, and several major questions remain unanswered: Is the catalytic water molecule bound to the zinc ion before proton abstraction by the conserved active site Glu base catalyst? Is the water-bound molecule pushed and replaced by the carbonyl oxygen of the peptide, or is a penta-coordinative complex consisting of both water and peptide formed? Finally, although it is thought that protein function depends on protein flexibility, precisely how the molecule dynamics contribute to the catalytic mechanism remains unclear. Although internal protein dynamics are connected intimately to enzymatic catalysis, enzyme motions linked to substrate turnover remain largely unknown. Because of the scarcity of adequate experimental tools, the field is presented with a great challenge regarding quantification of protein conformational transitions during catalysis. In the next section we discuss the use of a real-time multidisciplinary structural-spectroscopic approach to study reactive sites in metalloproteases during catalysis.

**Time-Resolved, Structural Analysis—Correlation of Reaction Kinetics–Protein Conformations and Evolution of Reaction Intermediates**

Enzymes are flexible molecules whose structures exhibit dynamic fluctuations on a wide range of timescales. This inherent mobility of a protein fold was shown to be manifested in the various steps constituting the catalytic cycle. The nature of this linkage between protein structure movement and function undoubtedly is complex and might involve the formation of a coupled network of interactions that bring the substrate closer, orient it properly, and provide a favorable electrostatic environment in which the chemical reaction can occur (45). However, the molecular details that link the catalytic chemistry to key kinetic, electronic, and structural events have remained elusive because of the difficulties associated with probing time-dependent, structure-function aspects of enzymatic reactions.
Figure 4  A schematic representation of the experimental approach for time-resolved XAS measurements. XAS provides local structural and electronic information about the nearest coordination environment surrounding the catalytic metal ion within the active site of a metalloprotein in solution. Spectral analysis of the various spectral regions yields complementary electronic and structural information, which allows the determination of the oxidation state of the X-ray absorbing metal atom and precise determination of distances between the absorbing metal atom and the protein atoms that surround it. Time-dependent XAS provides insight into the lifetimes and local atomic structures of metal–protein complexes during enzymatic reactions on millisecond to minute time scales. (a) The drawing describes a conventional stopped-flow machine that is used to rapidly mix the reaction components (e.g., enzyme and substrate) and derive kinetic traces as shown in (b). (b) The enzymatic reaction is studied by pre-steady-state kinetic analysis to dissect out the time frame of individual kinetic phases. (c) The stopped-flow apparatus is equipped with a freeze-quench device. Sample aliquots are collected after mixing and rapidly freeze into X-ray sample holders by the freeze-quench device. (d) Frozen samples are subjected to X-ray data collection and analysis.

By using this argument, a single crystal structure generally is insufficient to enable the elucidation of enzymatic catalysis reaction mechanisms at an atomic level of detail. Typically, the catalytic cycle involves a series of intermediates and transition states, and for many of these states, no detailed structural information is available. Furthermore, determining the energies of the various stationary points in the cycle is highly nontrivial, from a theoretical or experimental point of view. For these reasons, as of today, a complete characterization of reactive enzymatic chemistry is unavailable.

Concentrating on metalloenzymes, we have developed a strategy based on stopped flow X-ray absorption spectroscopy (XAS) to elucidate in detail the molecular mechanisms at work during substrate turnover (Fig. 4). Importantly, XAS provides local structural and electronic information about the nearest coordination environment surrounding the catalytic metal ion within the active site of a metalloprotein in solution. When the X-rays hit a sample, the electromagnetic radiation interacts with the electrons bound in the metal atom. The radiation can be scattered by these electrons, or it can be absorbed, thereby exciting the electrons. At certain energies, the absorption increases drastically and gives rise to an absorption edge. Such edges occur when the energy of the X-ray beam is just sufficient to cause excitation of a core electron of the absorbing atom (in this case zinc) to a continuum state, for example, to produce a photoelectron. The energies of the absorbed radiation at these edges correspond to the binding energies of electrons in the K (or L) shell of the zinc ion. When the photoelectron leaves the absorbing atom, its wave is backscattered by the neighboring atoms (in this case, the zinc-bound protein atoms). The constructive and destructive interference of these outgoing photoelectrons with the scattered waves from atoms surrounding the central metal atom gives rise to the extended X-ray absorption fine structure (EXAFS) oscillation pattern. Spectral analysis of the edge and EXAFS regions yields complementary electronic and structural information. Analysis of the edge region enables us to determine the oxidation state of the X-ray absorbing metal atom (in other words, the position of the absorption edge), whereas analysis of the EXAFS region provides precise information regarding distances between the absorbing metal atoms and the...
The design and production of highly selective inhibitors for this highly structurally homologous catalytic sites residing in these enzymes may aid in the process of discovering new drugs. One may therefore inquire as to whether structurally homologous enzymes display similar dynamic profiles during binding and catalysis or whether functionally related enzymes that share similar structures can be distinguished based on their dynamic nature.

So far only a few reports have indicated distinct differences regarding the reaction chemistry and mechanisms among highly structural homologous MMP active sites. Studies by Fasciaglione et al. demonstrate that the protein-linked behavior (in different pH environments) for $k_{cat}$, $K_m$, and $K_i$ is different among various MMPs, including MMP-9 and MMP-2. In addition, the authors found that to have a fully consistent description of the enzymatic action of the various MMPs, they had to apply three protonating groups that are involved in the modulations of substrate interaction and catalysis. These results indicate that the details regarding active site modifications at the zinc ion are different for various MMPs and are of enormous importance for elucidating the mode of action of individual MMPs. Similar conclusions may be drawn from the recent study of Solomon et al. (52), which compares the inhibition mode of the mechanism-based inhibitor SB-3CT with the highly homologous catalytic sites of TACE and MMP-2. Importantly, it was shown that the zinc ion of TACE is different regarding the length of the $Zn$-$S(SB-3CT)$ bond distance and the total effective charge of the catalytic zinc ion. In addition, SB-3CT inhibits TACE in a noncompetitive fashion by inducing marked conformational changes in the structure. For MMP-2, SB-3CT behaves as a competitive inhibitor, and no significant conformational changes are observed. Examination of the second shell amino acids surrounding the catalytic zinc ion of these enzymes indicated that the active site of TACE is more polar than that of MMP-2 and other MMPs. On the basis of these results, it was proposed that, although a seemingly high structural similarity exists between TACE and MMP-2, these enzymes are significantly and other MMPs. On the basis of these results, it was proposed that, although a seemingly high structural similarity exists between TACE and MMP-2, these enzymes are significantly diverse in the electronic and chemical properties within their active sites.

Comparative Structural-Dynamic Analysis and Its Relevance to Drug Discovery

Conducting real-time structural analysis on metalloproteinases was found to be a novel and effective approach for studying the highly structurally homologous catalytic sites residing in these enzymes via comparative structural–dynamic analysis. Remarkably, such high structural homology (Fig 2.) often hinders the design and production of highly selective inhibitors for this enzyme family. Thus, linking distinct protein conformational transitions with catalysis of individual enzymes is of great importance for revealing the molecular mechanisms underlying individual enzymes that may aid in the process of discovering new drugs. One may therefore inquire as to whether structurally homologous enzymes display similar dynamic profiles during binding and catalysis or whether functionally related enzymes that share similar structures can be distinguished based on their dynamic nature.

Overall, these results raise the possibility that the highly structural analogous catalytic centers in MMPs exhibit different structural conformations and electronic behavior during catalysis and inhibition. Yet, an open question is as follows: How can such differences be quantified? Comparative TRXAS analysis might provide the answer. By performing a comparative TRXAS analysis of functionally related enzymes that share similar structures, it might be possible to identify the intermediate states, the active site key amino acid residues, and the thermodynamic parameters of individual enzymes that are critical for efficient catalysis as well as for inhibition. Furthermore, the mobility so fundamental to protein structures is a major complicating factor to the general structure-based drug design approach. MMP’s active site dynamics were observed in the NMR structures of inhibitor-free MMP-1, MMP-3, and...
MMP-13 (25). The impact on drug design was clearly illustrated in MMP X-ray structures that demonstrate the ability of side chains in the active site to undergo conformational changes that accommodate a bound inhibitor. Thus, an inhibitor predicted to have poor inhibition activity against a specific MMP based on a poor fit in the specificity pocket (S1) may be accommodated in the binding site because of the mobility of the protein. By applying the time-resolved structural–dynamic approach to study catalysis as well as inhibition pathways, we can characterize unique reaction intermediates at the active site, as well as distinct conformational intermediates of the protein side chains, and can provide critical information about protein flexibility that additionally might be used for the design of potent and selective inhibitors. This approach departs from traditional drug design strategies used for metalloenzymes that target the catalytic metal with potent zinc-chelating peptidomimetic compounds.

Concluding Remarks

Here we reviewed the advances that have been made in the field of metalloproteinase chemistry over the past decades. Focusing on the critical need to understand better the chemical workings of these biologic machines at atomic detail, we noted the importance of quantifying the biophysical properties and the structural dynamic behavior of these enzymes to reveal their underlying molecular mechanisms. A fundamental challenge for better understanding the functioning of proteins/enzymes is to characterize proteins as dynamic objects. In this article, we have presented novel experimental approaches that go beyond static structures, with the ultimate goal of characterizing macromolecules reacting at atomic resolution. More specifically, we have combined time-resolved X-ray spectroscopy with X-ray crystallography, enzymology, and computational mechanistic investigations to determine the structures and chemistry of transient intermediates as they evolve during the catalytic cycle. Our recent progress in the field of MMP reaction mechanisms demonstrates how a combined approach that uses a variety of biophysical techniques advances our fundamental understanding of complex biologic molecules.

References


Protease Inhibitors, Mechanisms of
Christopher J. Farady, University of California, San Francisco
Charles S. Craik, University of California, San Francisco
doi: 10.1002/9780470048672.wecb474

Relatively few design principles underlie the mechanisms of inhibition of a myriad range of protease inhibitors. Protease inhibitors tend to be competitive and to compete with substrate binding, either through direct competition or deformation of the protease active site. Although protein inhibitors can gain potency through the burial of a large surface area and specificity through contacts with specific exosites, small-molecule inhibitors primarily gain potency through interactions with the catalytic machinery of the enzyme and specificity through interactions with the substrate binding sites. Incorporation of these design principles into chemical probes and drugs have improved greatly our ability to create potent and specific protease inhibitors.

Proteolytic enzymes are ubiquitous in all organisms and constitute 2–4% of the encoded gene products. They are critical for diverse biologic processes such as digestion, blood clotting, host defense, pathogenic infection, viral replication, wound healing, and disease progression, to name a few. Because proteases trigger an irreversible event—the cleavage of a protein—their activity must be controlled tightly. Dysregulated proteolytic activity causes a disruption in the homeostatic balance of a biologic system and can result in any number of poor biologic outcomes. As a result, nature has developed several strategies for inhibiting proteases to control proteolysis. Similar approaches have often been employed in the development of synthetic protease inhibitors. To a large extent, the same design principles that work well for naturally occurring protease inhibitors work well for inhibitors developed in the laboratory.

This review aims to survey the mechanisms by which protease inhibitors function. To achieve this goal, we have divided inhibitors into categories based on their mechanism, to illustrate that a relatively small number of design principles can be combined to develop new and effective protease inhibitors. These divisions are somewhat arbitrary, as many inhibitors could be grouped into several classes. Because of space limitations, the list of mechanisms is not exhaustive in its treatment of all inhibitors, but it aims to be illustrative of the many ways proteases can be inhibited. Although beyond the scope of this review, it is also important to keep in mind that spatial and temporal regulation of proteolytic activity is critically important in biology. Beyond the many levels of transcriptional and translational control, proteases are expressed as inactive (or nearly inactive)zymogens, and they are not activated until needed. Furthermore, they are often localized to cellular structures such as the cell membrane or stored in specific organelles such as lysosomes or granules to minimize unwanted proteolysis. Background on the four major classes of proteases (serine, cysteine, aspartic, and metalloproteases) as well as on the basic mechanisms of enzyme inhibition is abundant; please consult the “Further Reading” list at the end of the chapter. Figure 1 provides an overview of basic substrate and protease nomenclature, whereas Tables 1 and 2 list many of the inhibitors discussed in the text (1–21).

Mechanisms of Naturally Occurring Protein Protease Inhibitors

Competitive inhibition

Standard mechanism

The most thoroughly studied mechanism of protein protease inhibitors is that of the standard mechanism (or Canonical or Laskowski mechanism) inhibitors of serine proteases (1) (Fig. 2). Standard mechanism inhibitors include the K unit, Kunitz, and Bowman-Birk family of inhibitors and bind in a lock-and-key fashion. All standard mechanism inhibitors insert a reactive loop into the active site of the protease, which is complementary to the substrate specificity of the target protease and binds in an extended β-sheet with the enzyme in a substrate-like manner. While bound to the protease, the “scissile bond” of standard mechanism inhibitors is hydrolyzed very slowly, but products are not released and the amide bond is re-ligated. The standard mechanism is an efficient way to inhibit serine proteases, and it is thus used by many structurally

Advanced Article

Article Contents

• Mechanisms of Naturally Occurring Protein Protease Inhibitors
• Mechanisms of Small-Molecule Protease Inhibitors
• Inhibitors Developed via Protein Engineering
• Conclusions and Future Directions
Protease Inhibitors, Mechanisms of

Figure 1  Diagram of a protease active site. A protease cleaves a peptide at the scissile bond, and has a number of specificity subsites, which determine protease specificity. Substrates bind to a protease with their non-prime residues on the N-terminal side of the scissile bond and their prime-side residues C-terminal to the scissile bond. The catalytic residues determine the class of protease. Serine, cysteine, and threonine proteases hydrolyze a peptide bond via a covalent acyl-enzyme intermediate, and aspartic, glutamic and metalloproteases activate a water molecule to hydrolyze the peptide bond in a non-covalent manner.

Table 1  Naturally occurring protein protease inhibitors

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>Target protease</th>
<th>Mechanism</th>
<th>Specificity</th>
<th>$K_i$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPTI</td>
<td>S1 serine proteases</td>
<td>Standard M mechanism</td>
<td>broad specificity</td>
<td>femtomolar to low nanomolar</td>
<td>(1)</td>
</tr>
<tr>
<td>SMPI</td>
<td>M4 metalloproteases</td>
<td>Standard F mechanism</td>
<td>fold-specific</td>
<td>thermolysin -0.1 nM</td>
<td>(3)</td>
</tr>
<tr>
<td>Cystatin A</td>
<td>papain family cysteine</td>
<td>Competitive</td>
<td>specific</td>
<td>&lt;5 nM</td>
<td>(4)</td>
</tr>
<tr>
<td>TIMP 1</td>
<td>matrix metalloproteases</td>
<td>Competitive</td>
<td>broad specificity</td>
<td>picomolar to low nanomolar</td>
<td>(19)</td>
</tr>
<tr>
<td>Ascaris papain</td>
<td>aspartic proteases, cathepsin E, pepsin, gadiacin</td>
<td>Competitive</td>
<td>some specificity</td>
<td>1-100 nM</td>
<td>(7)</td>
</tr>
<tr>
<td>Inhibitor 3</td>
<td>S1 serine proteases</td>
<td>Competitive, Exosite binding</td>
<td>fold-specific</td>
<td>picomolar to low nanomolar</td>
<td>(44)</td>
</tr>
<tr>
<td>Hirudin</td>
<td>thrombin</td>
<td>Competitive, Exosite binding</td>
<td>specific</td>
<td>0.2 pM</td>
<td>(11)</td>
</tr>
<tr>
<td>XIAP-BIR3</td>
<td>caspase-9</td>
<td>Allosteric, Competitive</td>
<td>specific</td>
<td>20 nM</td>
<td>(14)</td>
</tr>
<tr>
<td>α-2-macroglobin</td>
<td>most proteases</td>
<td>Activity dependent</td>
<td>non-specific</td>
<td>N.D.</td>
<td>(18)</td>
</tr>
<tr>
<td>α-1-antiprotease</td>
<td>serine, occasionally</td>
<td>Activity dependent</td>
<td>broad specificity</td>
<td>elastase -6.5 x 10^7 M^-1s^-1</td>
<td>(17)</td>
</tr>
</tbody>
</table>

disparate protein scaffolds to create potent inhibitors. However, most standard mechanism protease inhibitors tend to have relatively broad specificity within subclasses of serine proteases. For example, the bovine pancreatic trypsin inhibitor (BPTI) efficiently inhibits almost all trypsin-fold serine proteases with P1-Arg specificity, and it can also inhibit chymotrypsin (Phe P1 specificity) with a $K_i$ of 10 nM (22).

Standard mechanism inhibitors are classified strictly as inhibitors of serine proteases. There have been reports of inhibitors of other classes of proteases that have similar mechanisms to those of standard mechanism inhibitors, though. Initial studies on the streptomyces metalloprotease inhibitor (SMPI) suggest that it inhibits the metalloprotease thermolysin through a substrate-like binding mechanism (2). Similarly, staphostatin B, a cysteine protease inhibitor from Staphylococcus aureus, binds in a substrate-like manner in the active site of diphosphopain cysteine proteases. However, staphostatin B has a glycine P1 residue, which adopts a backbone conformation that seems to prevent nucleophilic attack of the scissile bond (3).

Noncanonical competitive inhibitors

Several protease inhibitors bind in the active site of the protease, but they do not bind in a substrate-like manner, instead forming interactions with the catalytic residues in a noncatalytically competent manner, and thus, they are not considered standard mechanism inhibitors.

The cystatins, which are a superfamily of proteins that inhibit papain-like cysteine proteases, are a classic example of these inhibitors. The cystatins (Fig. 3) insert a wedge-like face of the inhibitor that consists of the protein N-terminus and two hairpin loops into the V-shaped active site of a cysteine protease. The N-terminal residues bind in the S3-S1 pockets in a substrate-like manner, but the peptide then turns away from the catalytic residues and out of the active site. The two hairpin loops bind to the prime side of the active site, which provides most of the binding energy for the interaction. Thus, both the prime and the nonprime sides of the active site are occupied, but no interactions are actually made with the catalytic machinery of the enzyme (23).

Tissue inhibitors of metalloproteases (TIMPs)

TIMPs inhibit matrix metalloproteases (MMPs) via a two-step mechanism in a manner somewhat similar to that of cystatins (Fig. 3). While the N-terminal residues of cystatins bind to the nonprime side of cysteine proteases, TIMPs N-termini bind in the P1-P3 pockets of the protease, coordinate the catalytic...
### Table 2: Small-molecule protease inhibitors

<table>
<thead>
<tr>
<th>Inhibitor</th>
<th>Target protease</th>
<th>Mechanism</th>
<th>Specificity</th>
<th>$K_i$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bortezomib</td>
<td>proteosome</td>
<td>competitive, transition state analog</td>
<td>specific proteosome</td>
<td>0.62 nM</td>
<td>(43)</td>
</tr>
<tr>
<td>Pepstatin A</td>
<td>aspartic proteases</td>
<td>competitive, transition state analog</td>
<td>class-specific</td>
<td>chymotrypsin</td>
<td>320 nM</td>
</tr>
<tr>
<td>Idinavir</td>
<td>HIV protease</td>
<td>competitive, transition state analog</td>
<td>specific HIV1 pr</td>
<td>&lt;0.35 nM</td>
<td>(42)</td>
</tr>
<tr>
<td>Leupeptin</td>
<td>cysteine, threonine proteases</td>
<td>competitive, transition state analog</td>
<td>P1 arginine specificity</td>
<td>trypsin</td>
<td>130 nM</td>
</tr>
<tr>
<td>E64</td>
<td>serine, cysteine proteases</td>
<td>irreversible, alkylation</td>
<td>C1 class-specific</td>
<td>cathepsin B</td>
<td>6 nM</td>
</tr>
<tr>
<td>Phosphonates</td>
<td>serine proteases</td>
<td>irreversible, phosphonylation</td>
<td>P1 residue specificity</td>
<td>thrombin</td>
<td>70 M</td>
</tr>
<tr>
<td>Cephalosporin</td>
<td>elastase</td>
<td>mechanism-based</td>
<td>some cross-reactivity</td>
<td>161,000 M</td>
<td>(48)</td>
</tr>
<tr>
<td>E64</td>
<td>matrix metalloproteases</td>
<td>competitive, active site</td>
<td>inhibits many MMPs</td>
<td>IC$_{50}$</td>
<td>20 nM</td>
</tr>
<tr>
<td>Captopril</td>
<td>angiotensin converting enzyme</td>
<td>competitive, active site</td>
<td>specific</td>
<td>ACE</td>
<td>23 nM (IC$_{50}$)</td>
</tr>
</tbody>
</table>

1. Inhibition constants for covalent inhibitors are given as $k_{	ext{inact}} / K_i$.

2. Values are for D-Phe-Pro-(4AngGly)P(OPh)$_2$NH$_2$.

Zn$^{2+}$ ion, and exclude a catalytic water molecule from the active site. Meanwhile a second loop of the TIMP binds in both the P3 and the P2 pockets, and it binds to the N-terminus of the MMP. Despite the similarities in mechanistic architecture between TIMPs and cystatins (hairpin loops and N-terminal residues in substrate binding pockets), TIMPs interfere with the catalytic machinery of MMPs by chelating the catalytic Zn$^{2+}$ (5).

The ascaris pepsin inhibitor-3 is an aspin, which is a family of inhibitors of aspartic proteases that protect worms from host gastric enzymes. Like the cystatins, the aspins are competitive inhibitors that bind in the substrate-binding subsites, but they do not have an amide bond that is available for nucleophilic attack. They gain most of their inhibitory activity by inserting their 3 N-terminal residues in the S1'–S3' subsites of the protease (6) (Fig. 3). Although cystatins and aspins do not interact directly with the catalytic residues of cysteine proteases, many protease inhibitors, such as cytotoxic T-lymphocyte antigen 2-α and the cathepsin propeptides, do interact with the catalytic machinery of the enzyme, but they do so in a proteolitically noncompetent manner. These inhibitors have long lengths of peptide that span the active site cleft in the reverse orientation (from C-terminus to N-terminus).
Competitive inhibition with exosite binding

Several protease inhibitors are competitive, and they bind in the protease active site, but also they have secondary binding sites outside the active site, which are critical to inhibition. Exosite binding provides two major benefits: 1) It increases the surface area of the interaction, which leads to a greater affinity, and 2) it can provide a greatly increased amount of specificity.

Ecotin is a dimeric serine protease inhibitor found in E. coli, which effectively inhibits many trypsin fold serine proteases, regardless of primary specificity, and is thought to protect E. coli from attack by host proteases. It inhibits serine proteases through a standard mechanism at a primary binding site, but it also has a secondary binding site that can contribute 5 kcal/mol of binding energy to the very tight enzyme-inhibitor complex (Fig. 4). Surprisingly, the individual binding energies of the two binding sites are not additive; the effect of the secondary binding site on affinity was found to be inversely proportional to the strength of binding at the primary site. The secondary binding site seems to provide compensatory effects that can overcome suboptimal binding at the primary binding site. If binding at the primary site is not optimal, the secondary binding interaction tends to be stronger. In this case, the secondary binding site actually makes the inhibitor less specific, or more capable of inhibiting a broad range of proteases, and it protects bacteria from several host proteases (25).

Many blood-meal parasites have developed specific inhibitors of clotting enzymes to prevent blood clotting of the host. These inhibitors often use mechanisms of inhibition described above, but they have domains that bind to protease exosites and provide...
a high degree of target specificity. Rhodnin, which is a thrombin inhibitor from the assassin bug Rhodnius prolixus, has two Kazal-type inhibitory domains and a common standard mechanism to inhibit serine protease inhibitor domains. Although the N-terminal domain binds and inhibits via the standard mechanism, the second Kazal-type domain has evolved to bind to exosite I on thrombin. The binding affinities of the individual domains are roughly additive, and the resultant inhibitor has a K<sub>i</sub> of 0.2 mM and exquisite specificity for thrombin (26). Hirudin, from Hirudo medicinalis and tick anti-coagulant peptide (TAP) from Ornithodoros moubata, specifically inhibit thrombin and factor Xα (Xkα), respectively (8, 27). They do so by similar mechanisms; they insert the N-terminal tail of the protein in the protease active site (analogous to apo-1 inhibition of pepstatin), whereas the body of the inhibitor binds to specific exosites, either exosite I on thrombin (Fig. 4) or the autolysis loop on FXa. Because the secondary binding sites are specific to each clotting factor, the inhibitors show a high degree of specificity.

Inhibitor of apoptosis (IAP) proteins inhibit caspases, which are dimeric cysteine proteases responsible for programmed cell death, or apoptosis. IAPs are multidomain proteins that have multiple BIR domain repeats. One family member, the X-linked IAP (XIAP), has three BIR domains, and uses different BIR domains to inhibit different caspases through disparate mechanisms. The XIAP-BIR2 domain and its N-terminal peptide extension are responsible for inhibition of the “executioner” caspases-3 and caspase-7. The N-terminal peptide extension binds in the active site in reverse orientation, which is similar to the inhibition mechanism of cathepsin proteases. Meanwhile, the BIR2 domain binds to an exosite on the caspase dimer (Fig. 4). The BIR2 domain needs both exosite-binding capability and the N-terminal extension to inhibit efficiently its target caspsases (28).

The XIAP-BIR3 domain is responsible for inhibition of the initiator caspase-9, but it functions via a completely different mechanism. The BIR3 domain is an allosteric inhibitor of caspase-9: it binds to the dimer interface and prevents dimerization and subsequent activation of the enzyme (9) (Fig. 4). Caspase-9 is at the apex of the apoptotic cascade that leads to the activation of executioner caspases. As such, BIR3 can provide an extra level of regulation by sequestering monomers in a catalytically inactive conformation and ensuring that no unwanted caspase-9 activity occurs.

Activity-dependent inhibition
Sometimes called suicide substrates, several protein inhibitors of proteases require proteolytic activity of the enzymes they inhibit, which leads to either covalent modification of the enzyme or release of charged groups that inhibit the catalytic machinery. In either case, this sort of activity-dependent inhibition is powerful and fundamentally different than the competitive mechanisms outlined above; the inhibitor acts as a substrate and then uses the enzymes’ catalytic machinery to trap and then inhibit the enzyme.

The potato metallocarboxypeptidase inhibitor (and metallo-carboxypeptidase inhibitors from leeches and ticks) inhibit carboxypeptidase B after a proteolytic processing event. These inhibitors bind their four C-terminal residues in the protease subunits S3–S1'. The C-terminal residue, Gly39, is processed, but it does not diffuse from the active site. In a type of product inhibition, Gly39 instead stays in the S1' pocket and chelates the catalytic Zn<sup>2+</sup> – which creates a protease-activated reversible inhibitor (29).

The inhibitor α-2-macroglobin and its relatives are responsible for clearing excess proteases from plasma. Less an inhibitor than a “protease sponge,” α2-M is a large protein, which is a tetramer of about 600 kD that has four ball loops on its surface. When a protease cleaves one of these reactive loops, it triggers a conformational change, and the protease becomes cross-linked to the inhibitor through surface lysines and arginines. The enzyme is still active; small-molecule substrates can still be hydrolyzed by proteases complexed with α2-M, but protein substrates are occluded from the active site and the complex is cleared quickly from the blood (30).

The serpins are a family of inhibitors that inhibit covalently and irreversibly primarily serine proteases (1) (the serpin CmL inhibits cysteine proteases). Serpins have a large reactive center loop (RCL) that is presented to a protease for proteolytic processing. During productive cleavage of the RCL, the N-terminal half of the RCL, which is still attached to the protease as an acyl–enzyme intermediate, is inserted into a β-sheet in the body of the inhibitor. The resulting free-energy change is enough to translocate the protease (which is still covalently attached to the RCL) to the distal side of the inhibitor, and the resulting steric collisions completely deform the protease active site, which thus leaves the protease tethered to the serpin and completely inactive (Fig. 5). The serpin inhibitory mechanism is completely irreversible. Because of the drastic nature and irreversibility of this mechanism, serpins function as protease scavengers, which protect cells and tissues from unwanted proteolytic activity. These types of inhibitors, which take advantage of the catalytic activity of a protease to trap and inhibit the enzyme, are effective and powerful inhibitors. As discussed in the following section, many small-molecule irreversible and protease-activated inhibitors have been developed that rely on the same fundamental mechanism of using enzyme activity to trap and inactivate a protease.

Mechanisms of Small-Molecule Protease Inhibitors
To develop both tools for chemical biology and possible drugs, significant effort has gone into the discovery and development of small-molecule protease inhibitors. The two critical components in the design of a protease inhibitor, and indeed any enzyme inhibitor, are potency and specificity. It has been challenging to create potent and specific protease inhibitors for clinical use because of the high degree of similarity among families of proteases, but even relatively nonspecific protease inhibitors have been invaluable in teasing apart the roles proteases play in complex biologic processes. A novel effective strategy for developing protease inhibitors has been to take peptide substrates that target the active site and to
Protease Inhibitors, Mechanisms of  

Serpin Inhibition of a Serine Protease

Figure 5  Serpins inhibit serine proteases by binding a reactive center loop in the active site, forming a covalent complex with the enzyme, undergoing a large conformational change, and irreversibly distorting the active site of the protease.

Figure 6  An effective strategy for developing synthetic protease inhibitors is to combine a peptide or peptidomimetic specificity element with a warhead that binds (either reversibly or irreversibly) to the catalytic machinery of a protease.

and have been developed into therapeutic agents; the proteosome inhibitor Bortezomib (Fig. 7) has been approved for the treatment of multiple myeloma (31).

Class-specific transition-state analogs have been developed to interfere specifically with the catalytic residues of each class of proteases. Aspartic protease inhibitors have long been designed around substrate polypeptides, with a replacement of the scissile amide bond with a noncleavable transition-state isostere. The first specific inhibitor for aspartic proteases, pepstatin A, was discovered from Actinomycetes, as an inhibitor for pepsin. It also showed strong inhibitory activity against several other aspartic proteases. Pepstatin A is a peptide, but the scissile bond is replaced with a statine group [(3 S,4 S)-4-amino-3-hydroxyl-6-methyl heptanoic acid)] (Fig. 7). Instead of a trigonal carbonyl, statines have a chiral hydroxyl group, which give it the ability to mimic the tetrahedral state of the substrate transition state (32). Other transition state isosteres, including homostatines, hydroxyethenies, diols, and phosphinates have been developed for aspartic proteases, and functionalizing these classes of inhibitors with peptidomimetic groups to bind to the substrate binding regions, has resulted in the development of numerous drugs, most notably HIV protease drugs (33) (Fig. 7).

Transition-state inhibitors

Transition-state inhibitors stably mimic the transition state of the enzymatic reaction, and thereby interact with the substrate-binding and catalytic machinery of the enzyme in a low-energy conformation. Transition-state analogs are competitive, reversible inhibitors, although some have extremely low Kᵢ's and very slow off-rates. All proteases activate a nucleophile to attack a carbonyl, which leads to the formation of a tetrahedral intermediate that then collapses to form the enzyme products—two peptides. Thus, synthetic small molecules that mimic the tetrahedral intermediate of the protease reaction are attractive transition-state analogs. A classic class of protease transition-state inhibitors uses a boronic acid scaffold (4, 10). Boronic acid adopts a stable tetrahedral conformation in the protease active site that is resistant to nucleophilic attack. Boronic acid inhibitors, which are derivatized with different specificity elements, have been developed against every class of protease and turn them into inhibitors by interfering with the catalytic machinery of the enzyme (Fig. 6). Although sometimes hampered by specificity problems, these types of inhibitors can be useful biologic probes, can help validate an enzyme as a drug target, and can act as lead compounds for additional drug development.

Transition-state inhibitors stably mimic the transition state of the enzymatic reaction, and thereby interact with the substrate-binding and catalytic machinery of the enzyme in a low-energy conformation. Transition-state analogs— inhibitors that stably adopt a conformation that mimics the transition state of an enzyme-substrate intermediate—have traditionally been an effective way to develop inhibitors. In addition, several irreversible inhibitors have been developed that bind covalently to cysteine, serine, and threonine proteases, which are proteases that form a covalent acyl–enzyme complex during peptide bond hydrolysis.

Transition-state inhibitors

Transition-state inhibitors stably mimic the transition state of the enzymatic reaction, and thereby interact with the substrate-binding and catalytic machinery of the enzyme in a low-energy conformation. Transition-state analogs are competitive, reversible inhibitors, although some have extremely low Kᵢ's and very slow off-rates. All proteases activate a nucleophile to attack a carbonyl, which leads to the formation of a tetrahedral intermediate that then collapses to form the enzyme products—two peptides. Thus, synthetic small molecules that mimic the tetrahedral intermediate of the protease reaction are attractive transition-state analogs. A classic class of protease transition-state inhibitors uses a boronic acid scaffold (4, 10). Boronic acid adopts a stable tetrahedral conformation in the protease active site that is resistant to nucleophilic attack. Boronic acid inhibitors, which are derivatized with different specificity elements, have been developed against every class of protease and turn them into inhibitors by interfering with the catalytic machinery of the enzyme (Fig. 6). Although sometimes hampered by specificity problems, these types of inhibitors can be useful biologic probes, can help validate an enzyme as a drug target, and can act as lead compounds for additional drug development.

Transition-state inhibitors

Transition-state inhibitors stably mimic the transition state of the enzymatic reaction, and thereby interact with the substrate-binding and catalytic machinery of the enzyme in a low-energy conformation. Transition-state analogs are competitive, reversible inhibitors, although some have extremely low Kᵢ's and very slow off-rates. All proteases activate a nucleophile to attack a carbonyl, which leads to the formation of a tetrahedral intermediate that then collapses to form the enzyme products—two peptides. Thus, synthetic small molecules that mimic the tetrahedral intermediate of the protease reaction are attractive transition-state analogs. A classic class of protease transition-state inhibitors uses a boronic acid scaffold (4, 10). Boronic acid adopts a stable tetrahedral conformation in the protease active site that is resistant to nucleophilic attack. Boronic acid inhibitors, which are derivatized with different specificity elements, have been developed against every class of protease and turn them into inhibitors by interfering with the catalytic machinery of the enzyme (Fig. 6). Although sometimes hampered by specificity problems, these types of inhibitors can be useful biologic probes, can help validate an enzyme as a drug target, and can act as lead compounds for additional drug development.

Transition-state inhibitors

Transition-state inhibitors stably mimic the transition state of the enzymatic reaction, and thereby interact with the substrate-binding and catalytic machinery of the enzyme in a low-energy conformation. Transition-state analogs are competitive, reversible inhibitors, although some have extremely low Kᵢ's and very slow off-rates. All proteases activate a nucleophile to attack a carbonyl, which leads to the formation of a tetrahedral intermediate that then collapses to form the enzyme products—two peptides. Thus, synthetic small molecules that mimic the tetrahedral intermediate of the protease reaction are attractive transition-state analogs. A classic class of protease transition-state inhibitors uses a boronic acid scaffold (4, 10). Boronic acid adopts a stable tetrahedral conformation in the protease active site that is resistant to nucleophilic attack. Boronic acid inhibitors, which are derivatized with different specificity elements, have been developed against every class of protease and turn them into inhibitors by interfering with the catalytic machinery of the enzyme (Fig. 6). Although sometimes hampered by specificity problems, these types of inhibitors can be useful biologic probes, can help validate an enzyme as a drug target, and can act as lead compounds for additional drug development.
Protease Inhibitors, Mechanisms of

Bortezomib

H
H
Pep
R
O

Enzyme

Aldehyde Transition State Inhibitor

Statine Group of Pepstatin

Hydroxyethylene Group of Idinavir

Figure 7  Various transition-state protease inhibitors. Bortezomib is an approved drug for the treatment of multiple myeloma. It is a boronic acid analog that inhibits the proteosome, a threonine protease. The boronic acid moiety can adopt a tetrahedral conformation in the active site. Pepstatin is a peptidyl aspartic acid inhibitor. The reactive statine group binds to the catalytic machinery, and the chiral hydroxyl group of the statine mimics the tetrahedral geometry of the transition state. Idinavir is an approved HIV 1 Protease inhibitor that binds to the active site via a hydroxyethylene transition state isostere. Aldehydes are also transition state analogs, which are susceptible to nucleophilic attack. In cysteine, serine and threonine proteases, this results in a covalent, reversible inhibition mechanism.

substrate analogs that form a covalent hemiacetal linkage (or hemithio acetal linkage in cysteine proteases) between the aldehyde of the inhibitor and the active site nucleophile of the protease. The resulting tetrahedral adduct mimics the transition state of the normal enzymatic reaction. The tetrahedral intermediate can collapse and regenerate a free inhibitor; therefore, the enzyme-inhibitor complex is in equilibrium with a free enzyme and a free inhibitor. Many aldehyde inhibitors, such as leupeptin, are nonspecific and inhibit many proteases, but medicinal chemistry efforts taking advantage of unique protease specificity elements have led to several very specific aldehyde protease inhibitors (10). Transition-state inhibitors, especially those with peptidyl or peptidomimetic extensions, are slow-binding inhibitors, and the protease-inhibitor binding mechanism includes one or more weakly bound intermediates before the formation of the tightly bound E•I complex. This slow-binding inhibition is a hallmark of inhibitors that bind in the active site in a substrate-like manner. In this way, transition-state analogs mimic the association mechanism of many of the naturally occurring protein inhibitors described above, particularly standard-mechanism inhibitors.

Irreversible inhibitors

Serine, cysteine, and threonine proteases, which perform a direct nucleophilic attack on the scissile bond (as opposed to the water-mediated nucleophilic attack performed by aspartic and metalloproteases) are excellent targets for covalent, irreversible inhibitors. Although pharmaceutical companies have been hesitant to pursue covalent inhibitors as drugs because of concerns over their possible cross-reactivity and potential for the development of an unwanted host immune response, it has been shown that a high degree of specificity can be built into these compounds. Furthermore, they are excellent biologic tools: when functionalized with fluorescent dyes or biotin, covalent inhibitors are effective imaging and pull-down reagents and they have been used to determine the presence and roles of
active proteases in many biologic systems. Powers and coworkers (15) authoritatively reviewed these inhibitors, with detailed analysis of their mechanisms of action, structure-activity relationships, and their effectiveness in vivo.

Covalent irreversible inhibitors of cysteine, serine, and threonine proteases are capable of either alkylating or acylating their target enzymes (Fig. 8). Alkylating agents are very effective cysteine protease inhibitors, and they include chloromethyl ketones, fluoromethyl ketones, diazomethyl ketones, acyloxymethyl ketones, epoxides, and vinyl sulfones. The nucleophilic active site cysteine attacks an activated carbon and forms an irreversible carbon-sulfur bond. Because the active site serine of serine proteases is generally less nucleophilic than the corresponding catalytic cysteine, alkylating agents primarily target cysteine proteases. An exception is the chloromethyl ketones, which are capable of inhibiting serine proteases, although they have a slightly different mechanism of inhibition. After a nucleophilic attack of the carbonyl of the inhibitor, the inhibitor alkylates the catalytic histidine. The enzyme can then be deacylated, but the alkylation of the catalytic histidine results in a covalently bound inhibitor.

Inhibitors that acylate proteases present a carbonyl bond for nucleophilic attack, but because of either the poor electrophilicity of the acyl-enzyme intermediate or the inhibitor adopting a conformation unfavorable for deacylation, the enzyme reaction coordinate is trapped in the acyl-enzyme intermediate state. These inhibitors can inhibit both serine and cysteine proteases, and they can generally be sorted into two classes. Peptidyl-acylating agents have a peptide-like specificity element to target the inhibitor to a specific enzyme, but they have a modified scissile bond, such as an aza-group, a carbamate, or an acyl hydroxamate, that are resistant to deacylation. Several heterocyclic acylating agents also exist, such as isocoumarins, β-lactams, cephalosporins, and penems, which have their carbonyl groups enclosed in a ring structure. During nucleophilic attack, the ring is sprung open, and geometric or electrostatic effects can cause a stable acyl–enzyme complex.

Figure 8. Irreversible inhibitors of proteases. Serine and cysteine proteases can be acylated by aza-peptides, which release an alcohol, but cannot be deacylated due to the relative unreactivity of the (thio) acyl-enzyme intermediate. Reactive carbons, such as the epoxide of E64, can alkylate the thiol of cysteine proteases. Phosphonate inhibitors form covalent bonds with the active site serine of serine proteases. Phosphonates are specific for serine proteases as a result of the rigid and well-defined oxyanion hole of the protease, which can stabilize the resulting negative charge. Mechanism-based inhibitors make two covalent bonds with their target protease. The cephalosporin above inhibits elastase (23). After an initial acylation event that opens the β-lactam ring, there are a number of isomerization steps that eventually lead to a Michael addition to His57. Therefore, even if the serine is deacylated, the enzyme is completely inactive.
Protease Inhibitors, Mechanisms of

Phosphonates (Fig. 8) and sulfonates represent a third class of covalent irreversible inhibitors. These inhibitors adopt a stable tetrahedral geometry and are covalently bound transition-state analogs. They often have a peptide-like specificity element, and the electrophilicity of the leaving groups can be modified to tune the reactivity of the inhibitor. These inhibitors are specific for serine proteases, because the serine protease active site has a well-defined oxyanion hole, which stabilizes the transition-state mimic.

A final group of covalent small-molecule inhibitors of proteases are mechanism-based inhibitors. These inhibitors are enzyme-activated irreversible inhibitors, and they involve a "two-hit" mechanism that completely inhibits the protease. Some isocoumarins and β-lactam derivatives have been shown to be mechanistic inhibitors of serine proteases. A classic example is the inhibition of elastase by several cephalosporin derivatives developed at Merck (Fig. 8). The catalytic serine attacks and opens the β-lactam ring of the cephalosporin, which through various isomerization steps, allows for a Michael addition to the active site histidine and the formation of a stable enzyme–inhibitor complex (34). These mechanism-based inhibitors require an initial acylation event to take place before the irreversible inhibitory event. In this way, these small molecules have an analogous mechanism of inhibition to the naturally occurring serpins and α-2-macroglobin, which also act as suicide substrates.

Reversible, competitive inhibitors

Concerns about toxicity, cross-reactivity, and immunogenicity have hampered the development of irreversible therapeutics that target proteases. Therefore, many noncovalent, reversible, competitive protease inhibitors have been developed. Most of these inhibitors interact with both the catalytic residues and the substrate binding sites, as binding to either individual element usually will not provide enough specificity or potency. In some cases, such as small-molecule peptidomimetic inhibitors of thrombin (35) (Fig. 9), binding to the substrate binding sites of the protease has provided sufficient potency, but this is likely because of the buried nature and unique specificity of the thrombin active site. In most cases, though, binding to the catalytic machinery provides potency, whereas the substrate binding sites provide opportunities for specificity. Metalloprotease inhibitors provide a representative example of reversible, competitive, small-molecule inhibitors. Metalloproteases have been the targets of multiple large-scale drug discovery efforts, and many inhibitors have been developed, with varying results. Many inhibitors of the angiotensin-converting enzyme (ACE) have been brought to market to combat hypertension and myocardial infarction. Conversely, the first generation of matrix metalloprotease (MMP) inhibitors failed in clinical trials for the treatment of cancer, in part because the inhibitors had cross-reactivity with other metalloproteases. The common
Inhibitors Developed via Protein Engineering

Protein engineering has allowed for the development of many new protease inhibitors with increased potency and specificity and for diverse mechanisms of action. Because of the relatively shallow active sites, homology, and broad specificity of many proteases, larger molecules are attractive inhibitors in that they can bury more surface area during binding and hopefully gain more potency and specificity.

One strategy has been to improve the specificity of naturally occurring protease inhibitors, either through rational design or via phage display. Mutations to residues that interact with the protease active site have drastic effects on inhibitor affinity, but specificity tends to be gained through evolution of secondary interactions. The standard mechanism serine protease inhibitors ecallin (40) andeginol (41) have been refined at both of their primary and secondary interaction sites, which drastically improve their specificity for a single protease. Although not altering the primary mechanism of action of the inhibitors, the engineering of secondary binding sites gives these inhibitors a mechanism of inhibition similar to that of the anticoagulant inhibitors such as hirudin, and greatly improves inhibitor specificity.

A similar strategy has been to develop polypeptide-based inhibitors of proteases. Typically consisting of 10–20 amino acids, and often containing disulfide bonds to rigidify the inhibitors and decrease the entropic cost of binding, constrained peptides have been developed to inhibit aspartic, cysteine, serine, and threonine proteases. Although peptides are not ideal drug molecules because of their susceptibility to proteolysis, the relatively small size of constrained peptides allows for the creation of extremely diverse libraries. Furthermore, they are amenable to the incorporation of non-natural or D-amino acids, which thus greatly increases potential diversity. The mechanisms of action of these inhibitors have sometimes mimicked known biologic mechanisms, and sometimes they have been completely novel. Constrained peptide phage display libraries have yielded standard mechanism inhibitors of the serine protease chymotrypsin (42) and urokinase-type plasminogen activator (uPA) (43) with moderate potency and specificity. Cyclic peptides have also been shown to inhibit competitively the aspartic protease renin, and they are thought to bind to the enzyme in a substrate-like manner (44).

Constrained peptides that mimic natural inhibitors are essentially a reduction of naturally occurring inhibitors to just their reactive elements. But several allosteric peptide inhibitors have been developed that have novel intramolecular inhibition, which reveal information about enzyme function, and suggest new ways of regulating proteolysis. Constrained peptide libraries have yielded two extremely potent exosite inhibitors of the clotting enzyme factor VIIa (FVIIa) (45, 46). The two inhibitors are bound to two different sites outside of the active site of the enzyme, and they have unique mechanisms of inhibition. One inhibitor, A-183, functioned by forcing a loop near the active site into an inactive conformation and by occluding substrate binding to the enzyme. The other inhibitor, E-76, was a non-competitive inhibitor of FVIIa’s natural substrate, factor X, and...
it seems to work by locking the enzyme in a zymogen-like conformation. In another example of allosteric inhibition, an α-helical peptide was designed to disrupt—and thus prevent activation of—the dimerization of the protease from Kaposi’s sarcoma-associated herpes virus (KSHV) (47). The mechanisms of inhibition of these peptide inhibitors clearly overlap with those of both small-molecule and naturally occurring allosteric inhibitors; namely they lock their target enzymes in an inactive, closed, or zymogen state. The allure of allosteric inhibition is founded in the idea that it is possible to find multiple sites on an enzyme to regulate activity, and these molecules have done this by using established mechanisms to inhibit new proteases. A third approach has been to mature specific protease inhibitors on other natural protein scaffolds, such as antibodies. Antibody inhibitors have been raised against metalloproteases (48), cysteine proteases (49), and serine proteases (50), and to this point, characterized inhibitors have either been monoclonal antibodies raised from hybridomas or from phage-display libraries. The mechanisms of inhibition are familiar; protease antibody inhibitors either interfere with multimerization (and thus activation) of a protease (51), bind to loops and protein–protein interaction sites to occlude substrate binding, or bind in the protease active site. The benefit of using antibodies as inhibitors is that they are exquisitely specific—antibodies have evolved to bind specifically to their antigen—and are useful biologic tools for imaging and in vivo experiments. And that they have been able to mimic the mechanism of inhibition of naturally occurring protease inhibitors suggests that specific antibody inhibitors can be developed for many proteases.

Conclusions and Future Directions

Relatively few design principles underlie the mechanisms of inhibition of an awe-inspiring range of protease inhibitors. Protease inhibitors tend to compete with substrate binding, either through direct competition or deformation of the protease active site. Although protein inhibitors can gain potency through the burial of a large surface area and specificity through contacts with specific exosites, small-molecule inhibitors primarily gain potency through interactions with the catalytic machinery of the enzyme and specificity through interactions with the substrate binding sites. The search for novel modes of enzyme control, such as allosteric regulation, is particularly exciting, with the hope that these regulatory sites will be more amenable to the design of specific inhibitors. The design of new inhibitors based on these principles is critically important. A great deal remains to be learned about the role of proteases in biology, and effective, reliable chemical tools are needed to tease apart these processes. The failure of MMP inhibitors in clinical trials stemmed from toxicity problems (cross-reactivity that could be eliminated with more specific inhibitors), and just as critically, an incomplete understanding of which MMP was important to what stage of cancer growth and metastasis. A more thorough understanding of the role of individual proteases might have alleviated some of these issues. Because of their critical roles in biology, proteases are attractive drug targets. Although only a handful of anti-protease drug targets have approved inhibitors on the market thus far (DPPIV, HIV, ACE, and prostate-specific protease inhibitors), several drug discovery efforts that target proteases are underway, which aim to treat, for example, viral infections, parasitic infections, thrombosis, osteoporosis, neurodegenerative diseases, and cancer. As our understanding of the targets and the design principles of their inhibitors improve, more treatments are sure to follow (7, 12–14, 16–21).

References


Further Reading

See Also
Approaches to Enzyme Inhibition A spartic Proteases and A spartic Protease Inhibitors Cysteine Proteases and Cysteine Protease Inhibitors Metalloproteinases, Biological and Chemical Properties of Serine Proteases and Serine Proteinase Inhibitors
Serine proteases and Serine Protease Inhibitors

Michael J. Page and Enrico Di Cera, Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri

doi: 10.1002/9780470048672.wecb534

Serine proteases are among the largest group of proteolytic enzymes in the human genome that play vital roles in health and disease. Regulation of their activity in vivo is mediated by a diverse group of serine protease inhibitors. An overview of the interplay between serine proteases and their inhibitors is provided. In addition, approaches to characterize this relationship are discussed with subsequent emphasis on how such measures apply to pathologies that result from defect of serine proteases or their inhibitors.

Cylindrical life is brokered on the activity of proteins, and, in turn, control over their concentration and state is vital. Enzymes that hydrolyze peptide bonds, the proteases, are therefore critical ingredients of the genome. Proteases may act as nonspecific agents of digestion or high-selectivity mediators of posttranslational modification. Proteases are a diverse group of enzymes. Over 180 phylogenetically distinct families of proteases have been identified by the MEROPS protease classification system on the basis of protein fold and additionally separated through phylogenetic relationship (1). Of these families, the largest contingent comprises serine proteases, which are named based on their application of the hydroxyl group of a serine side chain as catalytic nucleophile. Regulation of protease function in vivo is mediated by a minimum of 90 families of protease inhibitors. Together, these two groups control biological functions both inside and outside of the cell. Genetic defects in either protease or inhibitor, therefore, can result in significant pathology with potentially multiple biological pathways impacted.

Serine Proteases and their Inhibitors

More than a third of all known proteolytic enzymes are serine proteases (2). The family name stems from the nucleophilic serine residue within the active site, which attacks the carbonyl moiety of the substrate peptide bond to form an acyl-enzyme intermediate. Nucleophilicity of the catalytic serine is commonly dependent on a catalytic triad of aspartic acid, histidine, and serine—commonly referred to as a charge relay system (3). First observed by Blow over 30 years ago in the structure of chymotrypsin (4), the same combination has been found in four other three-dimensional protein folds that catalyze hydrolysis of peptide bonds. Examples of these folds are observed in trypsin, subtilisin, prolyl oligopeptidase, and ClpP protease. Many other enzyme families use the same catalytic triad, such as asparaginases, esterases, acylases, and β-lactamases (5). Mutagenesis of the aspartic acid to alanine impacts peptide bond hydrolysis to a greater extent than ester hydrolysis, which indicates that a complete catalytic triad is required for the hydrolysis of the stronger peptide bond. It should be noted that several serine protease families use a dyad mechanism in which lysine or histidine is paired with the catalytic serine. Yet, other serine proteases present novel catalytic triads, such as a pair of histidines combined with the nucleophilic serine. In nearly all reported cases, the active site serine can be rendered inactive by generic inhibitors such as diisopropylfluorophosphate and phenylmethanesulfonyl fluoride. We will focus on the most abundant serine protease and serine protease inhibitor families in the human genome. The reader is also referred throughout the text to other recent and more expansive reviews and to the MEROPS database for a more detailed description of the impressive diversity of serine protease and inhibitor structure, function, and activity.

A typical genome contains 2–4% of genes that encode for proteolytic enzymes. The entire complement of peptidases within a genome is referred to as the degradome (6). Of these proteases, a select subset of peptidases underwent considerable gene duplication and divergence. The trypsin-like serine peptidases (Clan PA Family 51 Subfamily A in MEROPS nomenclature) are the largest group of homologous peptidases in the human genome responsible for various critical biological processes. Similar degradation composition is observed in all vertebrates, which indicates that expansion of the S1A peptidase family occurred before emergence of the lineage roughly 450 million years ago. Of 699 peptidases in humans, 178 are serine peptidases, and 138 of them belong to the S1A peptidase family. The chymotrypsin-like fold of the S1A peptidase family presents an ideal catalytic platform that enables high turnover, substrate...
selectivity, and various modes of regulation in a package readily combined with additional protein domains.

Clan PA: S1 peptidases—the trypsins

Chymotrypsin-fold proteases are the largest family of peptidases known. Pioneering studies used digestive enzymes such as trypsin and chymotrypsin, which cleave polypeptide chains on the C-terminal side of a positively charged side chain (arginine or lysine) or large hydrophobic residue (phenylalanine, tryptophan, or tyrosine), respectively. In Schechter & Berger nomenclature, the peptide bond hydrolyzed lies between the P1 and P1’ residues of the substrate and is numbered increasingly toward both the N- and C-terminal directions. Interaction occurs in the similarly numbered subsite (S) in the protease. For example, the S2 pocket binds the P2 residue, which is the second amino acid on the N-terminal side of the scissile bond. Determination of other polypeptide sequences of several serine proteases revealed a large family of enzymes. Trypsins and chymotrypsins belong to the S1A family of clan PA, whereas the S1B family comprises various bacterial proteases and the HTRA subgroup of proteases responsible for intracellular protein turnover.

Both subfamilies share the two β-barrel architecture. Two Greek-key β-barrels comprise the chymotrypsin fold and are homologous in a nontraditional manner (Fig. 1a). β-strand topology of the fold reveals a Greek-key architecture in both barrels, yet this topology originates from sequences that run in the opposite direction.

Hence, the two halves of the structure are mirror images in protein fold space. Both barrels are functionally partitioned with one end involved in catalysis and a second in regulation. The active site lies in the cleft between them.

The conventional catalytic triad in S1 family peptidases mediates peptide bond hydrolysis (Fig. 1b). Hydrogen bonding between residues Asp-102 and His-57 (chymotrypsin numbering is used) facilitates abstraction of the proton from Ser-195 and creates a potent nucleophile. Stabilization of the catalytic triad is mediated through a network of additional hydrogen bonds provided by several highly conserved amino acid residues that surround the triad, particularly Thr-54, Ala-56, and Ser-214, and buttress it along a disulfide bridge between residues 42 and 58. The reaction proceeds via pair tetrahedral intermediates. The first step, nucleophilic attack by the serine yields an oxoacyl tetrahedral intermediate stabilized by the backbone amides of Gly-193 and Ser-195. Collapse of the tetrahedral intermediate generates an acyl-enzyme intermediate, and stabilization of the newly created N-terminus is mediated by His-57, Hartley and Kilby provided evidence for the acyl-enzyme intermediate in 1954 (9). In these initial experiments, a pre-steady state burst of product correctly identified that a bond to a hydroxyl moiety.

The second half of the reaction is a reversal of the first step, in which a water molecule displaces the free polypeptide fragment and attacks the acyl-enzyme intermediate. Again, the oxoacyl intermediate stabilizes the second tetrahedral intermediate, and collapse of this intermediate liberates a new C-terminus in the substrate. Central to the regulation of peptidase activity is zymogen activation.

Activation of most chymotrypsin-like serine peptidases requires proteolytic processing of an inactive zymogen precursor protein. Cleavage of the proprotein occurs at an identical position in all known members of the family: between residues 15 and 16. The nascent N-terminus induces conformational change in the enzyme through formation of an intramolecular electrostatic interaction with Asp-194 that stabilizes the oxoacyl enzyme and substrate-binding site. Zymogen activation provides a powerful mechanism of regulation that endows temporal control over protease activity, an ability to escape premature enzyme inhibition, and places these enzymes within the context of chains of proteolytic events. Many proteases of the coagulation and immune pathways are regulated more through allosteric mechanisms that involve monovalent cations (Na+), divalent cations (Ca2+), glycosaminoglycans, and protein cofactors (11). These properties derive from the structure of the chymotrypsin fold and combine to produce proteolytic networks responsible for key biological processes responsible for human health.

Several vital processes rely on clan PA peptidases. Chief among them are blood coagulation and the immune response, which involve cascades of sequential zymogen activation. In both systems, the chymotrypsin-fold peptidase domain is combined with one more associated protein domains, including apoproteins, CUB, EGF, fibronectin, kringles, sufl, and von Willebrand factors. These protein domains are on the N-terminus as an extension of the propeptide segment of the peptidase. Such a trend of N-terminal-associated domains in the S1A peptidase family is common across all forms of life. The domain architecture pairs well with the zymogen activation mechanism, which liberates the proper N-terminus to enable catalytic activity. Often, the associated protein domains remain attached to
the peptidase domain through a covalent disulphide bond on the opposing surface of the protease active site. Many associated domains are entirely encoded by a single exon in their peptidase gene and suggest an important role for exon shuffling during molecular evolution of clan PA.

S1 peptidases in the human genome are, for the most part, primary selectivity toward amino acids in length derived from the two-domain architecture. An N-terminal proline and hydroxyproline residues from peptides (20). Substrate selectivity for peptides shorter than 30 amino acids in the digestive system signals proinflammatory responses typi- cally mediated by trypsin-like S1 peptidases. Trypsinases are major components in secretory granules of mast cells that are unique among clan PA peptidases because of their homote- trameric quaternary structure (13). Like trypsin, trypsinases mediate proinflammatory signaling through protease-activated recep- tors 2, yet definition of other substrates in health and disease states remain elusive. Matriptases are membrane-bound S1 pep- tidases that bear primary substrate selectivity similar to trypsin (14). A gain, physiological substrates of this subfamily of pep- tidases are largely unknown, yet high gene expression levels for matriptases are associated with various cancer types. Simi- lar association with cancer has led to great interest in the large family of kallikreins (15), a family commonly known for its role in regulation of blood pressure through the kinin system (16). The enzymes are mediators of directed apoptosis by natural killer cells and cytotoxic T cells that play key roles in the defense against viral infection (17). Notably, unique among clan PA is the primary selectivity of granzymes toward acidic residues in the P1 position of the substrates. Of the wide diversity of pro- teases in clan PA Family S1, the mediators of immunity and blood coagulation have been particularly well studied.

Clan SC: peptidase diversity in the α/β fold

Clan SC peptidases are α/β hydrolase-fold enzymes that con- sist of parallel β-strands surrounded by α-helices. The α/β hydrolase-fold provides a versatile catalytic platform that, in additional to achieving proteolytic activity, can either act as an esterase, lipase, dehalogenase, haloperoxidase, lyase, or pro- oxidase hydrolase (18). Six phylogenetically distinct families of clan SC are known, and only four of them have known structure. Catalytic amenability of the α/β hydrolase-fold may underlie why clan SC peptidases are the second largest family of ser- ine peptidases in the human genome. Other mechanistic classes need not use the catalytic serine and instead use cysteine or glutamic acid (19). Clan SC peptidases present an identical geo- metry to the catalytic triad observed in clans PA and S1, yet this conformation is ordered differently in the polypeptide se- quence. Substrate selectivity develops from the α-helices that surround the central β-sheet core. Within clan SC, carboxypep- tidases from family S10 are unique for their ability to maintain catalytic activity in acidic environments. Nearly all serine pep- tidases have activity restricted within the range of neutral to alkaline pH. Many clan SC peptidases hydrolyze substrates on the C-terminal side of a proline residue with several excep- tions. Both endoproteolytic and exoproteolytic activities occur in clan SC, which contrasts the trend in other serine pepti- dase families in which members are primarily one or the other. For examples of differing selectivity in clan SC, pro- pyl oligopeptidase from family S9 cleaves peptide bonds within peptides, and prolyl aminopeptidase from family S9 cleaves peptide bonds within proteins (20). N-terminal proline and hydroxyproline residues from peptides (20). Substrate selectivity for peptides shorter than 30 amino acids in length derived from the two-domain architecture. An N-terminal seven-bladed propeller restricts access to the C-terminal α/β hydrolase domain and, in turn, the site of peptide bond hydrolysis (21). On the basis of their selectivity toward smaller peptides and not full-length proteins, clan SC pep- tidases are thought to be particularly important in cell signaling mechanisms.

In humans, clan SC peptidases are often associated with proline-specific N-terminal processing of peptides and proteins, yet many present a nonproteolytic function. S9 is the largest family clan SC peptidases with 43 homologs in the human genome. Of these homologs, prolyl oligopeptidase (POP) and dipeptidyl peptidase IV (DPP-IV) are the best characterized. The crystal structure of POP revealed the two-domain architecture and basis for substrate selectivity. Notably, no naturally occurring inhibitor of this family of proteases has been found. A pu- taive role for POP has been suggested in the metabolism of vari- ous neuropeptides (22). DPP-IV presents a similar two-domain architecture (23). DPP-IV is a transmembrane protein responsi- ble for processing hormones, growth factors, and neuropeptides. POP family peptidases have been identified in the human genome, and their biological roles remain to be uncovered. Of three S28 family peptidases in human, dipeptidyl peptidase II (DPP-II) catalyzes release of two N-terminal amino acids when proline or alanine is in the P1 po- sition. Eight S33 family peptidases in the human genome, however, many of them do not display peptidase activity. For example, several of these enzymes catalyze hydrolysis of epox- ide bonds into diols and play a role in detoxification or cellular signaling (24).

Clan SB: family SB—subtilisins

Clan SB peptidases are prevalent in plant and bacterial genomes, with few representatives in a given animal genome. However, these proprotein convertases are vital for protein processing in all metazoans. The archetype of clan SB is subtilisin. Subtilisin was originally discovered in the gram-positive bacterium Bacil- lus subtilis and like chymotrypsin was one of the earliest protein crystal structures determined (25). The catalytic triad of aspartic acid, histidine, and serine is found in the exact geometric orga- nization observed in the peptidases of clans PA and SC, yet the surrounding protein fold bears no similarity (Fig 2). Clan SB also contains a second family of peptidases S53, the serine endoproteases. In these peptidases, the histidine general base is substituted by a glutamic acid, and the tetrahedral intermediate is stabilized by a negatively charged carbonyl group from an aspartic acid
Serine Protease and Serine Protease Inhibitors

Figure 2 Subtilisin (PDB 1SCN) presents an identical catalytic triad to that observed in other serine proteases and enzymes yet within an entirely different protein fold.

Serine Protease Inhibitors

Over 90 phylogenetically distinct families of protease inhibitors have been classified by the MEROPS database. We will focus the current discussion on the most abundant and well-characterized groups.

Canonical mechanism: kunitz- and kazal-type inhibitors

In general, protease inhibitors interact with the peptidase in a canonical substrate-like manner with the protease. In this situation, three to four residues interact in an antiparallel β-sheet fashion within the enzyme active site. Many of these protease inhibitors are small proteins typically 30 to 120 amino acids in length (29). Often, these smaller proteases inhibitors contain many disulfide bonds. Many proteases have extremely high melting temperatures (≥80 °C) and retain their native conformation in the presence of strong chaotropes. Despite heterogeneity of sequence, each of these inhibitors presents a nearly identical conformation in reactive site loop that restricts proteolytic activity. In this process, the reactive site loop does not undergo conformational change when in complex with the protease and the P1 residue is positioned to place the carbonyl oxygen at a very short distance to Ser-195. The carbonyl oxygen, in turn, points toward the oxyanion hole that forms H-bonds with the amide groups of Gly-193 and Ser-195. The amide nitrogen of the P1 residue is directed toward the Oγ of Ser-195 not facing Ser-214 as typically observed in natural substrates. The latter change is thought to shorten during the catalytic cleavage process (30). The loop may be combined with other structural elements to mediate inhibition, including a β-hairpin following the loop or a disulfide bond in close proximity to the scissile bond. During interaction with the protease, the scissile bond remains intact and may undergo a slight deformation from planarity or a more distorted Michaelis complex depending on the enzyme and inhibitor pairing. Few specific contacts define the interaction between protease and inhibitor outside of the active site, where most are typically hydrophobic. In turn, composition of the reactive site loop contributes the most significant energetic in the binding process. Altering the P1 residue by mutagenesis, therefore, can be used to shift the inhibition profile dramatically. Lack of additional contacts ensures that most serine protease inhibitors regulate activity of multiple proteases in vivo.

MEROPS identifies Kunitz-type inhibitors as families I2 and I3, yet they seem to have developed separately in evolutionary history. Families I2 and I3 are referred to as “Kunitz-A” and “Kunitz-P” for their origin from animals and plants, respectively. Aprotinin, also known as bovine pancreatic trypsin inhibitor, was one of the first protease inhibitors identified and isolated by Kraut and coworkers in 1930. The I2 family is considerably more homogenous and thought to inhibit only S1 peptidases. In contrast, the I3 family is split into two phylogenetic groups, I3A and I3B, both of which typically inhibit S1 peptidases. In addition to the I3 inhibitor, was the...
complex of soybean trypsin inhibitor with trypsin by Sweet and colleagues (31). The structure presents a β-barrel architecture capped by a pair of β-strands stabilized by two disulfide bonds (Fig. 3). The physiologic function of the Kunitz-type proteases remains unknown for many family members other than those in man. Prevention against digestion or invasion from pathogens has been suggested based on a common abundance in seeds. In man, tissue factor pathway inhibitor is a key Kunitz-type inhibitor responsible for regulating blood clot formation. On the basis of their specificity, Kunitz-type inhibitors were among the first examined for therapeutic application. Aprotinin was approved for clinical application in coronary-artery bypass graft surgery in 1993. Fifteen years later, considerable controversy has developed over its use given an associated risk of mortality and the availability of less expensive lysine analogs that achieve the same goals (32).

Kazal-type inhibitors are classified as family 11 by the MEROPS database. The name is derived from pancreatic secretory inhibitor, which is now termed SPINK1, originally isolated by Kazal and coworkers in 1948. The SPINK family (serine protease inhibitor, Kazal) plays important roles in the digestive system, lungs, skin, and likely many other tissues in the body. Six SPINKs can be identified in the human genome, and each contains multiple repeats of the Kazal-type fold. Mutations in SPINK1 are associated with hereditary pancreatitis (33). Netherton syndrome is a rare disorder that affects the skin of patients and results in ichthyosiform dermatosis and hair shaft abnormalities. Patients with Netherton syndrome are found to have a mutation on chromosome 5 in the SPINK5 gene, which encodes a protein that is a serine protease inhibitor. Kazal-type domains are often found within a single polypeptide chain, they need not inhibit the same type of protease or protease specificity. For example, dog bikazin contains two Kazal-type domains in which one domain prefers trypsin, and the other prefers chymotrypsin.

Bait-and-trap mechanism: serpins

Serpins are found in all kingdoms of life, yet their presence in a given organism is not, which suggests the family has undergone significant gene transfer and loss. The family name was coined by Carrell and Travis as an acronym of serine protease inhibitor (36). Serpins are the most abundant form of serine protease inhibitor in the human genome. With the exception of fungi, all multicellular eukaryotes seem to possess one or more serpin genes. However, despite their ubiquity, few physiological functions are ascribed to serpins outside those known in man, and, in particular, they are associated with pathologies. Most serpins are irreversible inhibitors of serine proteases of the S1 family of peptides with select family members inhibiting S9 of subtilisins and chymotrypsin. Several members of the serpin family have lost their ability to act as inhibitors and acquired new functions such as ovalbumin and pigment epithelium-derived factor. The unique mechanism of protease inhibition by serpins has received considerable attention. Serpins consist of a conserved core of three β-sheets and eight or nine α-helices that act collectively in the inhibitory mechanism. As with the Kazal- and Kunitz-type inhibitors, the mechanism involves a surface exposed loop that is termed the reactive center loop (RCL). The RCL presents a short stretch of polypeptide sequence bearing the P1–P1' scissile bond. Like other serine protease inhibitor families, the P1 residue dominates the thermodynamics that govern the interaction between protease and inhibitor. Exposure of the P1 residue to solvent is typically brokered by 15 amino acids N-terminal to the P1 residue and 5 residues on the C-terminal “prime” side of the scissile bond. Evidence for dramatic conformational change in the inhibitory mechanism was first provided by the crystal structure of the cleaved form of α2-antitrypsin (37). In this structure and unlike the native form, the reactive center loop was not solvent exposed but occurred as an additional β-strand within the core of the structure. Dramatic conformational change of both inhibitor and protease is the most recognizable feature of the serpin mechanism (Fig. 4). A first formation of the Michaelis-Menten encounter complex, the reactive center loop is cleared, and an acyl enzyme intermediate is formed as in the normal serine protease catalytic mechanism. However, after bond hydrolysis, the RCL rapidly inserts into the central β-sheet of the serpin, which yields an overall stability enhancement to the inhibitor and traps the acyl-enzyme intermediate. It largely unknown how this change in conformation occurs. Several studies suggest the complex undergoes transient exchange between expelled and partially inserted states (38). Integration of the β-strand into the structure flips the protease from the “top” of the structure to the complete opposite side of the serpin. Because of this, the local environment of the catalytic triad is distorted and therefore cannot complete the catalytic cycle (39). When the process is finished, the previously adjacent P1 and P1' residues are separated by over 70 Å. During this process, the protease has been
Serine Protease and Serine Protease Inhibitors

Figure 4 A large conformational change defines the mechanism of serpin inhibition. Conversion of the Michaelis complex (PDB 1OPH) into cleaved trapped conformation (PDB 1EZX) traps the RCL of serpin into the \( \beta \)-strand core of inhibitor. A significant gain in stability, therefore, is imparted to the entire serpin structure.

converted from a metastable free state into a more energetically favored relaxed bound state. Serpins have a considerably lower melting temperature \((T_m \sim 60^\circ C)\) in isolation than when cleaved \((T_m >100^\circ C)\). Unrelated in sequence or structure, the macroglobulin family of protease inhibitors similarly applies scissile bond cleavage, yet the subsequent step involves entrapment of the protease in a cage-like architecture \((40)\). Although most serpins apply this mechanism to inhibit serine proteases irreversibly, a select group has been shown to act reversibly. For example, protein C inhibitor, also known as PAI-1, acts as a reversible inhibitor to the single-chain urokinase-type plasminogen activator. Moreover, several serpins are known to integrate their cleaved RCL into another serpin molecule in trans \((41)\). In turn, serpins can undergo polymerization, which becomes relevant in several pathological conditions.

Serpins play key regulatory functions in man. \( \alpha_1 \)-antitrypsin serves a major role in protecting the connective tissue of the lungs from leukocyte-released elastase. The C1 inhibitor restricts proteases of the immune system from unwanted proteolysis and inflammation. Two plasminogen activator inhibitors control fibrinolysis. Viral serpins have also been described that broker their survival and propagation through restricting these same proteolytic pathways. Control of blood clot formation is through antithrombin. However, unlike other serpin family members, the interaction between clotting factor protease and antithrombin is greatly facilitated by heparin or heparin sulfate glycosaminoglycans, which bind to the inhibitor to mediate this effect \((42)\). In turn, antithrombin is directed toward regulation of protease activity at cell surfaces such as the vascular endothelium, which display heparin in various forms.

Exosite recognition: hirudin and other anticoagulant molecules

Numerous strategies have evolved in different pathogenic and parasitic organisms to alter the coagulation cascade for their own benefit. Snake and leech venoms have yielded a plethora of serine proteases and serine protease inhibitors that bear this trait. Many inhibitors function through hijacking the macromolecular recognition regions in blood clotting factors, the earliest known example of which is hirudin, a 66-amino acid residue protein that is an extremely tight binding and selective inhibitor of the blood coagulation protease thrombin. Indeed, the use of leeches in medicine dates back to the Greeks in 200 B.C. However, it was not until 1884 that Haycraft isolated the anticoagulant agent from medicinal leech saliva and termed this agent hirudin. Selective and tight binding results from the cooperative binding at both the anion binding exosite responsible for fibrinogen recognition and active site of thrombin. Notably, unlike nearly all other protease inhibitors, active site recognition involves formation of a parallel \( \beta \)-sheet. Hirudin has been derivatized and modified in various ways to develop direct thrombin inhibitors. Two recombinant forms, lepirudin and desirudin, are available for clinical use to prevent deep-vein thrombosis after surgery and treat heparin-induced thrombocytopenia. However, various problems have limited their use, including bleeding problems,
Characterization of Protease-Inhibitor Interactions

Defining the selectivity and potency of an inhibitor relies on accurate characterization of the protease. In particular, values of $k_{\text{on}}$ and $k_{\text{off}}$ are readily obtained using basic enzymology and various commercially available chromogenic and fluorogenic substrates. Proteases act on a single substrate during the catalytic cycle. Therefore, models to interpret data follow classical descriptions of competitive, noncompetitive, and mixed inhibition. As with traditional enzymology, curve-fitting measures combined with graphical validation of data is suggested as a more accurate approach than the use of initial rates analysis. Use of IC₅₀ is to be avoided as the values of this measure are only a crude comparison between reversible inhibitors. For serine proteases, the use of irreversible inhibitors, such as chloromethyl ketones, are extremely useful for determining the amount of active protease in a given protein preparation. Many serine protease inhibitors form stable complexes with their target proteases that can be resolved via gel electrophoresis as a simple and effective means of visualization. However, many protease-inhibitor interactions require more advanced data treatment. Some examples include slow tight-binding mechanisms and higher-order stoichiometries of inhibition (44). Lastly, conformational change can be measured through various biophysical techniques including UV and fluorescence spectroscopy, circular dichroism, and isothermal titration calorimetry.

Serine Proteases and their Inhibitors

As illustrated above, the human genome encodes a large and diverse population of serine proteases and serine protease inhibitors. Design of small-molecule inhibitors to restrict proteolytic activity continues to garner attention in the pharmaceutical industry. Given the many related proteases in the genome, this task is particularly challenging. Early work focused on active site-directed therapies. However, as evidenced by the naturally occurring serine protease inhibitors, active site recognition enables the regulation of multiple protease targets. Minimization of unwanted side effects is a significant hurdle. More recent effort has sought the development of therapeutics that focus on other regions of the protease. The crucial role for such regions in biological systems is demonstrated by the blood coagulation and immune system proteases, in which macromolecular recognition is dependent on exosites and the allosteric communication of these regions with the active site. Diversity of proteases and inhibitors results in a wide range of pathologies that result from disruption in either serine protease or serine protease inhibitor. The earliest descriptions of such imbalances are within the blood coagulation cascade. For example, hemophilia B results from deficiency in coagulation factor IX. In contrast, excessive activation of immune system serine proteases produces inflammatory states. Inhibitors of serine protease inhibitors can have consequences on multiple biological systems. However, overlapping inhibition by multiple families of inhibitor can, in certain instances, lessen the severity of the pathology. Genetic abnormalities in the serpins have also been associated with polymerization and therefore belong to the category of conformational disease. Emphysema, cirrhosis, and thrombosis may result from such aberrant conformational transitions. Neutrophil may also play a key role in familial enchephalopathy because of the formation of inclusion body-like material. Understanding the molecular mechanisms of limited proteolysis and their regulation in vivo remains a challenging and insightful venue to improve human health.

References

Serine Protease and Serine Protease Inhibitors


Further Reading

The MEROPS database of peptidases and inhibitors is an invaluable resource that can be found at http://merops.sanger.ac.uk/

See Also

Enzyme Kinetics
Protease Inhibitors, Mechanisms of Approaches to Enzyme Inhibition


Thrombin

Enrico Di Cera, Department of Biochemistry and Molecular Biophysics, Washington University Medical School, St. Louis, Missouri
doi: 10.1002/9780470048672.wecb678

Thrombin is a Na⁺-activated, allosteric serine protease that plays opposing functional roles in blood coagulation. Binding of Na⁺ is the major driving force behind the procoagulant, prothrombotic, and signaling functions of the enzyme, but is dispensable for cleavage of the anticoagulant protein C. The anticoagulant function of thrombin is under the allosteric control of the cofactor thrombomodulin. Thrombin exists in three forms, E*, E, and E:Na⁺, which interconvert under the influence of ligand binding to distinct domains. Transitions among these forms control key functional aspects of the enzyme.

Thrombin is a serine protease of the chymotrypsin family, which includes enzymes involved in digestion and degradative processes, blood coagulation, cell-mediated immunity and cell death, complement, fibrinolysis, fertilization, and embryonic development (1). Once generated in the blood from its inactive precursor prothrombin, thrombin plays two important and paradoxically opposing functions. It acts as a procoagulant factor when it converts fibrinogen into an insoluble fibrin clot that anchors platelets to the site of lesion and initiates processes of wound repair. This action is reinforced and amplified by activation of the transglutaminase factor XIII that covalently stabilizes the fibrin clot, the inhibition of fibrinolysis via activation of TAFI, and the proteolytic activation of factors V, VIII, and XI (2). In contrast, thrombin acts as an anticoagulant through activation of protein C (3). This function unfolds via an interaction with thrombomodulin, which is a receptor on the membrane of endothelial cells. Binding of thrombomodulin suppresses the ability of thrombin to cleave fibrinogen and PAR1, but it enhances -1,000-fold the specificity of the enzyme toward the zymogen protein C. Activated protein C cleaves and inactivates factors Va and VIIa, two essential cofactors of coagulation factors Xa and IXa that are required for thrombin generation, thereby downregulating both the amplification and progression of the coagulation cascade. In addition, thrombin is inhibited irreversibly at the active site by the serine protease inhibitor antithrombin with the assistance of heparin and by the thrombin-specific heparin cofactor II. Important cellular effects are triggered by thrombin cleavage of protease-activated receptors (PARs) (4, 5), which are members of the G-protein-coupled receptor superfamily. Four PARs have been identified and share the same basic mechanism of activation: Thrombin and other proteases cleave at a specific site within the extracellular N-terminus, which exposes a new N-terminal tethered ligand domain that binds and activates the cleaved receptor. PAR1 is responsible for platelet activation in humans at low thrombin concentrations, and its action is reinforced by PAR4 at high enzyme concentrations. Activation of PAR1 and PAR4 triggers platelet activation and aggregation, and it mediates the prothrombotic role of thrombin in the blood. PAR3 is not present on human platelets, but it is expressed widely and abundantly in other cell types. PAR2 is the target of other proteases and clotting factors. In the mouse, signaling in platelets is mediated entirely by PAR4, with PAR3 facilitating PAR4 cleavage at low thrombin concentrations. The efficiency of the coagulation cascade depends on the balance between the procoagulant and anticoagulant pathways. Thrombin is the key arbiter of this balance by virtue of its dual role and has therefore received utmost attention in structure-function studies and as a target of anticoagulant therapy (6). Recent advances demonstrate that thrombin uses the transition among different conformations to perform its numerous physiological roles.

Thrombin, Na⁺ and Thrombomodulin

Thrombin requires Na⁺ for optimal activity (2) and is a member of the large family of monovalent cation activated enzymes (7, 8). Na⁺ binding converts thrombin from a low activity E (Na⁺-free) to a high activity E:Na⁺ (Na⁺-bound) form. Na⁺ binding is required for optimal cleavage of fibrinogen and activation of factors V, VIII, and XI necessary for the explosive generation of thrombin in the coagulation cascade, but it is dispensable for cleavage of protein C (9, 10). This finding proves that Na⁺ is the major driving force behind the procoagulant role of thrombin in the blood. Na⁺ binding also promotes the prothrombotic and signaling functions of the enzyme by enhancing cleavage of PAR1, PAR3, and PAR4. Because of the allosteric nature of thrombin, any effect that destabilizes Na⁺ binding stabilizes the E form and produces an anticoagulant effect by prolonging the clotting time (reduced
fibrogen cleavage and reducing platelet activation (reduced PAR1 cleavage). Several naturally occurring mutations of the prothrombin gene affect residues linked to Na\(^{+}\) binding (11) and are often associated with bleeding (2). The differential effect of Na\(^{+}\) on fibrogen and protein C cleavage has been the major driving force behind the rational design of anticoagulant thrombin mutants.

The anticoagulant role of thrombin is under the control of the endothelial receptor thrombomodulin (3). In the microcirculation, the density of this receptor is enhanced greatly, and thrombin is sequestered by engagement of its exosite 1, which is a structural domain separate from the active site. Binding of thrombomodulin to exosite 1 precludes interaction of fibrinogen or PAR1 and therefore shuts down both the procoagulant and prothrombotic functions of the enzyme. At the same time, the thrombin-thrombomodulin complex becomes an efficient activator of the anticoagulant protein C by markedly increasing the k\(_{\text{cat}}\) and the rate of diffusion of substrate into the active site. Activation of protein C produces a protease that rapidly inactivates factors Va and VIIa necessary for additional thrombin generation.

Thrombin is composed of two polypeptide chains of 36 (A chain) and 259 (B chain) residues that are covalently linked through a disulfide bond (Fig. 3). The standard orientation (12) puts the A chain in the back of the molecule, opposite to the front hemisphere of the B chain that hosts the entrance to the active site and all known functional epitopes of the enzyme. Trypsin-like specificity for Arg residues at P1 is conferred to the active site and defines exosite 1. Structural data illustrate the mode of interaction of fibrogen (13), thrombomodulin (14), PAR1 (15), and PAR3 (16) with this thrombin exosite. Exosite 2 resides on the pole of the enzyme opposite to exosite 1 and offers a binding epitope for the platelet receptor GP\(_{\text{b,IIa}}\), as well as for polyvalent ligands required for interaction of antithrombin and heparin cofactor II. Na\(^{+}\) binding 20 Å away from residues of the catalytic triad and within 5 Å from D189 in the S1 site that occupies the bottom of the catalytic pocket. The autoinhibitory loop shapes the lower rim of access to the active site and contributes to recognition of fibrogen. A loop homologous to the Ca\(^{2+}\) binding loop of digestive serine proteases resides 25 Å away from the active site and defines exosite 1. Structural data illustrate the mode of interaction of fibrogen (13), thrombomodulin (14), PAR1 (15), and PAR3 (16) with this thrombin exosite. Exosite 2 resides on the pole of the enzyme opposite to exosite 1 and defines most of the aryl binding specificity pocket and the active site linked to the occupancy of the Na\(^{+}\) binding site (11). An important consequence of the large Na\(^{+}\) binding causes a significant increase in intrinsic fluorescence of the enzyme with an initial rapid phase with a k\(_{\text{cat}}\), which is followed by a slow phase with a k\(_{\text{cat}}\) in the ms time scale that decreases hyperbolically as Na\(^{+}\) increases (2, 18).

These examples are signatures of a two-step mechanism for Na\(^{+}\) binding as follows in Scheme 1:

Thrombin exists in equilibrium between two forms, E\(^{+}\) and E, that interconvert with kinetic rate constants k\(_{1}\) and k\(_{-1}\) in the ms time scale. Of these forms, only E can interact with Na\(^{+}\) with a rate constant k\(_{2}\) to populate E:Na\(^{+}\), which may dissociate into the parent components with a rate constant k\(_{\text{diss}}\). The fast phase is caused by the E-E:Na\(^{+}\) interconversion that involves Na\(^{+}\) binding/dissociation, and the slow phase is caused by the E-E\(^{+}\) interconversion that precedes Na\(^{+}\) binding. E and E:Na\(^{+}\) in Scheme 1 are the two active forms of thrombin that account for the dependence of k\(_{\text{cat}}\) on [Na\(^{+}\)]. However, a third form E\(^{+}\) is present in solution regardless of Na\(^{+}\) or any other ligand. The E\(^{+}\)-E equilibrium of thrombin is an intrinsic property of the enzyme uncovered by its effect on the kinetics of Na\(^{+}\) binding. This example demonstrates how studies of Na\(^{+}\) binding to thrombin and related clotting proteases can advance significantly our understanding of the function and regulatory interactions of an entire class of enzymes.

Thermodynamic signatures of Na\(^{+}\) binding are of mechanistic and physiological importance and were also resolved from temperature studies of the kinetic mechanism. A large enthalpy change of \(-22\) kcal/mol is compensated by a large entropy loss of \(-64\) cal/molK. The enthalpy change is caused by formation of the six ligating interactions in the coordination shell, and the entropy change reflects the uptake and ordering of water molecules within the channel embedding the primary specificity pocket and the active site linked to the occupancy of the Na\(^{+}\) site (11). An important consequence of the large enthalpy change is that the equilibrium association constant for Na\(^{+}\) binding becomes 10 M\(^{-1}\) at 37°C, which implies that under physiological conditions of temperature and (NaCl)] = 140 mM, thrombin is only 60% bound to Na\(^{+}\). The fraction of thrombin in the procoagulant E:Na\(^{+}\) form is about 60%, and the anticoagulant form E accounts for about 40%. E\(^{+}\) makes only a little contribution and represents \(-1%\) of the population of thrombin molecules at 37°C.

Single-site Phe mutants of each of the nine Trp residues of thrombin were used to identify fluorophores responsible for the spectral changes associated with Na\(^{+}\) binding (18). The W141F and W215F mutants lose \(-70%\) of the total fluorescence change and retain only the slow phase. The environments of W141 and W215 change in the E\(^{+}\) to E conversion and more drastically in the conversion of E to E:Na\(^{+}\). W215 is the closest Trp residue to the bound Na\(^{+}\) (13 Å) and defines most of the (aryl) binding site involved in substrate recognition (Fig. 4). W141 is buried in the active site.
Molecular Basis of Thrombin Allostery

Structural information has begun to develop on how the Na⁺ site and exosite 1 communicate allosterically with the active site and each other. Structures of E and E:Na⁺ (11) are highly similar (r.m.s.d. 0.38 Å), but significant differences are worth mentioning. D189 in the E:Na⁺ form is optimally oriented for electrostatic coupling with the P1 Arg of substrate, which may account for the lower Kₜₐₜ observed in the presence of Na⁺. Subtle changes affect the catalytic triad. In the E form, the nucleophilic S195 side chain is rotated about 35° relative to the E:Na⁺ form, and the critical H-bond with the catalytic H57 is broken. Integrity of the H-bond is important for catalysis, and the unfavorable position of S195 in the E form may explain the lower kₜₐₜ observed in the absence of Na⁺. Importantly, the structures of E and E:Na⁺ offer a plausible explanation for the long-range communication between the Na⁺ site and the active site. In the E:Na⁺ form, a network of 11 water molecules connect the bound Na⁺ to the Oγ atom of S195, which is located <15 Å away, through H-bonds. In the E form, only seven water molecules occupy positions in the network equivalent to those observed in the E:Na⁺ form, and the connectivity is radically altered. Binding of Na⁺ organizes a network of water molecules that spans the interior of the enzyme for over 15 Å up to the catalytic S195. Ordering of this network shapes the structure of the enzyme as a whole. Ala scanning mutagenesis of the epitopes of thrombin energetically linked to Na⁺ binding and ligand recognition confirm this scenario and vouch for long-range allosteric communications involving the active site, the Na⁺ site and exosite 1 (11, 19).
kinetic and thermodynamic signatures of Na$^+$ binding detected experimentally.

A major advance in our understanding of thrombin allosteric changes in the active site, the S1 site, and the Na$^+$ site that are unprecedented in thrombin and the entire realm of serine proteases and offers a compelling model of the E$^*$ form. The changes impact the conformation of the enzyme in a significant way and are consistent with a global perturbation of the structure predicted from mutagenesis and kinetic studies. The 215-219 β-stand collapses into the primary specificity pocket (Fig. 2) and W215 packs against the hydrophobic pocket in the active site formed by W60, Y60a, H57, and L99. R221a in the Na$^+$ loop brings its guanidinium group in contact with D102 in S1 site to mimic the P1-$^*$ of an incoming substrate. R187 penetrates the Na$^+$ site to occupy the space available to the cation. The structure of D102 N portrays a conformation of thrombin that cannot bind Na$^+$ and ligands to the active site, as expected for the E$^*$ form. The drastic movement of W215 confirms the major role of this residue as a reporter of the E$^*$ and E-E:Na$^+$ interconversions.

Binding of thrombomodulin to exosite 1 enhances the $K_{\text{cat}}/K_m$ of thrombin for protein C $\sim 1,000$-fold, because of a 3,000-fold increase of the rate of substrate diffusion into the active site. Notwithstanding this massive functional effect, the structure of thrombin bound to a fragment of thrombomodulin at exosite 1 fails to reveal significant conformational changes in the active site or other regions [14]. Such changes might have been obliterated by the presence of the active site inhibitor used in the crystalization. However, the structure of thrombin solved with hirudin bound to exosite 1 and the active site free also fails to reveal significant conformational changes in the active site [21]. However, a recent structure of thrombin bound to a fragment of PAR3 bound to exosite 1 [16] reveals a rearrangement of the 60-loop that lines the upper rim of the active site entrance (Fig. 1). The indole ring of W60 d partially occludes access to the active site and restricts specificity toward physiologic substrates and inhibitors. When PAR3 binds to exosite 1, the 60-loop shifts 3.8 Å upward and causes a 180° flip of W60 d that projects the indole ring into the solvent and opens up the active site fully. This structure reveals the basis of the allosteric communication between exosite 1 and the active site clef as well as the role of W60 d to move like a flap and regulate substrate diffusion into the active site. The allosteric effect is relevant for the cofactor function of cleaved PAR3 on PAR4 activation (4, 5), and the effect of thrombomodulin on the cleavage of protein C by thrombin [3].

More details on how exosite 1 communicates long range with the active site and the Na$^+$ site have developed from the structure of thrombin D102 N bound to a fragment of PAR3 at exosite 1 [15]. The E$^*$ conformation of D102 N is converted into the active E$^*$ form by a massive conformational change that restores access to the active site and patency of the Na$^+$ site.

The structure reveals the details of this allosteric transition in terms of a network of polar interactions that can be traced from exosite 1 to the opposite side of the molecule almost 30 Å away.

Conclusions and Future Directions

Thrombin continues to provide a biologically important platform for the study of the most basic aspects of protease structure, function, and regulation. Much has been learned about thrombin function and regulatory interactions, and structural details have begun to develop on the molecular basis of thrombin allosteroy mediated by Na$^+$ and thrombomodulin. Kinetics studies on Na$^+$ binding have revealed the true complexity of thrombin in terms of its ability to assume multiple conformations in equilibrium. The structure of E$^*$ confirms the prediction from kinetic studies and future studies will certainly broaden our understanding of the molecular basis of long-range communications among the active site, the Na$^+$ site and exosite 1.

Advances in our basic knowledge of mechanisms of substrate recognition and cofactor interactions will eventually foster new strategies to control or exploit the multiple functions of thrombin for pharmacological intervention.

Several questions remain on thrombin structure, function, and regulation. Future studies should focus on validation of the nature of E$^*$ and establish if this is indeed an inactive form of thrombin that cannot bind substrate at the active site. Structural studies should address more directly the role of thrombomodulin using thrombin in the E$^*$ or E$^*$ forms. The physiological role of E$^*$ should be elucidated. More direct assessment of the thermodynamic signatures of thrombin interactions and regulatory transitions should rely on calorimetry. Finally, future studies should address the molecular basis of Na$^+$ and substrate specificity of thrombin and how they can be reengineered. Of particular importance will be the design of more potent anticoagulant thrombin mutants or mutants with selective activation of PARs.

References


Further Reading


See Also

Serine Proteases and Serine Protease Inhibitors
Nucleobase Self-Assembly Codes, Constitutional Transcription of
Mihail Barboiu, Adaptive Supramolecular Nanosystems Group, Institut Européen des Membranes, IEM/UMII, Place Eugène Bataillon, Montpellier, France

doi: 10.1002/9780470048672.wecb527

The functional self-organization can be transcribed readily into hybrid nanostructures by using the sol-gel process. Accordingly, we have reported synthetic routes for preparing self-organized systems that have been "frozen" in a siloxane matrix, as a straightforward approach to design a novel class of solid hybrid nanomaterials. Nucleobases oligomerization can be an advantageous choice to reinforce the controlled communication between interconnected "dynamic supramolecular" and "fixing siloxane" systems. Moreover, the different interconverting outputs that nucleobases may form by oligomerization define a dynamic polyfunctional diversity that may be "extracted selectively" by sol-gel polymerization in solid state under the intrinsic stability of the different nucleobase-pairing and G-quadruplex-based systems. The nucleobase-type hybrid materials presented in this review unlock the door to the new self-organized materials world paralleling that of biology.

Introduction

In a broadest sense, the self-assembly and the self-organization via noncovalent interactions must play an important role in critical areas as genetic code, biological information storage, transfer biomolecule (protein, DNA, RNA, etc.) synthesis, and so on. The formation of duplex DNA from its single-stranded constituents, the stabilization of the high-ordered hairpins loops in RNA, and the functional self-assembly of protein-nucleic acid complexes are a result of a large collection of intermolecular forces. These forces include hydrogen bonding, aromatic π-stacking, charge interactions, van der Waals forces, or hydrophobic effects. Moreover, the high fidelity observed in these self-organized biomacromolecular architectures is largely caused by the high selective molecular recognition processes of natural base-pairing interactions via Watson-Crick H-bonding or of specific protein folding via amide H-bonding, and so on. Among these systems, the nucleobases (1–5) and the nucleosides (6, 7) as well as DNA or RNA (8–10) are well-known, fascinating compounds with a high ability to form controlled multiple intermolecular H-bonding of complementary nature, −C-H–O, hydrophobic, and stacking interactions.

The adenine-uracil interaction that involves two hydrogen bonds ($K_a \approx 10^2 - 10^3 \text{ M}^{-1} \text{ CDCl}_3$) is weak and nonspecific compared with the guanine-cytosine interaction, which involves three hydrogen bonds ($K_a \approx 10^3 - 10^5 \text{ M}^{-1} \text{ CDCl}_3$) that are usually paired via Watson-Crick interactions (4, 5). Homopairing and heteropairing of adenine-uracil and guanine-cytosine derivatives, which result in the formation of interconverting dimers, trimers, and oligomers via the combination of H-bond pairings, seem inadequate to function in any predefined recognition scheme. Amazingly, a very diverse set of interconverting supramolecular entities (oligomers) may be generated by using only these four nucleobases.

Their remarkable self-association properties, via Watson-Crick and Hoogsteen pairing, play a critical role in the stabilization of higher-order RNA hairpins loops, double or triple helix DNA, and G-quartets or G-quadruplexes (3–6). Even though the Watson-Crick (WC) base-pairing is prevalent in natural systems, other H-bonding motifs are present in natural and artificial systems; these motifs include: reverse Watson-Crick (rWC), Hoogsteen (H), reverse Hoogsteen (rH), Wobble (Wo), or reverse Wobble (rWo) (4, 5).

During the last decades, several studies reported the preparation of synthetic discrete supramolecular assemblies (4–8), polymers (1–9), and hybrid materials (10–12) that possess bases of nucleic acids as side groups or chain-end, which are used as precursors to conceive self-organized hybrid materials at nanometric scale.
Nanosized supramolecular materials have received increasing attention during the last two decades (15-19). The supramolecular synthesis provides a powerful tool for the noncovalent generation of such functional supramolecular architectures (15). The supramolecular polymers offer solutions for material molding at the macroscopic level, but their manipulation at the molecular (supramolecular) and nanoscopic levels is still difficult to control (16). This finding represents a nice extension from material science to biologically interesting component molecules such as nucleic acids that could be of interest for complementary binding (sewing) of nucleic acid strands.

For all these reasons, this article will highlight some recent accomplishments in the field of self-organized hybrid materials, and it will focus on the evolution of discrete nucleobase derivatives from self-assembled dynamic libraries of different devices exchanging in solution to constitutional functional hybrid solid materials. The primary purpose of this review is to describe our recent work on nucleobase-based self-organized hybrid materials. The article begins by describing some recent advances in the area of hybrid supramolecular materials. Then, constitutional dynamic amplification of supramolecular architectures will be examined with particular emphasis on self-organized nucleobase-type systems, presenting combined features of structural adaptation in a specific hybrid nanoevironment (11). The last part will emphasize the assembly behavior of G-quadruplex superstructures (6, 7) and the transcription of structural information in hybrid materials of functional complexity (12). The final structures of the hybrid materials develop solely from a balance of thermodynamic and kinetic factors as opposed to an iterative processing on a ribosome.

Hybrid Supramolecular Materials

Hybrid organic-inorganic materials produced by sol-gel process are the subject of various investigations, which offer the opportunity to achieve nanostuctured materials first from robust macrotemplates during sol-gel process on their supramolecular surface. Silsesquioxane-based precursors, in which the functional organic and siloxane inorganic groups are linked covalently, are employed extensively for the controlled generation of self-organized materials. Rigid aromatic molecules (9), which are urea H-bonding ribbons (20-23), are used to transpose a supramolecular self-organization in a siloxane matrix by a sol-gel process.

Despite such impressive progress, considerable challenges still lie ahead and the more significant one is the “dynamic marriage” between supramolecular self-assembly and the polymerization process, which might communicate kinetically and sterochemically to convert to supramolecular self-organization and functions in hybrid materials. The weak supramolecular interactions (H-bonds, coordination, or van der Waals interactions, etc.) that position the molecular components give the supramolecular architectures typically less robust than the cross-linked covalent bonds formed in a specific polymerization process. Accordingly, the sole solution to overcome these difficulties is to improve the binding (association) efficiency of molecular components that generate supramolecular assemblies. At least in theory, an increased number of interactions between molecular components and the right selection of the solvent might improve the stability of the templating supramolecular systems, which communicate with the inorganic siloxane network (35).

Three heteroditopic nucleobase-type silsesquioxanes, A<sub>12</sub>, U<sub>12</sub>, G<sub>12</sub> receptors have been reported by our group recently (Fig. 1, 11, 12). They generate self-organized continual superstructures in solution and in the solid state based on the following encoded features: 1) specific molecular recognition, 2) the supramolecular H-bond directing interactions, and 3) covalently bonded triethoxyxilyl groups that allow by sol-gel processes to transpose the solution self-organized dynamic superstructures in the solid heteropolysiloxane materials.

Constitutional Self-Organization of Adenine-Uracil-Based Hybrid Materials

As suggested in the introduction, the nucleobase building blocks generate a very complex dynamic pool of oligomeric ribbon-type or cyclic supramolecular architectures that exchange in solution when simple molecular precursors are used (Fig. 3). The A<sub>12</sub> and U<sub>12</sub> molecules were designed as rigid H-bonding modules. For instance, by introducing bulky blocking alkoxysilanepropylcarboxamido groups in N9 (A) and N1 (U) positions,
we limit only the Watson-Crick and the Hoogsteen interactions as preferential H-bonding motifs. The A_{Si} and U_{Si} precursors generate self-organized superstructures based on two encoded features: 1) they contain a nucleobase moiety that can form ribbon-like oligomers via the combination of H-bond pairings; 2) the nucleobase moiety is bonded covalently to siloxane-terminated hydrophobic groups packing in alternate layers.

Figure 1. (a) Supramolecular \( \mathbb{A} \) and supramacromolecular \( \mathbb{B} \) polymers resulted from intermolecular and intermacromolecular self-organization. (b) cross-linking and multiple outputs generation after the polymerization of monomers in supramacromolecular polymers (11).

Figure 2. Molecular structures of nucleobase ureido-silsesquioxanes A_{Si}, U_{Si}, and G_{Si}.

The dynamic self-assembly processes of such supramolecular systems (Fig. 3) that undergo continuous reversible exchange between different self-organized entities in solution may in principle be connected to a kinetically controlled sol-gel process to extract and to select an amplified supramolecular device under a specific set of experimental conditions. Such "dynamic marriage" between supramolecular self-assembly and in sol-gel polymerization processes that might communicate synergistically leads to "constitutionally-driven hybrid materials."

Based on this structural information and on the crystal structures of similar alkynucleobase derivatives, the relative arrangement of molecules of A_{Si} and U_{Si} in powders and in the M_{Si}, M_{U} and M_{Si}−M_{U} hybrid materials is similar to analogous ureidocrown-ether (20-23) and ureidoarene (36) superstructures. The generation of hybrid materials M_{Si}, M_{U}, and M_{Si}−M_{U} can be achieved using mild sol-gel conditions. In a typical procedure, we have prepared solutions of precursors A, U, or an equimolar mixture of A:U (1:1, mol/mol) in acetone, then we added deionized water and benzylamine as a catalyst. The mixture was kept at room temperature under static conditions for 15 days. The solvent was then evaporated at room temperature to yield the hybrid materials M_{Si}, M_{U}, and M_{Si}−M_{U} as the white powders.
Nucleobase Self-Assembly Codes, Constitutional Transcription of

Figure 3

Supramolecular dimers, trimers, and oligomers generated by H-bonding self-assembly of (a) adenine, (b) uracil, and (c) adenine-uracil base-pair. (R = sugar, alkyl, etc.)

The X-ray powder diffraction (XPRD) experiments presented in Fig. 4 show that well-defined, long-range order is present in the precursors \( \text{A}_0 \) and \( \text{U}_0 \). The long-range order is less pronounced than in the precursor materials in the hybrid materials \( \text{M}_0, \text{M}_1, \) and \( \text{M}_2 \) after the sol-gel step: less well-defined peaks are present than for the precursor and the average peak width increases, which indicates smaller domains in which coherent scattering occurs (37). The small angles XRD pattern of the precursor \( \text{A}_0 \) presents two well-resolved Bragg diffraction peaks that correspond to two crystallographically distinct phases: \( \text{A}_0\text{WC} \) and \( \text{A}_0\text{H} \) oligomers (Fig. 4a). A freshly synthesized solid sample of \( \text{A}_0 \) is crystallized predominantly as Watson-Crick-Hoogsteen \( \text{A}_0\text{WC} \) oligomer. A second nonpredominant polymorph of the all-Hoogsteen \( \text{A}_0\text{H} \) oligomer is present in powder as a
Figure 4: Constitutional (a) Hoogsteen packing of the Adenine-M₈ hybrid material; (b) Watson-Crick packing of the Uracil-M₈ hybrid material, and (c) A₂WCU₂H packing of the Adenine-Uracil-M₈ hybrid material.
result of breaking of Watson-Crick H-bonds and of creating the new Hoogsteen H-bonds (38). The small angles XRPD pattern of hybrid material $M_0$ presents a unique, well-resolved Bragg diffraction peak (Fig. 4a), which corresponds to a crystallographically distinct and unique all-Hoogsteen phase ($A_{2wc}$). Adenine (39) 9-methyladenine (40) and 9-ethyladenine (41) crystallize through the formation of unique Hoogsteen H-bonds in two-dimensional layers. These layers, which are stratified alternatively, exhibit two types of interfaces in between: one contact surface because of the $\pi-\pi$ stacking of adenine tapes and other surface that results from hydrophobic interactions of alkyl groups, which are in van der Waals contact. Similarly, the structure of the $A_n$ oligomers is most likely dictated by hydrophobic interactions between the grafted ethoxysilanepropylcarboxamide groups (Fig. Sb). In a freshly prepared $A$ sample, the solvent logged between hydrophobic groups favors the extended ($A_{2wc}$) oligomers. The condensation process between the ethoxysilane groups during the sol-gel process do favor the more compact ($A_{2w}$) oligomers in which the hydrophobic groups are interlocked, which stabilizes the interaction between Hoogsteen H-bonded layers.

The small angles XRPD patterns of the precursor $U$ as well as of the hybrid material $M_0$ present one Bragg diffraction peak (Fig. 4b), which corresponds to a characteristic Watson-Crick ($U_{2wc}$) dimer. The two other possible structures, the reverse Watson-Crick dimer ($U_{2hc}$) and the quartet ($U_4$) (see Fig. 3), do not correlate with the experimental distance.

A freshly evaporated solid sample of an equimolecular mixture in acetone of $A$ and $U$ presents two Bragg diffraction peaks that correspond to $A_{2hc}U_{2h}$ and $A_{2w}U_{2wc}$ oligomers, respectively. The small angles XRPD pattern of the hybrid material $M_{A-U}$ presents a unique Bragg diffraction peak that corresponds to a characteristic interplanar distance of $A_{2wc}U_{2w}$ oligomer (Fig. 4c). Amazingly, the unique structure of the resulting hybrid material $M_{A-U}$ that corresponds with experimental results is consistent with the formation of the Hoogsteen base pairing between $U$ and $A$ and with the formation of the Watson-Crick base pairing between two adenine
Figure 5  Toward a constitutional transcription of base-pairing codes in hybrid materials: (a) Guide to the eye interplanar $d_{\text{Si-Si'}}$ distances calculated from the geometry of minimized structures versus experimental interplanar Bragg diffraction distances. The squares correspond to the unpolymerized powders of precursors A, U, and their 1:1 mixture $\text{AU}_{\text{mix}}$, whereas the circles correspond to hybrid materials $\text{MA}$, $\text{MU}$, and $\text{MA-U}$. (b) Postulated model of self-organization of parallel $\text{H}$-bonded nucleobase aggregates and hydrophobic propyltriethoxysilane layers.

Nucleobase oligomerization by $\text{H}$-bonding can be an advantageous choice to reinforce the controlled communication between interconnected “dynamic supramolecular” and “fixing siloxane” systems. Moreover, the hydrophobic interactions can play an important role to stabilize compact packed superstructures that may be “extracted selectively” under the intrinsic stability of the system or external stimuli by polymerization in solid state.

Amplification and Transcription of the Dynamic Supramolecular Chirality of the G-Quadruplex

G-quartets are formed by the hydrogen bonding self-assembly of four guanosine and are stabilized by alkali cations. They play an important role in biology and in nucleic acid telomers in particular, which are of potential interest to cancer therapy (6, 7). The role of cation templating is to stabilize by coordination to the eight carbonyl oxygens of two sandwiched G-quartets: the G-quadruplex is the columnar device formed by the vertical stacking of four G-quartets.

The G-quartet architecture represents a nice example of dynamic supramolecular system that has been used as building molecules. Early contributions by Etter et al. (42), Castellano et al. (43), and others have been confirmed recently by Zimmermann calculations (44) for nearly exclusive preference for Hoogsteen binding, inside the classic A-U base-pairing, within 1:1 base-pairing complexes between alkyladenine and alkylthymine derivatives. Factors that contribute to the preference for Hoogsteen geometry are the shorter CH—O contacts, which results in a favorable alignment of the dipoles and is a greater distance between secondary repulsive sites (43, 44). Although many have recognized that in the solid state, the uracil or thymine preferentially bind adenine through Hoogsteen binding; therefore, this result is particularly interesting because the $\text{A}_{\text{mix}}\text{U}_{\text{mix}}$ oligomer is amplified quantitatively from a dynamic pool of oligomers in solution via sol-gel transcription.

As a general rule, which is proved by the differences between the values of interplanar Bragg diffraction distances, the condensation process between the ethoxysilane groups during the sol-gel process results in the formation of the more compact hybrid materials $\text{MA}$, $\text{MU}$, and $\text{MA-U}$ compared with the unpolymerized A, U, and AU mix powders (Fig. 5).
Figure 6  (a) The cation-templated hierarchic self-assembly of guanine alkoxysilane gives the G-quartet and the G-quadruplex (b) transcribed in solid hybrid materials by sol-gel in the presence of templating K⁺ cation.

Nucleobase Self-Assembly Codes, Constitutional Transcription of

block for gelators (45), columnar polymeric aggregates (46), self-organized surfaces (47), prototypes of chemical dynamic devices (48), and so on.

In the last decades, G-quartets (6) and a similar folic acid quartet (49) have been proposed as powerful scaffolds for building synthetic ion channels. Although stable in organic solvents, they do not seem to have defined transport functions in hydrophobic membranes. Barrel-stave-, lipophilic calix(4)arene-8-aromatic-guanosine conjugates have been used to stabilize the formation of G-quartets (7). Very recently, new strategies based on reversible metathesis were used successfully by Kaucher et al. (50) to generate a rich array of interconverting ion-channel conductance states of a unimolecular G-quartet in a phospholipid membrane. Polymeric guanosine hydrogels that can be interconverted reversibly between gel and sol states may be used to synthesize adaptive functional nanostructures (45).

Despite such impressive progress, considerable challenges still lie ahead, and the more significant challenge is to improve the stability of G-quartet dynamic aggregates in polymeric devices such as films or membranes to extend (address) the transport studies at macroscopic level. Several studies reported the preparation of discrete supramolecular assemblies of nucleobases. However, the “dynamic communication” between the
supramolecular self-assembly of nucleobases and the polymerization processes, which kinetically and sterochemically might communicate, is not so trivial.

On the other hand, the G-quadruplex with a twisted supra molecular architecture represents a nice example of a dynamic chiral supramolecular system, when guanine and guanosine molecules are used. Molecular chirality may be used as a tool to assemble molecules and macromolecules into supramolecular structures with disymmetric shapes. The supramolecular chirality, which results from both the properties and the way in which the molecular components associate, is by constitution dynamic and therefore examples of large-scale transcription of such virtual chirality remain rare.

For all these reasons, the guanine building blocks and the sol-gel chemistry were used as molecular precursor to conceive hybrid chiral materials at nanometric and micrometric scales. Our efforts involved the synthesis and the self-assembly of a guaninesiloxane monomer \( \text{G}_{x} \text{O}_{y} \) in the G-quartet and G-quadruplex supramolecular architectures (Fig 6), which are fixed in a hybrid organic-inorganic material by using a sol-gel transcription process, followed by a second inorganic transcription in silica, by calcination.

The generation of G-quadruplex hybrid material can be achieved by mixing \( \text{G}_{x} \) derivative with potassium triflate in acetone, followed the sol-gel process performed at room temperature using benzylamine as catalyst. Then, the hybrid materials were calcined at 400°C to transcribe their superstructural features into inorganic silica replica materials.

We have observed a long-range amplification of the G-quadruplex supramolecular chirality into hybrid organic-inorganic twisted nanorods followed by the transcription into inorganic silica microsprings (12). We believe that in the first sol-gel step, the polycondensation reactions of the inorganic alkoxysilane network take place around tubular twisted superstructure of G-quadruplex. The dynamic G-quadruplex is fixed in a covalently bonded siloxane network, and the structural (constitutional) memory of G-quadruplex is transcribed in the hybrid materials. These fixed ("frozen") objects are chiral and self-correlate with a helical order to generate anisotropic mesophases interconnected via condensed siloxane bridges. We obtained by sol-gel process a hybrid material that features a twisted hexagonal rod-like morphology of about 2 µm length and 350–850 nm diameter (Fig 7). The mixture of these entities contains left and right twisted nanorods, which are a result of the nonpreferential disymmetric orientation of the G-quartets. They are chiral, and no inversion centers have been observed within the same entity. Amazingly, these materials are at nanometric or micrometric scale topologically analogous to its G-quadruplex supramolecular counterpart. Similar "communication processes" have been identified in the DNA transcription into inorganic materials (10).

After the sol-gel process, the preformed helical silica network has embedded probably enough chiral information to be amplified (reinforced) irreversibly during the calcination process when almost total condensation of Si–OH bonds occurs. By calcinations of the hybrid material, the templating twisted G-quadruplex architectures are eliminated, and inorganic silica anisotropic microsprings are obtained. They present the same helical topology, without inversion inside the helix. These objects have a different helical pitch, which depends strongly on...
the self-correlation between hexagonal twisted mesophase domains at the nanometric level.

Our findings showed a new way to transcribe the supramolecular chirality of a dynamic supramolecular architecture; the transfer of the supramolecular chirality of G-quadruplex at the nanometric and micrometric scale is reported, thereby creating nanosized hybrid structures or micrometerized inorganic superstructures, respectively. Moreover, we obtain chiral materials by using a starting achiral guaninesiloxane \( G_\text{G} \), as precursor of achiral G-quartet and of chiral supramolecular G-quadruplexes. Figures 7 and 8 represent the first pictures of the dynamic G-quadruplex transcribed at the nanometric level.

Finally, our results show a new way of embedding supramolecular chirality in materials, which is of interest for the development of a supramolecular approach to nanoscience and nanotechnology toward systems of increasing functional complexity.

Conclusions

The nanometric or the micrometric transcription of the supramolecular functional devices, although marked significant achievements, represent a young field. Whereas many elegant functional systems based on nucleobase self-assembly have been prepared in recent years, it is almost clear that the synthetic efforts might be developed to extend and to understand the key features of such self-organized systems and the nanometric and micrometric levels. As importantly, this work established that molecular precursors could be used to affect molecular recognition in organic-inorganic hybrid materials, differently to those found in typical biological DNA and RNA systems. Nanometric self-organization using molecular nucleobases is clearly different from biological nucleotide self-assembly. These initial hybrid nanomaterials presented in this paper paved the way to the development of new functional macroscopic materials that contain nucleobases as structural building blocks.

Some systems can be applied in the fields of ionics, electron, and energy transport studies (9) and in sensing technology. Finally, our results show a new way of embedding nucleobases self-assembly and supramolecular chirality of G-quadruplexes in hybrid materials, of interest for the development of a supramolecular approach to nanoscience (51). Likewise, the hybrid polymeric arrays with nucleobases functionalities represent an area where the best is surely best to come; it unlocks the door to the new materials world paralleling that of biology.

Acknowledgment

This work, which was conducted as part of the award “Dynamic Adaptable M materials for Separation and Sensing Mecysystems” made under the European Heads of Research Councils and European Science Foundation EURYI (European Young Investigator) Awards scheme in 2004, was supported by funds from the Participating Organizations of EURYI and the EC Sixth Framework Programme. See www.esf.org/euryi/.

I thank the past and the present members of my laboratory with whom we have collaborated in these studies: Carole Arnal-Herault, Mathieu M. Michau, Andreea Pasc, Arié van der Lee, Eddy Petit and Didier Cot.

References


Further Reading


Smell is an intriguing sense in many ways. Although it is not the dominant sense in humans, it has effects on emotion and memory which can be very profound. Odor has long baffled chemists who try to model it in mechanistic terms, and a great deal of research has been invested in trying to understand how various features of molecular structure provide different odor characters and intensities. Many models for prediction of odor properties from chemical structure exist, but none are accurate and precise consistently. The reasons for this are now becoming clearer. The greatest breakthroughs in our understanding of olfaction came with the discoveries of the gene family coding for the olfactory receptor proteins (1) and the combinatorial mode of operation of these receptors (2). Odorous molecules enter the nose either from the front, by inhalation, or from the back, by diffusion from the mouth and respiratory tract. In the nose, they interact with an array of receptors. Each receptor type responds to a range of odorants, and each odorant stimulates a range of receptors. The signals thus generated are coded onto centers, known as glomeruli, in the olfactory bulb. The olfactory signals pass on upward through the brain and are interpreted finally, in the cortex, as odor. During the neurotransmission process, many interactions occur between olfactory signals and cortical feedback mechanisms, which include effects from other senses. Therefore, it is not surprising that simple structure/odor correlations are so elusive.

Biologic Aspects of Olfaction

The process of olfaction begins when odorant molecules enter the nasal cavity either by inhaled air or by diffusion from the mouth. In the nose, these molecules activate receptors in the olfactory epithelium. The signals from the olfactory receptor cells are picked up by the olfactory bulb where mapping of receptor types onto glomeruli occurs. The signal patterns leave the olfactory bulb, travel upward through the brain, and are interpreted eventually as odor in the cortex. Within tens of milliseconds of presentation of an odorant to the nose, olfactory signals are generated in the epithelium; cortical activity is detectable within 100 milliseconds. Figure 1 shows the location of some key olfactory organs in humans.

What is odor?

Smell probably is the oldest of our senses. The ability to detect and to recognize specific chemicals gives even single-celled organisms useful information about their environment. For example, it is known (3) that sperm will swim toward certain odorous substances. The ability to analyze their environment...
by chemosensation allows organisms to identify opportunities and threats from chemical changes in their environment; therefore, chemosensation has an obvious survival value. Continual refinement of the mechanism throughout evolution has led to the highly developed sense that we humans know as smell. In this refined sense, usually odors are not derived from single chemical entities but from complex mixtures, and each mixture is perceived as a separate odor image (4, p. 63). Traditionally, the sense of smell is associated with the nose because that is where the signals are detected and research has tended to focus on the receptor events. However, it is now clear that the phenomenon of odor exists only in the higher brain and is a synthesis of inputs from the olfactory receptors, other senses, and various cortical feedback mechanisms. As stated by Wilson and Stevenson (4, p. 34), "With a relatively few exceptions, neither odor physico-chemical feature extraction at the receptor sheet, nor spatial maps of those features in the olfactory bulb, nor simple convergence of those features in cortical circuits are sufficient to account for the rich experience that is olfaction."

The nose

We tend to talk about the nose as a single entity; yet a septum exists that divides it into two physically separate cavities. On the roof of each of these cavities, and extending down onto the septum, is a patch of tissue known as the olfactory epithelium. Odorants can reach the epithelium either from in front, by inhalation of air from the environment (referred to as the orthonasal route), or by diffusion from the mouth and respiratory and gustatory tracts (referred to as the retronasal route). The latter is vital in flavor detection as the tongue contains receptors only for sweet, sour, salt, bitter, and umami tastes; the rest of what is referred to as taste, actually, is retronasal smell. On the sides of the nasal cavities are bony plates called turbinates that cause turbulence in the air flow and ensure that odorants reach the epithelium. The air flow through each nostril is always different because one is faster than the other; the fast flow alternates continually from one nostril to the other (5). This alternation provides a mechanism for increased sensitivity and discrimination based on transport phenomena because the timing and pattern of signal spread across the epithelium will depend, in part, on the air flow. It has been
shown that information on recognition features is exchanged at a higher level in the brain because a molecule whose odor has been learned using one nostril only will be recognized by the other nostril (6).

The olfactory epithelium and the olfactory receptors

The olfactory epithelium is a patch of greenish-yellow tissue several square centimeters in area and 100–200 micrometers thick. It contains receptor cells that run from the nasal cavity through the base of the skull (the cribriform plate) into the olfactory bulb. On the side of the nasal cavity, the receptor cells have hairs or cilia that are 20–200 micrometers long. These cilia are bathed in a mucus layer that is 35 micrometers thick and flows backward continually at a rate of 3–60 impulses/second; this rate of firing is increased when the cells are stimulated by an odorant.

The mucus also contains cytochrome P450s, which are oxidative enzymes, and proteins known as odor-binding proteins. The odor-binding proteins belong to the family of lipocalins. The role in olfaction of these two classes of proteins is not certain. It is possible that they serve only to remove excess odorants and therefore contribute to provision of the time-dimension of olfaction, which is an important feature in survival. However, other roles have also been postulated and will be discussed below.

The olfactory receptor proteins (ORs) belong to the family of 7-transmembrane G-protein coupled receptors (GPCRs). The gene family coding for the receptor proteins is the largest in the genome, which contains codes for over 1000 proteins (1). Most mammals express 800–900 of these proteins, but humans use a much reduced set, which contains only 350–400 proteins. Interestingly, the only other mammals (gorillas, chimpanzees, orangutans, and rhesus macaques) with such a restricted set of proteins are the only other mammals to have color vision. Variation between humans is such that statistically it is unlikely that any two humans use the same set of olfactory receptors (7).

Events that follow receptor activation involve the normal train of G-protein (G-off, in the case of olfaction), second messenger (cAMP or IP3, in the case of olfaction), and ion channel chemistry, which leads to polarization of the cell and hence generation of a discharge at the synapse that leads to the olfactory bulb (8). Intriguingly, it has been shown that the signal pattern elicited by a mixture of two substances is not necessarily a simple additive of the signals elicited by the two components individually, but receptors that are not activated by either component can be activated by the mixture (9).

The olfactory bulb

The olfactory bulb sits at the base of the brain on the cribriform plate. Humans have two olfactory bulbs, the right bulb that receives signals from the left olfactory epithelium, and the left bulb that receives signals from the right olfactory epithelium. In the bulb, centers called glomeruli exist, each of which receives signals from only one type of receptor cell, irrespective of where those receptors are located on the epithelium. Olfactory receptor cells have short lifetimes (about 2 weeks), and new receptor cells develop in such a way that each makes its connection with the correct glomerulus. Receptor types with similar substrate selectivities tend to be associated with glomeruli that are found close to each other in the olfactory bulb.

Higher processing

Olfactory signals that leave the olfactory bulbs travel by several routes to the higher centers of the brain where the phenomenon of odor develops eventually. Remarkably, the architecture of the olfactory parts of the brain is consistent across all mammalian species; research on other animals throws considerable light on the function in humans. The entire process is well reviewed by Wilson and Stevenson (4) and by Delano and Sobel (10).

The signals travel initially to the piriform cortex and the amygdala; hence, both the thalamus and the limbic system are involved. It is postulated that the direct link to the limbic system accounts for the influence of odor on memory and emotion. From the piriform cortex, signals go to the amygdala, then to the thalamus, and directly to the orbitofrontal cortex. Processed signals from the thalamus also go to the orbitofrontal cortex. This structure also receives signals directly from the amygdala. Many of these signal pathways are two-way in nature with signals that come down from higher centers that affect the ascending signals. This accounts for the well-known effects of experience, expectancy, and context in distorting odor perception. A classic example is the inability of experts to describe the aroma of a white wine correctly to which a tasteless red dye has been added (11). Interference from other senses is also important to determine the ultimate odor percept. For instance, 70% of odorants also stimulate the trigeminal system in the nasal cavity and visual input has been shown to affect olfactory signal processing (12). Essentially, recognition of “odor” is a process of matching pattern of this final signal combination against reference patterns stored in the brain. For example, in perfumery, it is well known that development of an odor language is essential to train perfumers, and their powers of discrimination improve as their language improves.

Anosmia

Anosmia, the inability to smell, can be divided into two classes. General anosmia, the inability to smell any odors at all, usually is the result of disease or accident. More common is specific anosmia, in which an individual cannot detect a specific chemical substance that most people can detect or displays a threshold of detection for it which is significantly above the normal range. At one time, specific anosmias were linked to the concept of primary odors (13), but confirmation of the combinatorial mechanism of olfaction has put paid to this concept. Interestingly, it has been demonstrated that exposure to the substance can affect anosmia and individuals can begin to smell materials to which they were previously anosmic. This effect has been demonstrated for androstenone, amyl acetate, geranyl nitrile, and isoborneol (14–18).
Chemical Aspects of Olfaction

For various reasons, it is very difficult to develop an understanding of the chemistry involved in olfaction. The olfactory receptors are found in the membrane of the receptor cells; therefore, their active states are not amenable to structural determination by X-ray diffraction or other physical tools. Odor is a mental image rather than a physical property that can be measured and quantified. Odor perception is a multistep process. Olfaction is combinatorial in nature. All of these facts indicate that building of either substrate or receptor models are fraught with significant difficulties.

Olfaction, however, is a very obvious property of any chemical compound, and thus, speculation about the mechanism of perception is very tempting. Experienced fragrance chemists can predict odor type with much better than random accuracy, and a commercial driver exists in terms of design of novel materials for the fragrance industry. Therefore, it is not surprising that many structure/odor correlations and olfaction models have been reported and debated, often very hotly, in the literature.

Not all of the olfaction models in the literature are far too simplistic and too mechanical in nature, and none of them have succeeded in accounting for all of the observations about olfaction. As described, recent advances in our understanding have confirmed that odor perception, as predicted by Polak (19), starts with a combinatorial mechanism at the receptor level (1) and involves pattern recognition in the higher brain (4). No single odorant-receptor interaction will be the sole determinant of odor perception, and even knowledge of the pattern elicited at the olfactory bulb is insufficient to enable prediction of the cortical image of odor. Therefore, structure/odor models are and, for the foreseeable future, will remain statistical tools rather than mechanistic indicators.

Odor measurement

To understand olfaction at a chemical level, it is necessary to have good data that link chemical structure to odor properties. This task is much more difficult than it would seem, for instance, to a chemist who sniffs a sample that he has just synthesized in the laboratory and applies an odor descriptor to the molecular structure of his synthetic target. These difficulties stem from both chemical and sensory issues. Techniques for odor measurement and the difficulties involved have been reviewed by Neuner-Jehle and Etzweiler (20).

Chemical purity and organoleptic purity are not synonymous. For example, the aldehyde (Structure 1) was discovered when a sample of the alcohol (Structure 2) was found to have the expected muguet (lily of the valley) odor (21). The alcohol was prepared from 4-tert-amylcyclohex-1-yl)acetaldehyde.

\[
4\text{-tert-amylcyclohex-1-yl)acetaldehyde} \rightarrow 2\text{-tert-amylcyclohex-1-yl)alcohol} \rightarrow 2\text{-tert-amylcyclohex-1-yl)acetaldehyde}
\]

Figure 2: Synthesis of 2-(4-tert-amylcyclohex-1-yl)acetaldehyde.

A good example of subjectivity in odor character measurement is provided by O’Hoff et al. (22). When 27 panelists were asked to allocate the odor of the cyclic ether (Structure 5) to one of various odor categories, 14 participants described it as minty/amphicaceous; 6 participants described it as fruity; 3 participants described it as musky/woody. Therefore, classification as minty (based on the largest subject group) would only be correct for 50% of the panel. Similarly, it is easy to demonstrate that Bangalol (Givaudan, Vernier, Geneva, Switzerland) (Structure 6) is perceived by some subjects as sandalwood in character but by others as musk (23). Odor intensity is also subjective. For example, the average odor threshold for (-)-geosmin (Structure 7) was found to be one tenth that for the (+)-enantiomer (Structure 8). However, some individuals were
40 times more sensitive to one enantiomer than to the other, some experienced similar thresholds for both enantiomers, and some were more sensitive to the \( (+) \)-enantiomer (24). As with all sensory magnitude estimation, odor intensity measurement must take into account the fact that humans adjust mental scales unconsciously to suit the task in hand.

Odor classification is particularly difficult. For the senses of touch or sight, it is easy to pick physical reference points (hardness of standard substances, wavelength of light, etc.) and then to classify sensory properties in relation to these points. No such classification exists in odor. No primary odors and no physical reference points exist. Consequently, all odor classification is by comparison with other odors. For example, it might make sense to see apples and pears as subclasses of fruit in botanical terms, but in terms of their odors, putting apple and pear under the general heading of fruity odors leads to difficulties in structure/odor correlation (25). Indeed, the brain sees each new odor as a new percept rather than as a combination of existing percepts (4). Therefore, odor classification, although a useful tool in perfumery, is essentially meaningless scientifically. For example, mixing together in suitable proportions, hexylcinnamic aldehyde (Structure 9) (fatty odor), benzyl acetate (Structure 10) (floral odor), and indole (Structure 11) (fecal odor) will give a perfume that is recognizable as jasmine in character. However, a sample of pure cis-jasmone (Structure 12) would also attract the descriptor of jasmine. So, of what value is the term "jasmine"?

**Transport to the receptors**

It is self-evident that transport properties must be of importance in olfaction because if odorants cannot reach the olfactory epithelium, they will not be detected by the olfactory receptors and no odor will result. Volatility is the most obvious requirement, and for organic compounds, this requirement results in a cutoff point at about 18–20 carbon atoms in the molecule equivalent to a molecular weight of about 300 Daltons. Larger molecules are simply not volatile enough to reach the olfactory epithelium in sufficient concentration to be perceived. Solubility is also important; partially because water solubility implies a polar molecular structure, and this, in turn, implies a low vapor pressure relative to the molecular weight because of intermolecular hydrogen bonding. However, solubility properties per se also seem to be important, which is perhaps related to the ability of molecules to cross the aqueous mucus layer to reach the receptor proteins. Fragrance molecules generally have \( \log P \) in the region of 2–5.

**Odor-binding proteins**

Proteins that belong to the lipocalin family and are present in the olfactory mucosa were first identified as involved in binding of pyrazines and so were first named pyrazine-binding proteins (26). However, it soon became clear that they are capable of binding a very wide range of odorant molecules, and the name odorant-binding proteins (OBPs) was coined (27–30). Possible roles of odor-binding proteins include transport of odorants across the mucus to the receptor, signal attenuation, or removal of excess odorant, or it is also possible that the receptor proteins distinguish between free and liganded lipocalins rather than detecting free odorants. However, because no lipocalins are present in the experiments of Spehr et al. (1), it is clear that odorants can be detected by receptor proteins in the absence of OBPs. OBPs are more important to insects than to mammals. For instance, Drosophila melanogaster has about 50 different types of OBPs and 70 different types of ORs, whereas mammals have only about five types of OBPs and a potential pool of about 1000 different ORs (31). No kinetic studies have been carried out on OBP solubilization of odorants, but if simple model systems are relevant, then the work of Rebek on self-assembling clathrates (32–35) would suggest, by analogy, that the time required for trapping and release of odorants by OBPs might be rather long in terms of the total time involved in olfaction. Overall, insufficient information exists to enable a clear picture of the role of OBPs to be drawn.

**Cytochromes**

The most obvious role for oxidative enzymes in the olfactory mucosa would be removal of excess or spent signal material. However, some evidence suggests that products of cytochrome oxidation are detected by the olfactory receptors and thus constitute a part of odor perception (36).

**The receptor event**

Based on homology models in humans and in mice, Lancet and coworkers have suggested that odorant binding to olfactory receptor proteins occurs in the transmembrane part of the receptor protein, which bridges between amino acid residues of helices 2 to 7 (37, 38). Interestingly, this region is the same region in which 11-cis-retinal binds in rhodopsin and also corresponds with the binding sites of other GPCRs (39). Furthermore, it is also in broad agreement with binding sites proposed in several reported model studies on specific odorant/olfactory receptor binding (vide infra).

The olfactory receptors are tuned broadly because each receptor type responds to a range of odorants, and each odorant fires a range of receptor types (16). For some receptor/protein/odorant combinations, the binding affinity is concentration dependent (40, 41), and this can correlate with observed changes in character as concentration varies (2, pp. 71–74).
Structure/activity approaches

The relationship between molecular structure and odor has been the subject of much research during the last 150 years. The motivation has come partly from commercial interest to learn how to design improved ingredients for fragrances and partly from the scientific interest to understand the sense of smell. Attempts to translate structure/odor models into mechanism and vice versa have led to much confusion and to very heated debates. It is now clear that structure/odor models are useful tools to aid the design of novel odorants but that understanding of the mechanism of olfaction will come from biochemistry, molecular biology, and neuroscience rather than from these models. Conversely, developments in our understanding of receptor events in olfaction will not necessarily improve the structure/odor models.

The structure/activity relationship (SAR) tools employed in odor research are essentially the same standard tools used in all applications, and the models developed fall into the categories of substrate and receptor models. The pharmaceutical industry is the leader in SAR techniques, and the fragrance industry tends to follow its lead. Early models were substrate based, but the discovery of the genes that code for the olfactory receptor proteins has also allowed receptor models to be constructed.

Substrate modeling

For substrate models, classic chemical methods such as Hansch analysis are used, as are statistical techniques such as principal components analysis (PCA). Hansch uses regression analysis to correlate electronic, steric, and hydrophobic properties with the biologic activity in question, whereas PCA is a statistical technique that reduces multidimensional input (physical properties of molecules) to two or three dimensions that aids in correlation with the biologic activity. Molecular modeling tools such as COMFA (comparative molecular field analysis) and the olfactophore approach, the odor equivalent of pharmacophores, have also been used successfully. In these techniques, the stereoelectronic properties of a test set of molecules are used to build a model of an idealized substrate with which putative novel materials can then be compared. A comparison of Hansch analysis, PCA and COMFA in the correlation of structure with fruity odors provides a useful introduction to the three techniques and shows that, in this instance at least, they provide similar results (42). A good illustration of the use of the olfactophore approach is K. Hart and Eichenberger’s design of a novel marine odorant using the technique (43). The approaches used in structure/odor modeling have been comprehensively reviewed by Rossiter et al. (44), Frater et al. (45), and Kraft et al. (46).

The commercial driver for substrate models has been the search for novel fragrance ingredients, and odor has tended to be the defined biologic activity. However, the combinatorial mechanism of olfaction presents a serious obstacle for such substrate modeling and renders it almost meaningless in terms of understanding the mechanism of olfaction. Medicinal chemists who work on drug SARs usually are targeting a specific active site in a single protein. Fragrance chemists must target interactions of odorants simultaneously with the active sites of an unknown number of proteins because no single protein–single percept relationship exists. This lack of relationship can be the case even with animals simpler than humans. For example, it has been shown that the fruit fly Drosophila requires simultaneous activation of two different receptors to detect carbon dioxide (47).

Frustrating factors abound in structure/odor correlation. Sometimes, a small structural change in a molecule produces a large odor change whereas in other cases, gross structural changes produce little change in odor (44). Sometimes the chemical functional group in a molecule is important, as in the ester group and the fruity odor (25), whereas as other times the shape of the molecule is more important, as for the camphor odor (48). Absolute stereochemistry of a molecule sometimes affects its odor and sometimes it does not (49). Similarly, the effects on odor of a given change in molecular structure are unpredictable (50). In general, SARs suffer from the limitations of being interpolative and their reliability is proportional inversely to the number of steps in the process being modeled. In view of all of this and all of that has been said above about the process of olfaction, its combinatorial nature, its subjectivity, the difficulty in measuring odor and the fact that odor is a mental construction rather than a physical reality, it might be expected that structure/odor models would be impossible to find. However, many useful models do exist. The following accounts provide a few examples of useful models. For a more comprehensive list, the reader is referred to the reviews cited above.

The most consistently accurate structure/odor model is Amoore’s camphor model (48). As shown in Fig. 3, the model indicates that hydrophobic molecules with an ellipsoidal shape that have a long axis of 9.5 Å and a short axis of 7.5 Å will possess a camphoraceous odor. Another example of a simple but effective model is Boelens’s model for jasmine odors (51). Shown in Fig. 4, this model proposes that, to possess a jasmine odor, a molecule should contain a central carbon atom surrounded by a strongly polar group, a weakly polar group, and an alkyl chain.

The first published model for the ambrengis odor is that of Ohloff’s triaxial rule (52). This model proposes that, to possess an ambrengis odor, a molecule should have a trans-decalin structure with three axial substituents in a 1,2,4-relationship, as shown in Fig. 5 and that one of these should be an oxygen function. A more sophisticated model is that of Brusker et al., which is based on the electron topological theory of odor (53).

They propose that ambrengis odors have two hydrogen atoms

Figure 1

Amoore’s model for camphoraceous odorants.
that the charge on Hi should be negative, and that the charge atom should be between 0.24 and 0.31 of that of an electron, the molecule, that the coefficients of both of these hydrogen three atoms make a significant contribution to the LUMO of odorant successfully (61).

(55–60). Bajgrowicz et al. built an olfactophore model around hydrophobe and the distance between it and the alcohol function techniques to define more closely the exact requirements of the model has been refined by many workers that use different SAR center of hydrophobic bulk at a set distance from it. This basic suggests a model that involve s an alcohol function with a representative for decalin systems similar to the natural ambergris odor (54). Ohloff’s model is reasonably repre-
sions of which are shown in Figure 5.

and one atom located at the corners of a triangle, the dimen-
sions of which are shown in Fig 5. They also require that all three atoms make a significant contribution to the LUMO of the molecule, that the coefficients of both of these hydrogen atoms should coincide, that the negative charge on the oxygen atom should be between 0.24 and 0.31 of that of an electron, that the charge on H, should be negative, and that the charge density over the triangle should be $-0.1\text{e/Å}^2$. More recently, Bajgrowicz and Broger constructed an olfactophore model for the ambergris odor (54). Ohloff’s model is reasonably repre-
sentative for decalin systems similar to the natural ambergris chemicals. However, many ambergris odorants are known now for which this model in inapplicable. Therefore, the more recent models are of more use when studying the ambergris odor.

Even a cursory inspection of the known sandalwood odorants suggests a model that involves an alcohol function with a center of hydrophobic bulk at a set distance from it. This basic model has been refined by many workers that use different SAR techniques to define more closely the exact requirements of the hydrophobe and the distance between it and the alcohol function (55-60). Bajgrowicz et al. built an olfactophore model around this basic concept and used it to design a potent new sandalwood odorant successfully (61).

Receptor modeling

The ability to insert active olfactory receptors into functioning cells means that it is now possible to profile the activity of receptor proteins and move from the realm of gueswork into experimental reality. For example, Sans et al. have profiled and compared the selectivity of two human olfactory receptors that belong to different phylogenetic classes (62). They found that the class I receptor, OR51D1 has a relatively narrow range, which accepts alcohols, esters, ketones, and acids with a molecular length that corresponds to a chain of 8 or 9 carbon atoms; whereas the class II receptor, OR1G1, is much more broadly tuned and responds to a wide range of functional groups and with a preference for slightly longer carbon chains of 8 or 9 atoms. They also found that some odorants are capable of acting as antagonists and of blocking receptor activity from other molecules that would function normally as agonists. This finding is similar to that of Spehr et al. (1) who observed the same phenomenon with hOR17-4. Therefore, it seems likely that such antagonism is widespread and this will have implications for the perception of odor in mixtures. Because almost everything we smell is a mixture, such interactions are of considerable importance.

Models of the receptor sites are based on analogy with those of other GPCRs and of rhodopsin in particular. Using homology modeling, it would seem that the odorants are most likely to be bound in the regions between the 3-, 4-, 5-, and 6-transmembrane sections, and the region of rhodopsin is where 11-cis-retinal is bound. A good example of this approach is that of Pilpel and Lancet (63).

Several assumptions are made when building such models. For example, it is assumed that the tertiary structure adopted by rhodopsin in crystalline form is similar to that which it adopts when in the cell membrane; that olfactory receptor pro-
top's distance for receptor modeling, it would seem that the odorants are most likely to be bound in the regions between the 3-, 4-, 5-, and 6-transmembrane sections, and the region of rhodopsin is where 11-cis-retinal is bound. A good example of this approach is that of Pilpel and Lancet (63).

Several assumptions are made when building such models. For example, it is assumed that the tertiary structure adopted by rhodopsin in crystalline form is similar to that which it adopts when in the cell membrane; that olfactory receptor proteins adopt a similar tertiary structure to that of rhodopsin, and that ligand docking in olfactory receptors is similar to cofac-
top docking in rhodopsin. However, Vaidhevi et al. (64) have used MembStruk and HierDock software (which can be ob-
tained from William A Goddard III at Caltech, Pasadena, CA) to show that consistency exists between GPCRs of different types (mouse and rat olfactory receptors, the human sweet receptor, endothelial differential gene 6, and the $\beta$-adrenergic receptor) (64). They also showed that these modeling techniques work across this range of receptors and they used these models to predict the tertiary structure and binding site of rhodopsin with a reasonable degree of accuracy. Moreover, it is possible to test receptor models experimentally by comparing the predic-
tions of the model with the measured selectivity of the receptor in a live cell. Research in this area over the last decade has yielded significant results and we can be confident that we now have a reasonable concept of how olfactory receptors recognize their agonists.

One of the first exercises of this type was the work of Singer and Shepherd on the rat receptor OR5 which is known to respond to Lyral (International Flavors and Fragrances of Union Beach, New Jersey) (65). They proposed a model for the binding site of Lyral that involved transmembrane helices III to VII. However, a later model identified a somewhat different site using only helices IV, V, and VI (66). Floriano et al. investigated the mouse receptor OR525 (67). Using HierDock software, they identified binding sites and cal-
culated binding energies for 24 potential agonists. The binding site was found to involve 10 amino acids from transmembrane helices III-VI and the energies indicated that hexanol and hep-
tanol should bind most strongly. Experiments with the receptor in cells showed that, indeed, of the 24 test materials, only these two alcohols elicited a response. They then went on to screen an additional six mouse receptors (56, S1B, S19, S25, S18, S19, S25,
Also supported by mutation data. Extrapolating these results to examples studied, the selectivity of the receptors were determined largely by two of these positions. This latter result is consistent with the HierDock calculated binding energies and experimental receptor activation. They confirmed that the crucial transmembrane helices are III–VI and that extra-cellular loops II and III also contribute to binding. In the results from modeling were confirmed by materials remained in the binding site whereas others migrated out. Correlation with testing of the results was 100%. Those molecules found to be agonists whereas those that migrated out failed to activate the receptor in vivo. Moreover, the model elucidated the route into the binding site from the extracellular side of the protein. Similarly to the results from modeling were confirmed by in vivo testing of the receptors.

A new dimension in modeling was introduced by Lai et al. who built a dynamic model of rat receptor OR17 (72). They incorporated 10 potential aldehydic ligands into the binding site and then set the whole assembly into normal motion. Some test materials remained in the binding site whereas others migrated out. Correlation with in vivo results was 100%. Those molecules that the model predicted would remain in the binding site were found to be agonists whereas those that migrated out failed to activate the receptor in vivo. Moreover, the model elucidated the route into the binding site from the extracellular side of the protein. Similarly to the results from modeling were confirmed by in vivo testing of the receptors.

A consistent feature exists in all of these models, in that, for a good odorant/receptor fit, each model requires a polar group in the odorant that can form a hydrogen bond or similar interaction with a donor site in the receptor and the rest of the fit is determined by a spatial match with the shape of the (largely hydrophobic) binding pocket. Saturated hydrocarbons presumably lack the polar interaction and, in some cases at least, it would seem that weaker nonbonded interactions, such as π-stacking exist in the hydrophobic pocket.


Further Reading

See Also
G-Protein Coupled Receptor (GPCR) Signaling
Cytochrome P450 Monooxygenases, Chemistry of Neurotransmission, Chemical Events in Taste: Topics in Chemical Biology
Adaptation to different environmental temperatures establishes specific requirements on the stability of DNA and protein macromolecules. Organismal strategies of thermophilic adaptation, structure- and sequence-based, and their physical origins provide a consistent picture of the evolution of protein thermostability. A strong correlation between the optimal growth temperature (OGT) and the frequency of ApG dinucleotides in both sense and antisense strands of genomic DNA along with the absence of any “thermophilic” bias in the nucleotide composition highlights a key role of base stacking in the thermostabilization of the DNA double helix. The codon bias provides an excess of ApG pairs, which ensures the thermophilic adaptation of genomic DNA. The concentration of seven amino acids, Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL), serves as a universal proteomic predictor of the OGT prokaryotes. The IVYWREL combination manifests a generic “thermophilic” trend in amino acid composition: the increase of hydrophobic and charged residues at the expense of polar ones. This so-called “from both ends of the hydrophobicity scale” trend is a result of the positive (stabilizing the native state) and the negative (destabilizing misfolded conformations) components of protein design. The pressure to preserve energies of important native and non-native contacts results in a correlation in mutations of amino acid residues involved into these contacts. A comparison of energy (Myiazawa–Jernigan potential) and substitution (BLOSUM62) matrices reveals a high rate of substitutions between amino acids that strongly attract each other (native contacts) and between residues that strongly repel each other (non-native contacts).

What makes thermophilic adaptation so attractive for researchers from the very beginning of protein and DNA molecular studies (1, 2)? Although life exists in different extreme conditions, such as temperature, pressure, salinity, radioactivity (3), the adaptation to extreme temperatures is an outstanding phenomenon. Indeed, organisms belonging to one level of organization, prokaryotes, thrive under environmental temperatures that cover the entire range from −10 to +110 °C, one third of the absolute temperature interval. A significant difference in the optimal growth temperature of prokaryotes results in a distinct stability of their proteins and DNA, which makes them a central subject in the studies of mechanisms of molecular adaptation.

The thermostabilization of biomolecules is a result of the mutual contribution from fundamental interactions [e.g., hydrophobic forces (4, 5) or ionic interactions (3, 6, 7)] that stabilize individual molecules and prevent their aggregation (6), structural modifications [such as DNA superhelicity (8, 9) and posttranslational modification of proteins], interactions with an environment (10), intermolecular interactions (11), and oligomerization (12). The possible dependence of fundamental interactions, for example, hydrophobic forces, on temperature may also affect stability. However, it remains a subject of controversy as to how and to what extent the dependence of the interaction strength on temperature should be taken into account (13–16). This article reviews the very basic level of protein and DNA thermostability,
fundamental interactions, and their sequence/structure determinants.

**Basics of Protein and DNA Thermostability**

Various factors that contribute to protein thermostability, such as van der Waals interactions (17), core hydrophobicity (18–20), networks of hydrogen bonds (4, 5, 21), amount of secondary structure (4, 22), ionic interactions (6, 7, 23), packing density (24), and decreased length of surface loops (25), have been a subject of intense study for several decades. The major challenge, however, is to find out how the above factors are chosen and their combinations are formed by natural selection responding to the environmental temperature and depending on the evolutionary history of the organism (26).

Thermostabilization of double-stranded DNA is provided by base pairing (1) and base stacking (see Reference 27 and references therein) complemented by positive supercoiling by reverse gyrase (in hyperthermophiles (8, 9, 28)) and by stabilization via interactions with histone-like proteins (29, 30). The relative contribution of base pairing and base stacking into the thermostability of double-stranded DNA has been a subject of extensive studies for more than four decades (1, 27, 31). We will consider here this question, based on the results of recent experimental and computational works (31, 32).

**Sequence/Structure Signals of Thermophilic Adaptation**

Major “recipes” for increasing the thermostability observed in previous computational sequence/structure analysis and confirmed in experiment vary from the optimization of hydrophobic core interactions (18–20) to the introduction of additional ion pairs (7, 32). Respectively, thermophilic trends known so far include a large difference between the proportion of charged (DEKR) versus polar (noncharged, N/Q/G) residues (34, 35), an increase of long and branched side chain hydrophobic residues (36), an excess of some aromatic amino acid (35), and an over-representation of Pro (37). Recently, Satterson/mine et al. (37) illuminated the importance of structure dependence in the relationship between amino acid composition and optimal growth temperature (OGT) of the organism. In particular, the difference in amino acid frequencies between core and surface residues is getting more pronounced under higher temperatures (38), but not during adaptation to a cold environment (37). It also was demonstrated that amino acid biases in thermophilic adaptation are independent of the (G+C) content of coding nucleotide sequences, and the (G+C) content itself is not a determinant of the thermophilic adaptation of the double-stranded DNA (39–41).

**Experimental (Re)Design of Thermostable Proteins**

All experimental techniques of protein thermostabilization can be related to one of three major directions (40). First, the rational design concept is based on using previously known stabilizing factors. The limited predictive power of the rational design concept prompts one to test all potentially thermostabilizing mutations by using site-directed mutagenesis. The second approach is a directed evolution approach. Selective pressure or screening for a desired trait applied after random mutagenesis and/or DNA shuffling provides another possibility for engineering protein stability. Limited sequence space, amenable to testing in directed evolution, makes it necessary to eliminate more effectively the neutral and deleterious mutations, to increase the number of recombination events, and to improve the selection tools. Third, the “consensus concept” is based on the assumption that consensus amino acid contributes more to the stability of the protein chain than the nonconsensus amino acid at a given position in the alignment of the amino acid sequences.

**Physics and Evolution of Thermophilic Adaptation**

The tight connection between the “recipes” for thermostability immediately raises a question about the common evolutionary and/or physical basis for the variety of mechanisms of thermophilic adaptation. To address this question, one has to go beyond the analysis of specific stabilizing interactions and their various combinations. Conceptually, then, two major directions lead to the selection of proteins with high thermal stability. First, thermostable proteins have a structural bias such as enhanced packing. Second, stabilization is achieved by a few particularly strong, strategically placed interactions. The choice between these directions is affected by several evolutionary and environmental factors, and thermostabilization is a result of the intimate interaction between the physics of protein stability and the phylogenetics of the host organism.

**Physical Basics of Protein Thermostability**

Given a structural similarity between meso- and thermophilic orthologs, the variation of the stability across different proteins stems from the differences in the physical mechanisms of thermostability (3). Sequence/structure analysis and the unfolding simulations of hyperthermophilic proteins and their mesophilic homologues (43) reveal two major mechanisms of thermostabilization (Table 1). The first mechanism is “structure-based.” Some hyperthermophilic proteins are significantly more stable than their mesophilic homologues because of their high compactness (44). In this case, no single type of interaction is extremely strong and dominates stabilization, but the sheer number of interactions provides enhanced stability. Structure-based thermostability is nonspecific in the sense that no or minimal special features of sequences are needed to achieve thermostability, which makes it robust under a wide range of environmental conditions. A possible evolutionary disadvantage of such a robust stabilization mechanism is that it makes the protein less adaptable to rapid and specific changes in environmental conditions. The second mechanism is “sequence-based.” This way, several substitutions made in sequences of mesophilic proteins provide the formation of “staples,” that is, specific and strong...
Designability of structure is reflected in the certain properties of protein structure that indicates how many sequences can fold into that structure at various levels of stability (46, 47). Designability is a property of protein structure that correlates with the second order of C (Tr²C), of the contact matrix of a structure, C (44, 46). In particular, designability correlates with the compactness of a structure (number of contacts per residue). It was demonstrated (44) that more designable structures provide initial advantage because a greater number of sequences can fold into them with low energy. Therefore, a sequence search in the design of a thermostable protein will be less severe given a highly designable structure. A high contact density of LUCA domains (48) is suggestive, which shows that nature used high designability in the creation of the first thermostable proteins for ancient species. The role of designability in the design of ancient thermostable proteins is corroborated additionally by the high-throughput analysis of major folds (43). The Van der Waals contact density in hyperthermophilic archaea Pyrococcus is higher than in hyperthermophilic (T. maritima and A. aeolicus) or mesophilic (E. coli) bacteria. It indicates that on the organismal level, archaea used a structure-based mechanism and developed a respective strategy of thermophilic adaptation. What evolutionary scenario can one imagine for the emergence of another, sequence-based, strategy of adaptation? When mesophilic organisms recolonized in a hot environment, it was necessary to find a fast and effective way of tuning protein stability. To increase the stability of the protein without a redesign of the whole structure is possible via making sequence substitutions that introduce “staples,” a restricted set of a strong specific interaction (e.g., ion pairs). Hyperthermophilic bacteria (T. maritima and A. aeolicus), which recolonized in hot conditions, exemplify a sequence-based strategy. A high-throughput comparison of T. maritima and A. aeolicus proteomes with those of hyperthermophilic archaea shows the crucial role of sequence-based strategy in achieving the thermostability of proteins in hyperthermophilic bacteria (43). An analysis of the phylogenetic relationships between hyperthermophilic archaea and bacteria provides additional evidence for different organismal strategies of adaptation. 24% and 16% of the genes of T. maritima and A. aeolicus, respectively, were transferred to bacteria via lateral gene transfer (LGT) from archaea (49, 50), and corresponding bacterial proteins are the most similar to those of archaea. The importance of LGT in specific biochemical and environmental adaptations was demonstrated undoubtly by the comparison of complete genomes, codon analysis within genomes, and phylogenetic trees based on single gene families (see Reference 51 and references therein). Alternatively, it may be problematic to assess the relative contributions of LGT and vertical inheritance. For example, T. maritima and A. aeolicus belong to two lineages (Thermotogales and Aquificales) believed to have diverged earliest from the rest of bacteria. Therefore, it is possible that T. maritima and A. aeolicus retained ancestral genes and share some primitive features with archaea, whereas these genes were lost in the rest of the bacterial species. However, regardless of the
scenario working in Thermotoga and Aquilae (genes are received via LGT or, alternatively, are descendants of retained ancestral ones), the so-called "archaeal" parts of their genomes are reflective of the hyperthermophilic lifestyle and the distant evolutionary past (51). In particular, the archaean parts of the above bacterial proteomes (extracted according to the listing in the taxonomic distributions of the homolog TaxMap, available at http://www.ncbi.nlm.nih.gov) exhibit compositional features typical for structure-based strategy, whereas the bacterial parts follow a sequence-based strategy of thermophilic adaptation (43). Later events in protein evolution affected structures/sequences of both archaean and bacterial species which combine strategies of adaptation (52) or use complementary mechanisms of stability (53).

Genomics/Proteomics of Thermophilic Adaptation

A better understanding of how nature adapts life to elevated temperatures of the environment helps us to get a deeper insight into the basic physical laws that govern protein design. In particular, the explosion of information on crystalized proteins and complete genomes/proteomes makes it realistic to perform high-throughput analysis of sequences and structures. In this part, we will show how 1) alignments of proteomic sequences reveal a signal of a new entropic mechanism of thermostability, 2) exhaustive enumeration of all possible combinations of amino acid residues identifies a particular combination of them, which can serve as the best predictor of the optimal growth temperature (OGT) prokaryotes, and 3) correlation analysis which can serve as the best predictor of the optimal growth temperature of hyperthermophiles.

2) exhaustive enumeration of all possible combinations of amino acid residues identifies a particular combination of them, which can serve as the best predictor of the optimal growth temperature (OGT) prokaryotes (43). Later events in protein evolution affected structures/sequences of both archaean and bacterial species which combine strategies of adaptation (52) or use complementary mechanisms of stability (53).

Entropic Mechanism of Protein Thermostability

The compositional bias toward increasing charged residues in hyperthermophilic proteomes compared with mesophilic ones is well documented. However, the enrichment in positively charged residues is almost entirely because of lysines (34) (see Table 2). If only the total content of arginine (Arg) plus lysine (Lys) residues would matter in determining the stability of hyperthermophilic proteins, then no preference for the Lys over the Arg should exist. Arg and Lys are similar residues by their physical and chemical features; both residues are charged and have the same maximal number [81] of possible rotamers. An examination of the substitutions of types Arg/Lys versus Lys/Arg in the alignments of mesophilic sequences versus hyperthermophilic ones (Fig. 1) sheds light on the relationship between Arg and Lys content. The number before the slash (Table 3) is the percentage of amino acid residues in the mesophilic sequence, for example, Arg that was replaced by the other amino acid in the hyperthermophilic sequence, for example, Lys. The number after the slash reflects the same data for the opposite replacement, for example, Lys in the mesophilic sequence replaced by Arg in the hyperthermophilic sequence. Numbers in parenthesis show the ratio of forward to backward substitutions. The control groups are pairs Leu/Thr and Ser/Thr. Residues in each pair possess similar physical and chemical features (Leu/Thr are hydrophobic; Ser/Thr are polar), and both have the same maximal number of possible rotamers (9 and 3, respectively). In nine hyperthermophilic organisms, the pairs RK/KR demonstrate a remarkable bias toward the replacement of arginine in the mesophilic sequence with lysine in the hyperthermophilic sequence (Fig. 1) and hints to the specific role of lysine in protein stabilization. The complementary all-atom unfolding simulations show that lysines have a much greater number of accessible rotamers than arginines of similar degree of burial in folded states of proteins (53). Significant residual dynamics of lysine in folded states of proteins makes the entropic cost to fold lysine-rich proteins less favorable compared with arginine-rich ones. The arginine-to-lysine replacement stabilizes the folded state, preserving, however, the charged nature of the substitution.

Table 2 Percentage of charged amino acids and (G + C) content of 10 hyperthermophilic archaea (A), 2 hyperthermophilic bacteria (B), and mesophilic bacteria E. coli. A strong prevalence of lysine over arginine in proteomes of hyperthermophiles is obtained for nine organisms. A bold font marks the exception from the general trend

<table>
<thead>
<tr>
<th></th>
<th>EC</th>
<th>AA</th>
<th>AF</th>
<th>MJ</th>
<th>NE</th>
<th>PA</th>
<th>PH</th>
<th>PF</th>
<th>ST</th>
<th>TM</th>
<th>AP</th>
<th>MK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lys (K)</td>
<td>4.4</td>
<td>9.4</td>
<td>6.9</td>
<td>10.3</td>
<td>10.8</td>
<td>7.8</td>
<td>8.0</td>
<td>8.1</td>
<td>8.0</td>
<td>7.6</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>Arg (R)</td>
<td>5.5</td>
<td>4.9</td>
<td>5.8</td>
<td>3.9</td>
<td>3.9</td>
<td>5.7</td>
<td>5.6</td>
<td>5.3</td>
<td>4.2</td>
<td>5.5</td>
<td>7.8</td>
<td>8.3</td>
</tr>
<tr>
<td>Gin (D)</td>
<td>5.8</td>
<td>9.6</td>
<td>8.9</td>
<td>8.6</td>
<td>7.9</td>
<td>8.9</td>
<td>8.7</td>
<td>8.9</td>
<td>7.0</td>
<td>8.9</td>
<td>7.1</td>
<td>10.0</td>
</tr>
<tr>
<td>Asp (D)</td>
<td>5.1</td>
<td>4.3</td>
<td>5.8</td>
<td>5.5</td>
<td>5.0</td>
<td>4.6</td>
<td>4.4</td>
<td>4.4</td>
<td>4.6</td>
<td>5.0</td>
<td>4.2</td>
<td>5.8</td>
</tr>
<tr>
<td>G + C content</td>
<td>50.8</td>
<td>43.5</td>
<td>48.6</td>
<td>31.4</td>
<td>31.6</td>
<td>44.7</td>
<td>41.9</td>
<td>40.8</td>
<td>32.8</td>
<td>46.2</td>
<td>56.3</td>
<td>62.1</td>
</tr>
</tbody>
</table>

A, archaea; AA, A. aeolicus; AF, A. fulgidus; AP, A. pernix; B, bacteria; EC, E. coli; MJ, M. jannaschii; NE, N. equitans; MK, M. kandleri; PA, P. abyssi; PH, P. horikoschii; PF, P. furiosus; ST, S. solfataricus; TM, T. maritima.
Protein and DNA Thermostability, Physics and Evolution of

Proteome Sequence Determinant of Thermophilic Adaptation

The availability of complete genomes/proteomes makes it possible to search systematically for the combination of amino acid residues, which is most important for protein thermostability (41, 54). An exhaustive enumeration of all possible subsets of 20 amino acids is performed by representing sets of amino acids by vectors, where each component $a_i$ of the vector takes the value of 1 if the amino acid of type $i$ is presented in the set and 0 otherwise. Thus, $2^{20} - 2 = 524,286$ linearly independent non-trivial combinations exist. Given $F_{i,j}$, the fraction of amino acid $i$ in proteome $j$, total fraction $F$ of the amino acids from a particular subset, $F = \sum_i a_i F_{i,j}$, linear regression between the values of $F$ and the optimal growth temperature (OGT) of the organism allows us to determine the best predictor of OGT.

For 86 complete proteomes of prokaryotes thriving under temperatures from $-335$ to $110^\circ$C, the combination of Ile, Val, Trp, and Leu (IVWL) amino acids, which is presented in the set and 0 otherwise. Thus, $2^{20} - 2 = 524,286$ linearly independent non-trivial combinations exist. Given $F_{i,j}$, the fraction of amino acid $i$ in proteome $j$, total fraction $F$ of the amino acids from a particular subset, $F = \sum_i a_i F_{i,j}$, linear regression between the values of $F$ and the optimal growth temperature (OGT) of the organism allows us to determine the best predictor of OGT.

For $86$ complete proteomes of prokaryotes thriving under temperatures from $-335$ to $110^\circ$C, the combination of Ile, Val, Trp, and Leu (IVWL) amino acids, which is presented in the set and 0 otherwise. Thus, $2^{20} - 2 = 524,286$ linearly independent non-trivial combinations exist. Given $F_{i,j}$, the fraction of amino acid $i$ in proteome $j$, total fraction $F$ of the amino acids from a particular subset, $F = \sum_i a_i F_{i,j}$, linear regression between the values of $F$ and the optimal growth temperature (OGT) of the organism allows us to determine the best predictor of OGT.

For $86$ complete proteomes of prokaryotes thriving under temperatures from $-335$ to $110^\circ$C, the combination of Ile, Val, Trp, and Leu (IVWL) amino acids, which is presented in the set and 0 otherwise. Thus, $2^{20} - 2 = 524,286$ linearly independent non-trivial combinations exist. Given $F_{i,j}$, the fraction of amino acid $i$ in proteome $j$, total fraction $F$ of the amino acids from a particular subset, $F = \sum_i a_i F_{i,j}$, linear regression between the values of $F$ and the optimal growth temperature (OGT) of the organism allows us to determine the best predictor of OGT.

For $86$ complete proteomes of prokaryotes thriving under temperatures from $-335$ to $110^\circ$C, the combination of Ile, Val, Trp, and Leu (IVWL) amino acids, which is presented in the set and 0 otherwise. Thus, $2^{20} - 2 = 524,286$ linearly independent non-trivial combinations exist. Given $F_{i,j}$, the fraction of amino acid $i$ in proteome $j$, total fraction $F$ of the amino acids from a particular subset, $F = \sum_i a_i F_{i,j}$, linear regression between the values of $F$ and the optimal growth temperature (OGT) of the organism allows us to determine the best predictor of OGT.

Table 3: Percentage of the forward/backward replacements in alignments of hyperthermophilic genomes against mesophilic E. coli.

<table>
<thead>
<tr>
<th>Hyperthermophilic genome</th>
<th>RK/KR</th>
<th>L/I/L</th>
<th>TS/ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. aeolicus</td>
<td>20.0/8.1 (2.47)</td>
<td>14.2/19.3 (0.74)</td>
<td>7.5/6.8 (1.10)</td>
</tr>
<tr>
<td>A. fulgidus</td>
<td>14.5/10.6 (1.37)</td>
<td>14.4/17.5 (0.83)</td>
<td>8.4/6.6 (1.27)</td>
</tr>
<tr>
<td>M. jannaschii</td>
<td>22.4/8.0 (3.03)</td>
<td>20.6/7.2 (1.2)</td>
<td>7.0/6.5 (1.08)</td>
</tr>
<tr>
<td>M. kandleri</td>
<td>23.7/8.0 (3.95)</td>
<td>19.5/19.0 (1.03)</td>
<td>6.8/6.8 (1.00)</td>
</tr>
<tr>
<td>N. equitans</td>
<td>10.7/20.9 (0.52)</td>
<td>9.8/7.2 (1.36)</td>
<td>6.6/7.0 (1.00)</td>
</tr>
<tr>
<td>P. abyssi</td>
<td>10.7/20.9 (0.52)</td>
<td>9.8/7.2 (1.36)</td>
<td>6.6/7.0 (1.00)</td>
</tr>
<tr>
<td>P. horikoschii</td>
<td>16.5/9.9 (1.67)</td>
<td>16.3/18.2 (0.90)</td>
<td>7.8/7.5 (1.04)</td>
</tr>
<tr>
<td>P. furiosis</td>
<td>16.7/9.6 (1.74)</td>
<td>16.7/18.3 (0.92)</td>
<td>8.1/7.2 (1.13)</td>
</tr>
<tr>
<td>T. maritima</td>
<td>16.7/9.6 (1.74)</td>
<td>16.7/18.3 (0.92)</td>
<td>8.1/7.2 (1.13)</td>
</tr>
<tr>
<td>S. tokodaii</td>
<td>18.2/17.4 (2.46)</td>
<td>18.5/17.8 (1.04)</td>
<td>9.8/7.3 (1.34)</td>
</tr>
<tr>
<td>T. maritima</td>
<td>18.2/17.4 (2.46)</td>
<td>18.5/17.8 (1.04)</td>
<td>9.8/7.3 (1.34)</td>
</tr>
<tr>
<td>M. kandleri</td>
<td>8.1/15.8 (0.51)</td>
<td>9.8/7.2 (1.13)</td>
<td>6.6/7.0 (0.95)</td>
</tr>
</tbody>
</table>

when Ala (A) or a combination Ala, Gin (A,G) replaces Glu (E) or a combination Val, Gin (V,E). The correlation coefficient R is 0.18 and 0.23, respectively.

Finally, the fundamental question in thermophilic adaptation is a relationship between the amino acid composition of proteins and the nucleotide composition of coding DNA sequences (39-41). The availability of prokaryotic complete genomes, which consist mostly of coding DNA (on average ~85% of the total genome size), clarifies a relationship between the thermophilic adaptation of protein and DNA. If the NVYRELL predictor depends on nucleotide composition only, it must remain the same after the reshuffling of coding sequences given a nucleotide composition. However, the reshuffling of nucleotide sequences results in a non-NVYRELL thermostability predictor (32). Therefore, amino acid composition and thermal adaptation of proteins are not affected by the nucleotide composition of DNA sequences. The amino acid composition of the proteome, on the contrary, introduces a bias in the purine loading (A + G content) of nucleotide sequences. Indeed, purine loading of coding sequences reversely translated from protein sequences without codon bias, for example, by using synonymous codons with equal probabilities, is very close to a natural nucleotide sequence. The correlation coefficient between the purine loading and OGT is 0.48 and 0.6 in sequences without codon bias and natural ones, respectively (32).

Major Role of Stacking in DNA Thermostability

Pairing and stacking are two major factors of DNA stability. In a base pairing, the G+C pair contains three hydrogen bonds compared with the A+T pair that has two hydrogen bonds. The classical Marmur and Doty work (1) gives a linear relationship between the (G + C) content of the double-stranded DNA and its melting temperature, which strongly suggests that stability of the G+C and A+T pairs is different independently of their neighbors. This result originated a belief that DNA thermostability is provided mainly by pairing interactions and is achieved via (G + C) loading. Furthermore, the role of (G + C) content in establishing specific "thermophilic" biases in the amino acid composition of corresponding proteins was discussed extensively. A high-throughput analysis of genomic sequences conclusively demonstrates the absence of any connection between the (G + C) content and the OGT of the organism (39-41). The only bias in the nucleotide composition that correlates with the OGT is the (A + G) content (purine loading). It was shown, however, that purine loading is determined chiefly by the amino acid composition of proteins (32). Thus, thermostabilization of DNA does not work on the level of nucleotide composition.

The next step in the description of the DNA sequence is the analysis of the pair-wise nearest-neighbor correlations, for example, the normalized probabilities to find successive pairs of the nucleotides A and T. For all 16 possible successive dinucleotides in the coding strand of DNA, only the functions cCG and cCT correlate with OGT. The excess probabilities to find A+G and C+T pairs in the coding DNA are increasing significantly with OGT, correlation coefficient R = 0.68 and 0.601 (Fig. 2, upper row). Remarkably, the codon bias explains the observed sequence correlations in the coding parts of DNA. First, correlation in the nucleotide sequences does not depend on sequence correlations in amino acid sequences because removing an effect of the codon interface does not destroy a correlation between cAG, cCT and OGT, R = 0.736/0.574 (Fig. 2, middle row). Second, the correlation in DNA sequences stems from the neighboring nucleotides within a codon. Indeed, removing the natural codon bias results in eliminating the correlation between cAG, cCT and OGT, R = 0.177/0.216 (Fig. 2, bottom row). Thus, the codon bias establishes an excessive use of codons that contain successive A+G or C+T pairs, which is manifested in the correlation of OGT with the optimal growth temperature. The above sequence correlations in the coding strand of DNA sequences point to base stacking as a major factor of DNA thermostabilization. A+G dinucleotides have a low energy characteristic for a purine-purine stacking. The cCT correlation also shows, although indirectly, the role of stacking in thermostabilization. Indeed, the abundance of C+T pairs in the sense strand points to the equal enrichment of the antisense strand with A+G pairs because of the opposite directionality of sense and antisense strands of DNA. Therefore, the thermostabilization of double-stranded DNA is based on the stacking interactions provided by A+G pairs that are spread in different locations of both sense and antisense strands. This picture holds also for the whole DNA of prokaryotes, including its coding and noncoding parts. Therefore, in the scenario of thermophilic adaptation of double-stranded DNA, the stacking interactions play a major role. The codon bias provides an increase in the number of A+G dinucleotides with OGT in both sense and antisense strands of the DNA double helix. The necessity for A+G pairs can be explained by their low free energies of stacking obtained both theoretically (56) and experimentally (57). First, the study of the free energy contribution to the nucleic base stacking in aqueous solution shows that the free energy of stacking in order of decreasing stability follows the order purine-purine > purine-pyrimidine-pyrimidine-purine-pyrimidine-pyrimidine in general, and the free energy of A+G stacking is one of the lowest in particular (56). Second, the experimental study on the coaxial stacking contribution to the stabilization of gel-immobilized duplexes reveals that adenine stacking with other bases is significantly stronger than the stacking of other bases (57). The reasons for the discrepancy between the latter and the parameters of duplex stability obtained in the nearest neighbor approximation are yet to be explored (see Reference 27 and References 16-19 therein).

Recent experimental efforts also corroborate a major role of the base stacking (31) in DNA thermostability and the independence of the latter on G+C base pairing (1). In particular, DNA stacking parameters are determined directly (31) for the temperatures from below room temperature to close to melting temperature and for the salt concentrations from 15 to 100 mM Na+ . It seems that base stacking is the main stabilizing factor in the double-stranded DNA that determines the temperature and the salt dependence of DNA stability parameters. The A+T pairing is always destabilizing, and G+C pairing contributes almost no stabilization (31). It is important to note that base stacking interactions always are stabilizing for both A+T and G+C-containing contacts in double-stranded DNA. Bioinformatics studies display the importance of stacking...
by showing the independence of the DNA thermostability on (C + G) content (32, 39–41) and by illuminating a specific role of ApG stacking in the thermostability of the DNA double helix via a consideration of pair-wise nearest neighbor correlations (32) or a regression analysis of the dinucleotide composition of genomic DNA (41).

**Minimalist Physical Model of Protein Thermostability**

It is of great importance for protein design to elucidate how physical principles work in the evolution of natural proteins and how they provide viability and adaptation to different environments. A analysis of individual prokaryotic and eukaryotic proteins reveals a direct connection between their stability (expressed in melting temperature, \( T_{\text{melt}} \)) and the average living temperature of the organism, \( T_{\text{env}} \) (4); hence, environmental temperature should be incorporated in the model of protein thermophili design. In terms of statistical physics, the stability of the native state of a protein is determined by the Boltzmann factor exp(\( -\Delta E/RT \)), where \( \Delta E \) is the energy gap between the native state and the lowest energy completely misfolded structure (58, 59). This factor imposes a requirement on the energy gap; it must increase with the temperature (Fig. 3), and, as a result, the unique lowest energy native state will be preserved from the destruction by thermal fluctuations. The widening of the energy gap can be achieved by lowering the energy of the native state (positive design) and by increasing the energy of the misfolded structures (negative design) or by both processes working simultaneously.

**Design of Model Proteins with Selected Thermal Stability**

The first approach to simulation of protein thermophilic adaptation is to start from a purely statistical–mechanical analysis of protein thermostability. A specific Monte Carlo procedure [the so-called P-design (60, 61)] exists that maximizes the Boltzmann probability \( P_{\text{nat}} \) of being in the lowest energy (native state) conformation, \( P_{\text{nat}}(T_{\text{env}}) \approx \frac{e^{-\Delta E/RT_{\text{env}}}}{e^{-\Delta E/RT_{\text{env}}}} \), where \( E_3 \) is the lowest energy among all conformations and \( T_{\text{env}} \) is the environmental temperature. It takes the environmental temperature \( T_{\text{env}} \) as an input physical parameter, introduces mutations in the amino acid sequence, and accepts or rejects them according to the Metropolis criterion. As a result, this procedure designs proteins stable at given \( T_{\text{env}} \). The stability of designed proteins is characterized by their melting temperature \( T_{\text{melt}} \) that can be found numerically from the condition \( P_{\text{nat}}(T_{\text{env}}) = 0.5 \).
“From Both Ends of the Hydrophobicity Scale” Trend in Thermophilic Adaptation

The design of model protein for thermostability by using an exactly solvable lattice model (103,346) compact conformations of $3 \times 3 \times 3$ lattice proteins (62) discovers the fundamental rules of thermophilic adaptation. First, the amino acid composition of designed proteins reveals a specific “thermophilic” trend: Thermostabilization is accompanied by an increase of the amount charged (DEKR) and the hydrophobic amino acids (MPCLVWIF) at the expense of weak hydrophobic and polar (AGNQSTHV) ones. Importantly, the amino acid composition of β3 proteomes of psycho-, meso-, thermo-, and hyperthermophilic prokaryotes reveals similar trend. Thus, the “from both ends of the hydrophobicity scale” trend, that is combining amino acids with the maximum variance in their hydrophobicity, observed in simulations is, indeed, crucial for the thermostabilization of proteins.

The “from both ends” trend is related to the positive and negative components of the design. The positive design is a major contributor to the temperature-dependent energy decrease of the native state, and the negative design ensures an increase of the average energy of misfolded conformations (when an increase occurs in the $T_{m}$). Interactions between strongly hydrophobic residues in the protein core and ion pairs formed by amino acids of the opposite charge on the protein surface are responsible for the positive design. The repulsion between charged residues of the similar sign contributes to the negative design by raising the average energy of misfolded conformations (see Fig. 3). Importantly, both positive and negative components of the design are based on the conservative native and non-native contacts between residues that play an especially important role in the stabilization of the native state and the destabilization of the misfolded conformations (62). Whereas identities of amino acids that form such a contact may vary from sequence to sequence, the strength (or interaction energy) of the key native and non-native contacts is preserved: These contacts are either strongly repulsive or strongly attractive for all sequences that fold into a given structure (see Fig. 6 in Reference 62).

Design simulations confirm an existence of the energy-conserved strongly attractive (native) and most repulsive (non-native) contacts (62). The standard deviation of the contact energy is the lowest for these contacts. When the design is performed under hyperthermophilic temperatures, it results in stronger and more conserved (lower dispersion of the energy) native and non-native interactions compared with the design under the mesophilic temperatures (see Fig. 7 in Reference 62). As a result, the gap between the energies of native and misfolded structures is widening and the thermostability of the structure is increasing in response to the elevated environmental temperature $T_{m}$.

Positive and Negative Design in Evolution and Thermal Adaptation of Natural Proteins

The requirement to preserve energy of key contacts in multiple sequences that fold into the same structure implies that amino acids forming such contacts should mutate in a correlated way. For example, correlated mutations may occur as swaps to keep specific attractive native and repulsive non-native interactions (see Fig. 8 in Reference 62). This scenario invokes a peculiar dependence between the amino acid substitution rates [e.g., BLOSUM matrices (63)] and the interaction energy between corresponding amino acid residues [e.g., the Miyazawa–Jernigan quasi-chemical potential (64)]. Frequent substitutions are expected between amino acids that strongly attract each other (to preserve specific stabilizing native contacts) and between amino acids that strongly repel each other (to preserve specific non-native repulsive contacts). The dependence of elements of substitution matrix BLOSUM (62, 63) for 190 pairs of amino acids (synonymous substitutions are excluded) versus their interaction energy as approximated by the knowledge-based Miyazawa–Jernigan potential (64) has a nonmonotonic nature (Fig. 4, top chart: the dependence is highlighted by the parabolic fit). The most frequent substitutions are observed between the most attractive and most repulsive amino acids. A blow-up of the right top part of Fig. 4 (bottom chart) shows that along with conserved substitutions that reflect a positive design (arginine to lysine and glutamic acid to aspartic acid substitutions), frequent substitutions exist between mutually repulsive amino acids with vastly different physical-chemical properties and encoded by very dissimilar amino codons, such as glutamine to arginine, serine to asparagine, and so forth (Fig. 4, bottom chart). The high frequencies of substitutions between residues that strongly repel each other explain the correlated mutations observed between the residues that are distant in the native structure (62). These residues may form important repulsive contacts, which increase the energies of the misfolded conformations (see Fig. 10 in Reference 62).

Whereas a positive design is used widely in experiments (65), the big challenge in using the negative design originates from the difficulties in the modeling of relevant misfolded conformations (66). Nevertheless, charged residues were used effectively in negative design (65, 66). Site-directed mutagenesis provides other, although indirect, evidence of the contribution of charged residues to negative design: The mutation of polar groups to negatively charged ones on the protein surface leads to structure stabilization even in the absence of the salt-bridge partners of the mutated group. It also has been shown (67–70) that surface electrostatic interactions provide a marginal contribution to stability of the native structure, hence, the possible importance of charged amino acids is in making unfavorable high-energy contacts in misfolded structures. In the case of thermophilic adaptation, positive and negative components of design work concurrently and provide stabilization of the structure via an “opening” of the energy gap from both sides: A decreasing energy of the native state and, at the same time, an increasing energy of misfolded conformations can both exist.

Conclusions

A deep understanding of the physical mechanisms and the evolution of thermophilic adaptation is crucial for the engineering and design of biologic catalysts with desired stability (20). This
knowledge also is important for establishing a trade-off between the stability and flexibility in a directed evolution of protein function (66, 71). Current predictors of the stability effects of protein mutations are based on empirical potentials that are calibrated to fit experimentally observed $\Delta\Delta G$ values (20, 72, 73). Although predictions of $\Delta\Delta G$s during mutation in the native state are in good agreement with experimentally observed ones, they lack the effect of mutations on misfolded conformations, the structure-dependence of mutation effects (37, 38), and the dependence of mutations on the evolutionary strategy of thermophilic adaptation (43).

Recent computational studies of thermophilic adaptation described in this article make use of genomic/proteomic data (32, 43, 53, 62), simulations of model lattice proteins (62), and off-lattice all-atom simulations of natural proteins (43, 53). High-throughput analysis reveals signals of novel mechanisms of protein [entropic mechanism (53)] and DNA [purine–purine base stacking (32)] thermostability and urges us to consider what evolutionary strategy was followed in the process of thermal adaptation (43). Proteomic analysis and simulations of thermophilic adaptation also demonstrate that negative design necessarily should be taken into account to properly predict the effect of protein mutations (62).

References


Further Reading


See Also

The term “symbiosis,” from the Greek: συµβίον = living together, was defined by Anton de Bary in his monograph “Die Erscheinung der Symbios” as “the living together of unlike organisms” (1). His studies were based on the formation of lichens, which are the result of an association between a fungus and an alga or cyanobacterium. The definition was coined in the end of the nineteenth century but is regarded by most symbiosis researchers as largely valid today. Accordingly, any specific association between two or more species can be classified as symbiosis. It should be noted that many scientists use symbiosis in a more restricted way to denote a mutually beneficial relationship. This article will give an overview of various biologic manifestations of symbiosis and discuss selected examples, where primary or secondary metabolites play a crucial role in the association.

Symbiosis and Its Variants

If the partners in a symbiosis differ in size, the larger member is termed “host” and the smaller member is termed “symbiont” or “symbiote.” The more common term “symbiont” will be used here. One general way to distinguish between various forms of symbiosis is to identify the location of the attachment of the symbiont to the host. Symbionts that live on the host surface, including internal surfaces like the digestive tube, participate in ectosymbiosis (Greek: εκτός = outside). If a symbiont is localized within the tissues of the host, the association is termed “endosymbiosis” (Greek: ενδόν = within). Endosymbionts can be found either in the extracellular space or intracellularly.
with beneficial mycorrhizal fungi (2) or nitrogen-fixing bacteria (3) living on and in their roots. A mutual relationship can consist of partners that benefit from the association but do not depend exclusively on each other for survival. This mutual relationship is called protocooperation (Greek: πρωτ = first; Latin: cooperatio = cooperation). Such a facultative mutualism is known from the association of fungi and photosynthesizing microorganisms in lichens (4). In contrast, an obligate mutualism, where host and symbiont cannot survive after separation, is known as endosymbiosis (Greek: εν = good). Examples are mitochondria, chloroplasts, and other bacterial intracellular symbionts of higher organisms. These so-called primary or obligate symbionts usually are ancient, if specialized supracellular host structures exist that harbor such symbionts, they often are termed bacteriosomes. In insects, evidence exists for obligate symbioses that last 30–270 million years, and the symbionts usually coevolve with their hosts (5–7). During this long time of host dependency, the genomes of primary symbionts can undergo massive restructuring, which results in extremely small genome sizes. In addition to primary symbionts, many hosts harbor secondary or facultative symbionts. They participate in an intermediate form of symbiosis, where the symbiont is not strictly necessary for the survival of their host and is not exclusively localized in specific tissues. Because of occasional horizontal transfer, the phylogenetic trees of such organisms often are not congruent with those of their hosts (8). Many secondary symbionts are transmitted maternally and can influence, positively or negatively, various host traits, such as nutrition, survival against natural enemies, or reproduction rates (9, 10).

On the opposite side of mutualism in the interaction spectrum of symbiosis is parasitism (Greek: παρα = besides; σίτ = fatted). Here one species, called the parasite, has physiological or structural properties that force it to live temporarily or during its whole life on (ectoparasitism) or inside (endoparasitism) another species, the host. Only the parasite benefits from this relationship. Different types of parasitism exist, including biotrophy, in which the host survives during the entire time of the association; pathotrophy, in which the host lives for a short time after contracting a parasite, and necrotophy, in which the host is killed before the symbiosis starts. Biotrophs and necrotrophs are known also as parasitoids. This term was coined in 1913 by the entomologist D. M. Rauter to distinguish between true parasites that live at the expense of their hosts without actually causing their death and the parasites that always kill their hosts (11). Among the parasitoids are two categories, which describe the behavior of the parasitoids toward the host. A first the initial paratization, the idiobiont parasitoid prevents any additional development of the host. This parasitization is typical for an immobile host life stage like an egg or a pupa. Most of these parasitoids are endoparasites. A literally, a koinobiont parasitoid allows the host to continue its development and often does not kill or consume the host until it starts to pupate or become an adult. These parasites, therefore, mostly live in or on an active and mobile host. Some primary parasitoids serve as hosts for yet another parasitoid. The latter is termed "secondary parasitoid" or "hyperparasitoid" and usually kills both the host and the primary parasitoid. Several other commonly used terms exist for different kinds of parasitism. Kleptoparasitism is a relationship in which the parasite steals food that a host has caught or otherwise prepared. An example is the brood parasitism, which is known from many species of cuckoo that use other birds for parenting (12). Social parasites take advantage of interactions between members of a social host species like ants or termites (13). A rather special type of parasitism is cheating or exploitation. Here, the parasite uses situations of nonspecific mutualism to its advantage. Some myco-heterotrophic plants, for instance, establish mycorrhiza-like interactions with fungal symbionts and take carbon from the fungi that the fungi obtain from other plants (14). Opportunism is a kind of parasitism where harmless parasites under special circumstances cause diseases or death of their hosts. The fungus Pneumocystis carinii can initiate pneumonia if the host is infected by HIV (15). As in mutualism, obligate and facultative symbionts exist in parasitism. Obligate parasites are not viable without their host, and facultative parasites are free-living organisms that sometimes benefit from their host.

A third form of symbiosis, in addition to mutualism and parasitism, is commensalism (Latin: com = with; mensa = table). In commensal relationships, one symbiotic partner benefits from the association and the other is unaffected. Different variants of such relationships have been described. Inquilinism is a kind of commensalism where the symbiont uses a host for housing, such as birds living in the holes of trees. A more indirect dependency exists if a symbiont uses something its host has created or even the main determinant for a symbiosis, in which case the term "chemosymbiosis" often is used. A wide spectrum of substances can be involved, ranging from simple inorganic ions to highly complex secondary metabolites. Similarly diverse can be their function, such as nutrition, chemical protection, localization cues, developmental signals, and others. The chemical ecology of symbiotic associations is a relatively unexplored field.
field, and particularly systems that involve microorganisms can be difficult to study because cultivation requirements often are unknown or impossible to reproduce. However, with the advent of culture-independent methods and innovative cultivation techniques, this situation recently has begun to change.

The following sections will give some representative examples where small molecules play an important role in the symbiotic interaction. Cases involving inorganic substances and primary metabolites will be discussed; secondary metabolism will be the subject of the subsequent section.

Inorganic compounds

A best-studied example for a symbiosis between plants and bacteria is the association between leguminous plants and various members of the α-proteobacterial order Rhizobiales (3). The bacteria colonize root nodules of the plants and differentiate morphologically into bacteroids. These bacteroids fix nitrogen from the atmosphere, convert it to ammonia and amino acids, and supply it to the host plant. The bacteria receive organic acids like malate or succinate as a carbon and energy source from the plant. It has to be noted that this symbiosis is not a strict mutualism. The nitrogen fixation is energetically costly to the rhizobia and reduces the resources that could be allocated to their own growth and reproduction (17, 18). Some soils contain rhizobial strains that fix little or no nitrogen (19). Because the strains can coinfect an individual plant together with nitrogen fixers (17), any of these strains could redirect the resources from nitrogen fixation to its own growth and fitness at the expense of its host and the other lineages. This parasitism is a kind of cheating parasitism. Hence, the legume needs ways to guide the evolution of rhizobia toward greater mutualism (20).

Mycorrhiza, an association between soil fungi and plant roots, is one of the best-studied symbioses. The fungi, termed “mycorrhizal,” colonize the roots of 80% of all land plants carry (vesicular-)arbuscular mycorrhiza (32). The mycorrhiza is densely vascularized surface. It allows an efficient exchange of metabolites and waste products between the worm and the sea water. The major interior tissue of the worm is the trophosome that is localized in a large sac formed by the body wall (32). The cells of the trophosome, termed “bacteriocytes,” are colonized densely by the bacterial endosymbiont (33, 34). Its amount is estimated to reach between 15% and 35% of the whole cell volume (35). The bacteria are supplied with various nutrients by the blood circulation of the worm, which is driven by a heart-like pump. The blood contains an unusual multiheme-moglobin system, which can transport both sulfide and oxygen (36, 37). The mutual contributions of the bacteria are manifold. One function is to oxidize the transported sulfide into sulfite to produce metabolic energy for both partners (33, 38, 40). Furthermore, the bacteria are indispensable in several assimilation processes by providing various enzymes that the worms lack. An example is the CO₂ absorbed by the worm, which is transported to the bacteria by the blood cycle, either directly (41) or via synthesized malate (42). The bacteria, in turn, employ enzymes of the Calvin-Benson cycle to convert the CO₂ to small organic compounds (38, 42, 43). These metabolites then are used by the host for its own metabolism and production of adenosine triphosphate (ATP). The assimilation of ammonia and nitrate also depends on the endosymbiont. Although the glutamine synthetase and glutamate dehydrogenase required for ammonia assimilation are present in both partners (44), the nitrate-reducing enzymes are provided exclusively by the bacteria (43–45). The symbiont also provides the necessary enzymes for a de novo synthesis of pyrimidine but lacks the enzymes of the salvage pathways that allow pyrimidine synthesis from nucleic acid degradation products (45, 46). The host, however, only possesses enzymes that catalyze the final steps of the de novo pathway but provides enzymes of the salvage pathway (46). The de novo synthesis of pyrimidine, thus, is regulated by the bacterial enzymes. Arginine is metabolized in a similar manner. The worm does not possess arginine decarboxylase or ornithine decarboxylase activities (47), and it is likely that it depends on the bacteria to realize this pathway.

In another gillless marine worm, the oligochaete Olavius algarvensis, three partners participate in an unusual mutualism. The worm harbors two different primary symbionts in immediate proximity below its cuticle between extensions of epidermal cells. By performing comparative 16S rRNA sequencing and
fluorescence in situ hybridization, two dominant clone groups of the γ- and δ-subclasses of the proteobacteria were found (48). The larger γ-proteobacterial symbiont possesses a high sequence similarity with endosymbionts from other gutless oligochaetes, like Oligolimnus loisei and Inanidrilus leukodermatus (49, 50), and is thioautotrophic, i.e., possesses a sulfur-oxidizing, CO₂-fixing metabolism (48). The smaller δ-proteobacterial symbiont is genetically very similar to free-living sulfate-reducing bacteria, and the sulfate-reducing function also could be demonstrated in O. algarvensis (48). The coexistence of these bacteria indicates that they are engaged in a synthetic sulfur cycle in which oxidized and reduced sulfur compounds are exchanged back and forth (48). The oligochaete benefits from this arrangement by obtaining de novo-synthesized organic compounds from the symbionts, which in turn assimilate its anaerobic end products. The constant internal source of reduced sulfur compounds allows the worm to colonize new habitats without a requirement of high sulfide concentrations.

Numerous additional examples exist of intimate symbioses that involve the exchange of inorganic compounds, including cyanobacteria (providing fixed nitrogen to lichens, the water fern Azolla, Gunnera spp. plants, cycads and some mosses, liver worts, and horn worts) (51), clams of the genera Solen and Codakia (providing sulfate to bacterial) (52, 53), the sponge Haliclona cymiformis (providing ammonia and nitrate to a rhodophyte alga) (54), corals (inorganic carbon to photosynthetic dinoflagellates for carbon fixation) (55, 56), rumen Halicllona sponge is thioautotrophic, i.e., possesses a sulfur-oxidizing, CO₂-fixing metabolism (48). The smaller δ-proteobacterial symbiont is genetically very similar to free-living sulfate-reducing bacteria, and the sulfate-reducing function also could be demonstrated in O. algarvensis (48). The coexistence of these bacteria indicates that they are engaged in a synthetic sulfur cycle in which oxidized and reduced sulfur compounds are exchanged back and forth (48). The oligochaete benefits from this arrangement by obtaining de novo-synthesized organic compounds from the symbionts, which in turn assimilate its anaerobic end products. The constant internal source of reduced sulfur compounds allows the worm to colonize new habitats without a requirement of high sulfide concentrations.

Numerous additional examples exist of intimate symbioses that involve the exchange of inorganic compounds, including cyanobacteria (providing fixed nitrogen to lichens, the water fern Azolla, Gunnera spp. plants, cycads and some mosses, liver worts, and horn worts) (51), clams of the genera Solen and Codakia (providing sulfate to bacterial) (52, 53), the sponge Haliclona cymiformis (providing ammonia and nitrate to a rhodophyte alga) (54), corals (inorganic carbon to photosynthetic dinoflagellates for carbon fixation) (55, 56), rumen bacteria (providing hydrogen and CO₂ to methanogenic archaea) (57), and “Chlororhabdium aggregatum,” an association of a sulfate-reducing and a photosynthetic sulfide-oxidizing bacterium (58).

Primary organic metabolites

Among the best-investigated chemosymbioses that involve primary metabolites are mutualisms of eukaryotic cells with their mitochondria and chloroplasts. M. lateralis synthesizes ATP, as well as heme and steroids, by oxidative phosphorylation, whereas chloroplasts provide plant and algal cells with glucose generated via the Calvin cycle by using the energy of light. A according to the endosymbiotic theory, these organelles developed from independent prokaryotic organisms that were absorbed by their host cells (59). Several lines of evidence, such as the phylogeny of plastid DNA sequences and the structure of the photosynthetic apparatus, suggest that chloroplasts developed from cyanobacteria. The origin of the mitochondria has been traced back to an archaebacterium. The two associations are typical examples of eusymbioses: The organelles are genetically very similar to free-living sulfate-reducing bacteria, their genome size of 4.2 Mb (66), much larger than that of the primary symbiont, and shows only beginning signs of erosion (67). It is possible to cultivate the bacterium in the absence of host cells (68). A phidids feed on a plant sap diet rich in carbohydrates but poor in vitamins and amino acids. They possess a bacteriosome that contains up to 80 large cells filled with bacteria of the genus Buchnera (69). They are transmitted maternally through the host generations, and neither of the mutual partners can re-produce independently (70). The association of the aphids with the endobacterial-like ancestor of the symbiont was established 200–250 million years ago (71). The genome size of Buchnera aphidicola strains measures only 420–640 kb, which is less than one seventh of that of the close relative E. coli (72). During evolution, the symbiont has lost numerous genes, including those that encode the biosynthesis of cell-surface components, regulatory genes, and genes for its own defense (73). Notably, B. aphidicola lacks several pathways for amino acids that the host can synthesize, although it retains genes for the production of all essential amino acids (7, 73). Evidence exists that the aphid exports synthesized glutamate to the symbiont as a nitrogen source, which allows it to produce its essential amino acids (73, 74). The strain B. aphidicola BCC of the aphid Cinara cedri has an exceptionally small genome of only 422 kb that even lacks essential tyrosine and riboflavin genes (75). It has been hypothesized to gradually lose its mutual properties and to
Carsonella rud-symbiotica” (76). The smallest known genome among insect by-product of some protists is hydrogen and CO₂ (79). These in their gut (78). The protozoan converts cellulose to acetate for a complex community of protozoans, bacteria, and archaea (80). The bacteria are embedded in the host membrane and wave synchronously to assist in host locomotion. Symbioses based on the degradation of organic material are generally common in animals with a diet rich in cellulose, such as ruminants and wood-boring insects. Besides protozoa, bacterial and fungal cellulose degraders have been identified in their guts. Several insects are known to actively ingest fungi to use their cellulases. This behavior might represent the first stage in the establishment of a symbiosis with such fungi stably residing in the gut (81).

An economically important chemical protocooperation of two different bacteria exists in yogurt preparation. A widely used starter for the production of fermented dairy products is the thermophilic lactic acid bacterium Streptococcus thermophilus. For the manufacturing of yogurt, it is generally used in association with other microorganisms, in particular with Lactobacillus delbrueckii ssp. bulgaricus). To enable the intense and rapid acidification of milk, the fast growing capacity of these bacteria is crucial. The association causes a positive effect on the growth of the other species. The growth of L. bulgaricus is stimulated by pyruvic acid, formic acid, and CO₂ that are produced by S. thermophilus (82). L. bulgaricus is a producer of amino acids and peptides that stimulate the growth of S. thermophilus (83). The symbiosis additionally results in enhanced aroma formation compared with the pure cultures.

The exchange of primary metabolites also is known from several nonmicrobial symbioses. Carbohydrates play an important role in many of these symbioses. Besides the ubiquitous example of flowering plants offering nectar, several plants are known that produce extracellular nectar to attract beneficial insects, in particular ants (84). Ants colonizing suchy myrmecophytes or ant plants protect their nutrient source by efficiently removing herbivores and even pathogenic fungi. Other nutrient sources provided by ant plants are food bodies rich in proteins or lipids. In a well-known functionally related association, ants obtain carbohydrates from honeydew secreted by aphids that they regularly milk and defend against predators (85).

Secondary metabolites

Sessile marine animals, such as sponges, ascidians, and bryozoa, are the source of numerous compounds with diverse pharmacological activities (86, 87). It has long been suspected that many of these natural products are produced, in fact, by symbiotic bacteria because the chemical structures often resemble those of compounds isolated from prokaryotes (87-89). The existence of a bacterial source for an invertebrate-derived natural product should have important biomedical implications, as this ultimately could allow one to create sustainable fermentation systems for drug development and production. Sponges are the most important sources of marine drug candidates. These animals can contain remarkable numbers of diverse microorganisms from various taxa representing up to 40-60% of the total biomass (90, 91). The microflora in most bateria is highly distinct but shows little variation among sponge species or geographic location, which indicates a long period of coexistence (92). At least a part of the microbial community is transferred vertically via the larval brood chamber (93). As with many other symbiotic systems, cultivation attempts have failed in most cases. Therefore, for most sponge-derived natural products, the actual producer is still unknown. Among the few successfully studied examples are the protein biosynthetic inhibitors anammone (Fig. 2a) and theopederins (Fig. 2b), potent cytotoxic polyketides isolated from Theonella swinhoei (94). A bacterial source was identified by metagenomic techniques, i.e., methods employing the DNA of entire organismic associations without prior separation of species (95, 96). The screening of a large metagenomic library prepared from total T. swinhoei DNA led, in this case, to the isolation of the biosynthetic gene cluster, which, where attributed to a bacterial producer (95). Another method employed to identify biosynthetic sources of natural products is mechanical cell separation and subsequent extraction of each cell type. In this way, the likely bacterial sources of several compounds were identified, including the polyketide swinholide (97), the nonribosomal peptide theopelalactam A (97), chlorinated dippeptides (Fig. 2a) (98, 99) and brominated diphenylethers (Fig. 2b) (97). One patent exists on the successful cultivation of an actinomycete symbiont from the sponge Acanthasteroyphora sp. producing anti-infective manzamine alkaloids (Fig. 2) (100).

A further pharmacologically relevant group of marine invertebrates is the phylum Ctenopora, also known as Bryozoa or Polyzoa. They are colonial filter feeders and each member is housed in a separate unit called a zoecium. The bryozoan Bugula neriifrons contains complex polyketides of the bryostatin series (Fig. 4), protein kinase C activators with potent anticancer activity (102, 103). Bryostatin 1 is one of the most promising drug candidates from marine invertebrates and has reached phase II clinical trials for combination therapy (104). The concentration of bryostatins is particularly high in larvae of B. neriifrons, where it is used as a defense against predators (105). However, in adults the concentration is very low or even undetectable when the animals lack larval brood chambers (106). 16S rDNA analysis and in situ hybridization revealed the presence of a y-proteobacterium, “Candidatus Endobugula serulata,” in the larvae (105-108). This bacterium was shown to harbor polyketide synthase genes that are good candidates for bryostatin biosynthesis because their presence is correlated strictly with high bryostatin levels (108).
A scidians are marine filter feeders with a rich natural products chemistry that live commonly associated with symbiotic bacteria (88, 89, 109). A well-studied symbiosis consists of photosynthetic Prochloron spp. cyanobacteria that occur in ascidians of the family Didemnidae (110). Prochloron spp. also can be found in bacterial mat structures of stromatoliths (111) but so far have not been detected outside of such structured environments. From didemnid ascidians, numerous cytotoxic cyclic peptides of the patellamide group (Fig. 5) were isolated (109, 112, 113). Mechanical separation of the Prochloron sp. symbiont from its host Lissoclinum patella and subsequent genome sequencing revealed a set of biosynthetic genes that after transfer into E. coli enabled this bacterium to produce two different patellamides (114). The genes also were identified in an independent study by screening a library of Prochloron sp. DNA constructed in E. coli for the presence of clones that synthesize patellamides (115). The patellamide gene cluster also has been engineered to produce a drug-like, non-natural cyclic peptide (116). So far, the ecological function of the peptides in the natural environment remains unknown.

The gorgonian soft coral Pseudopterogorgia elisabethae harbors dinoflagellates of the genus Symbiodinium. The symbiotic function of the dinoflagellate is to contribute to the nutrition of its host by sharing photoassimilates (55, 56). Extracts of the corals contain the pseudopterosins (Fig. 6), a family of diterpene glycosides with antiinflammatory and analgesic properties (117). The substances modulate degranulation and release mechanisms in immune cells (118). They have been included as antiirritants in a widely marketed cosmetic. The amounts and the specific composition of the pseudopterosins vary depending on the location of P. elisabethae (119). Their total amount can be as high as 20% of the dry mass. By applying radioactively labelled geranylgeranyldiphosphate (GGPP) to purified symbiont preparations, it was demonstrated that the biosynthesis of the pseudopterosins occurs in the dinoflagellate (120). Elisabetha-triene synthase, the enzyme catalyzing the conversion of GGPP...
Several remarkable natural product symbioses have been reported from terrestrial habitats. Ants of the genera Atta and Acromyrmex have an important function in sustaining a healthy symbiosis. The protective effect of pederin then might have driven a mutual facilitative symbiosis. Within each beetle species, most, but not all, females contain the complex polyketide pederin (Fig. 7). This was thought to be a defense mechanism against predators (9). Structural studies have shown that pederin is highly similar to a range of compounds isolated from marine sponges, such as the onnamides and the theopederins mentioned above (94). Isolation of the genes that encode pederin biosynthesis revealed that the substance is produced by a bacterial symbiont closely related to Pseudomonas aeruginosa (131). The symbiont is present only in pederin-producing females (132) and also can be detected in the eggs (133). P. aeruginosa does not produce pederin, and partial genome sequencing suggests that the symbiont or its free-living ancestor might have acquired the pederin genes by horizontal gene transfer (134). This mechanism also could account for the occurrence of similar compounds in beetles and sponges. Because P. aeruginosa is a pathogen of insects and other organisms (including humans), it is likely that the symbiosis developed after the beetle was infected by a pathogenic ancestor of the symbiont.

The protective effect of pederin then might have driven a mutual adaptation of both partners toward symbiosis. The symbiont confers a greater resistance against fungal infection because of the production of an antibiotic (128). Panderus and Pandraeus spp. rove beetles participate in an unusual facultative symbiosis. Within each beetle species, most, but not all, females contain the complex polyketide pederin (Fig. 7). In their hemolymph (130). The symbiont is highly toxic and is used as a defense against predators (9). Structurally, pederin is highly similar to a range of compounds isolated from marine sponges, such as the onnamides and the theopederins mentioned above (94). Isolation of the genes that encode pederin biosynthesis revealed that the substance is produced by a bacterial symbiont closely related to Pseudomonas aeruginosa (131). The symbiont is present only in pederin-producing females (132) and also can be detected in the eggs (133). P. aeruginosa does not produce pederin, and partial genome sequencing suggests that the symbiont or its free-living ancestor might have acquired the pederin genes by horizontal gene transfer (134). This mechanism also could account for the occurrence of similar compounds in beetles and sponges. Because P. aeruginosa is a pathogen of insects and other organisms (including humans), it is likely that the symbiosis developed after the beetle was infected by a pathogenic ancestor of the symbiont.

The protective effect of pederin then might have driven a mutual adaptation of both partners toward symbiosis. The symbiont confers a greater resistance against fungal infection because of the production of an antibiotic (128). Panderus and Pandraeus spp. rove beetles participate in an unusual facultative symbiosis. Within each beetle species, most, but not all, females contain the complex polyketide pederin (Fig. 7). In their hemolymph (130). The symbiont is highly toxic and is used as a defense against predators (9). Structurally, pederin is highly similar to a range of compounds isolated from marine sponges, such as the onnamides and the theopederins mentioned above (94). Isolation of the genes that encode pederin biosynthesis revealed that the substance is produced by a bacterial symbiont closely related to Pseudomonas aeruginosa (131). The symbiont is present only in pederin-producing females (132) and also can be detected in the eggs (133). P. aeruginosa does not produce pederin, and partial genome sequencing suggests that the symbiont or its free-living ancestor might have acquired the pederin genes by horizontal gene transfer (134). This mechanism also could account for the occurrence of similar compounds in beetles and sponges. Because P. aeruginosa is a pathogen of insects and other organisms (including humans), it is likely that the symbiosis developed after the beetle was infected by a pathogenic ancestor of the symbiont.

The protective effect of pederin then might have driven a mutual adaptation of both partners toward symbiosis. The symbiont confers a greater resistance against fungal infection because of the production of an antibiotic (128). Panderus and Pandraeus spp. rove beetles participate in an unusual facultative symbiosis. Within each beetle species, most, but not all, females contain the complex polyketide pederin (Fig. 7). In their hemolymph (130). The symbiont is highly toxic and is used as a defense against predators (9). Structurally, pederin is highly similar to a range of compounds isolated from marine sponges, such as the onnamides and the theopederins mentioned above (94). Isolation of the genes that encode pederin biosynthesis revealed that the substance is produced by a bacterial symbiont closely related to Pseudomonas aeruginosa (131). The symbiont is present only in pederin-producing females (132) and also can be detected in the eggs (133). P. aeruginosa does not produce pederin, and partial genome sequencing suggests that the symbiont or its free-living ancestor might have acquired the pederin genes by horizontal gene transfer (134). This mechanism also could account for the occurrence of similar compounds in beetles and sponges. Because P. aeruginosa is a pathogen of insects and other organisms (including humans), it is likely that the symbiosis developed after the beetle was infected by a pathogenic ancestor of the symbiont.

The protective effect of pederin then might have driven a mutual adaptation of both partners toward symbiosis. The symbiont confers a greater resistance against fungal infection because of the production of an antibiotic (128). Panderus and Pandraeus spp. rove beetles participate in an unusual facultative symbiosis. Within each beetle species, most, but not all, females contain the complex polyketide pederin (Fig. 7). In their hemolymph (130). The symbiont is highly toxic and is used as a defense against predators (9). Structurally, pederin is highly similar to a range of compounds isolated from marine sponges, such as the onnamides and the theopederins mentioned above (94). Isolation of the genes that encode pederin biosynthesis revealed that the substance is produced by a bacterial symbiont closely related to Pseudomonas aeruginosa (131). The symbiont is present only in pederin-producing females (132) and also can be detected in the eggs (133). P. aeruginosa does not produce pederin, and partial genome sequencing suggests that the symbiont or its free-living ancestor might have acquired the pederin genes by horizontal gene transfer (134). This mechanism also could account for the occurrence of similar compounds in beetles and sponges. Because P. aeruginosa is a pathogen of insects and other organisms (including humans), it is likely that the symbiosis developed after the beetle was infected by a pathogenic ancestor of the symbiont.
stages that emerge from the carcass and search for new insect hosts. Bacteria of both genera can be cultivated easily. They have been shown to synthesize various broad-spectrum antibiotics, lipases, phospholipases, and proteases that are believed to become secreted into the insect hemolymph when the bacteria enter the stationary phase. From these, hormetic photobacterins (Fig. 8a) (140), antibiotica, and nematicidal hydroxystilbenes (Fig. 8b) (141) are synthesized. Various broad-spectrum antibacterial xenorhabdins (Fig. 9a) (142), fungal xenocoumacins (Fig. 9b) (143), cytoxic polyquiniones (Fig. 9c) (144), protein biosynthetic potential (159). Their exact ecological functions, however, in most cases are poorly understood.

References


A great diversity of endophytic fungi colonizes the interior of most plants, red algae, and brown algae (155). The fungi often can be isolated and subsequently studied by placing leaf fragments on solid media. In addition to exchanged nutrients, several secondary metabolites have been shown to play an important role in this interaction. Fungi-derived alkaloids, such as pederin (Fig. 11a), lolitrem (Fig. 11b), enigmatic alkaloids (Fig. 11c), and lolitrems (Fig. 11d) (159) are produced in Convolvulaceae in a symbiosis with a clavicipitaceous fungus that is transmitted via the seeds (158). Numerous additional compounds have been identified from isolated endophytes, which documents their remarkable biosynthetic potential (159). Their exact ecological functions, however, in most cases are poorly understood.
Figure 9: Examples of natural products from Xenorhabdus spp.

Figure 10: Rhizoxin, a polyketide from bacterial endosymbionts of the pathogenic fungus Rhizopus microsporus.

Figure 11: Peramine (a), Norloline (b), Elymoclavine (c), Lolitreme B (d), alkaloids from endophytic fungi.


Symbiosis: Topics in Chemical Biology


The mammalian sense of taste is crucial for evaluating food palatability and nutritional quality. To achieve the detection of relevant chemicals, evolution has shaped a set of receptor molecules that allow the detection of five basic taste qualities: sweet, umami, bitter, salty, and sour. Each taste modality has unique characteristics and serves a distinct function for an animal's nutrition. Therefore, we address the characteristics, the recent advances, the persisting difficulties, and the future perspectives of the basic taste qualities in separate paragraphs. Enormous progress has been made in the identification and the characterization of taste receptor molecules, and in the growing number of animal genomes accessible from databases, which has inspired us to devote a section of this review to discuss a series of sophisticated evolutionary studies on taste receptor molecules. Finally, evidence is accumulating to show that taste receptor and signal transduction molecules have nongustatory functions as well. The extragustatory expression of such genes and the resulting implications are summarized in the final section.

The chemical senses of olfaction and gustation were developed from phylogenetically old systems that enabled organisms to detect chemicals in their environment. The detectors were linked to behavioral patterns that enabled the organisms to escape noxious substances or to approach potential nutrients. These stereotypic mechanisms are still present in higher organisms, including mammals, although they compose more complex regulatory loops. The chemical sense of taste allows mammals to evaluate the food they consume. Each of the five basic taste modalities fulfills a particular task. Sweet and umami (glutamate and 5′-ribonucleotides) taste detects calorie-rich food that contains carbohydrates or protein. Therefore, both sweet and umami tastes are linked to pleasant feelings and to behaviors that facilitate food intake. Salty taste is part of a control loop that underlies electrolyte homeostasis. Salt intake compensates for salt loss through sweating and elimination. Like umami and sweet taste, salty taste is linked to liking and attraction promoting intake. Sour and bitter tastes are repulsive and seem to be part of a warning system. Sour taste prevents excessive intake of protons and balances the acid and bases in the body; also it prevents intoxication through consumption of spoiled food or unripe fruits. Bitter taste prevents ingestion of noxious compounds.

Taste sensation is initiated on contact of chemicals dissolved in saliva with cognate taste receptor molecules on the apical side of specialized epithelial cells (1). These taste receptor cells seem to be dedicated to only one of the basic taste modalities (2). They are assembled into groups of ~100 cells referred to as taste buds, which are structures embedded in the epithelium. On the tongue, taste buds are a part of morphologically, clearly visible epithelial protrusions and/or invaginations known as papillae. The contact of taste stimuli activates signal transduction cascades that result in the depolarization of the receptor cells and the release of the neurotransmitter ATP. ATP excites afferent nerves and allows taste information to be transmitted to the cerebral cortex, where neuronal activity creates the sensory perception (3, 4). In this scenario the taste receptor molecules have the important task of chemical recognition and discrimination. Organisms use these receptor molecules to convert chemical structures into biochemical reactions and, ultimately, to perceive taste.

In recent years, impressive progress has been made in the field of gustation, because of the discovery of the receptors for sweet, umami, bitter, and sour taste and the experimental tools that were created. Our objective here is to review the recent developments in the field with emphasis on taste receptors and their associated biochemical signal transduction cascades.
Sweet Taste

Sweet taste is elicited by many compounds of various chemical classes (Fig. 1). Sweeteners include monosaccharides and disaccharides such as glucose and sucrose; amino acids such as L-tryptophane, alanine, and glycine; proteins such as monellin and thaumatin; and many chemically diverse artificial sweeteners such as saccharin, cyclamate, aspartame, and alitame (5). This observation has fostered long-lasting speculations about how many receptors are necessary to detect the many structurally divergent compounds. Finally, an answer to this question was provided by the discovery of the TAS2R genes that encode a new family of the putative taste receptors (6). The gene family consists of the three members: TAS2R1, TAS2R2, and TAS2R3. TAS2R is the gene symbol proposed by the human genome project nomenclature committee for the gene family previously referred to as TRs; the corresponding mouse gene symbol is Tas1r. Genes are written in italics, whereas the protein is printed in normal letters. Many studies have demonstrated convincingly that the TAS2R2:TAS2R3 heteromer mediates the majority of human sweet taste perception.

The TAS2Rs are subclass 3 G-protein-coupled receptors that are related distantly to the Ca^2+ sensing receptor, the metabotropic glutamate receptors, GABA B and, and the V1R pheromone receptors. Consistent with their proposed role as taste receptors, behavioral experiments and neuronal recordings of Tas1r2 and Tas1r3 knockout mice showed that the deletion of either gene reduced strongly the nerve responses and the attractiveness of various sweeteners, whereas the deletion of the third family member Tas1r1 affected umami taste (7). Moreover, in situ hybridizations in rodents showed that Tas1r3 is co-expressed with either Tas1r2 or Tas1r3 in two nonoverlapping subsets of taste receptor cells (8). This finding led to the hypothesis that the functional sweet receptor could be a heteromer of two Tas1r subunits. Indeed, expression studies in HEK293 cells demonstrated subsequently that cells cotransfected with Tas1r2 and Tas1r3 or its human counterparts responded to various sweeteners (8, 9) and thereby confirmed that Tas1r2 and Tas1r3 form a functional sweet taste receptor. Both the human and the rodent receptors are activated by chemically diverse sweeteners such as monosaccharides, disaccharides, sweet amino acids, and artificial sweeteners (9). Notably, all tested compounds that taste sweet to humans activate the human TAS2R2:TAS2R3 receptor (9).

Interestingly, the human TA2R2:TA2R3, but not its mouse counterpart, are sensitive to the sweet proteins monellin, thaumatin, and brazzein, and to the artificial sweeteners neotame, cyclamate, and aspartame (9-11). This difference provides a molecular explanation for the previous observation that these compounds are sweet for humans but not attractive to rodents (9). The species difference also applies to the sweet inhibitor lactisole that blocks the sweet taste in humans but not in rats, and only inhibits the response of human TAS2R2:TAS2R3 to sweet stimuli (9).

Recently, these functional differences between the human and the rodent sweet receptor have been exploited to obtain insight into how this receptor can be activated by so many structurally different sweeteners. Replacement of the large, extracellular domain at the N-terminus of rat Tas1r2 by its human counterpart was sufficient to create a chimeric receptor that could be activated by the dipeptide derivatives aspartame, neotame, and the sweet-tasting protein monellin (10, 11). Similarly, replacement of the cysteine-rich region in mouse Tas1r3, which connects the N-terminal extracellular domain to the heptahedral domain by its human counterpart, created a receptor chimera that could be activated by the sweet protein brazzein (11). These findings suggest that the binding sites for aspartame, neotame, and monellin are located in the large extracellular domain of TAS2R1, whereas the binding site for brazzein may be located in the cysteine-rich domain of the TAS2R3 subunit. Additional analyses of receptor chimeras in combination with mutational studies and molecular modeling revealed that the sweet inhibitor lactisole and the sweetener cyclamate share an overlapping binding site in the heptahedral domain of the human TA2R3 subunit (12). Moreover, tryptophan fluorescence spectroscopy analysis of the purified extracellular N-terminal domains of Tas1r3 and Tas1r2 support the notion that sucrose, glucose, and sucralose interact with both domains (13). In summary, these results provide evidence that structurally diverse sweeteners use multiple binding sites to activate the sweet receptor (Fig. 1).

Umami Taste

In humans, umami taste (also referred to as amino acid taste) is elicited predominantly by L-glutamate and L-aspartate (14), whereas rodents respond to most L-amino acids (7). Interestingly, umami taste is enhanced by 5′-ribonucleotides such as inosine-5′-monophosphat (IMP) and guanosin-5′-monophosphat (GMP) (15). Thus, a genuine umami receptor should reflect these properties. Umami compounds are enriched during the ripening processes in many foods, including fruits, vegetables, cheese, and meat. Therefore, this taste quality helps us to choose the ripest fruits and the most palatable cheese for our meal.

In humans, some metabotropic glutamate receptor agonists such as ibotenate and L-AP4 elicit umami taste (15). Moreover, studies have demonstrated the expression of mGluR4 variants in taste buds of different species (16-21). These observations are consistent with the hypothesis that metabotropic glutamate receptors contribute to umami taste. In line with this assumption, the cDNA of an N-terminally truncated "taste" variant (mGluR4t) of the mGluR4 was isolated from rodent tongue tissue (20). Functional studies showed that it could be activated by L-AP4 and glutamate at concentrations that are typical for umami taste (20). Based on these data the truncated mGluR4 variant initially seemed to be an attractive candidate for an umami taste receptor, although several inconsistencies exist [c.f. (7)]. Most prominently, mGluR4 knockout mice show an increased preference for glutamate (22) instead of a reduced response as one would expect. Therefore, receptor activities cannot be enhanced by ribonucleotides.

Studies of Tas1r1 and Tas1r3 knockout mice provide clear evidence for their involvement in umami taste. The deletion of either receptor gene reduced strongly the attractiveness of the taste quality.
Figure 1  Schematic presentation of proposed binding sites for structurally different sweet tasting compounds at the human sweet taste receptor. Identification of the binding sites for aspartame, neotame, monnellin, brazzein, cyclamate, and lactisole are based on functional analysis of the rodent and human sweet receptor, chimeras created of them and specific receptor mutants, as well as data obtained through molecular modeling. Information about the sucrose, glucose, and sucralose binding sites have been derived from measuring agonist-induced conformational changes of the purified N-terminal ectodomains of the sweet receptor (see text for further details).

umami compounds in mice and in the corresponding nerve responses (7). In situ hybridizations in rodents demonstrated that Tas1r3 is coexpressed with Tas1r1 in a subset of taste receptor cells, which suggests that the umami receptor is a heteromer of TAS1R1 and TAS1R3. In vitro expression studies showed that cells cotransfected with cDNAs for human TAS1R1 and TAS1R3 responded to glutamate, aspartame, and l-AP4 (9), whereas cells transfected with the counterparts from mice acquired general sensitivity for l-amino acids (23). Remarkably, 5′-ribonucleotides such as IMP and GMP enhanced strongly the receptor responses, which is a hallmark of umami taste (9, 15). Thus, these functional properties of the TAS1R1/TAS1R3 receptor dimer of humans and rodents explain some of the most important properties of umami taste. It should be pointed out, however, that the response profiles of all the umami receptor candidates in transfected cells, i.e., TAS1R1/TAS1R3 and the various mGluRs found in taste tissue, differ from those observed in native taste cells (24). A complete description of umami taste transduction may involve combinations of the candidate receptors and/or as yet undiscovered taste receptors (24), or umami taste may be a delicious flavor formed by neuronal mechanisms in the brain (25).

Bitter Taste

Bitter compounds are numerous and structurally diverse (26). Estimates count thousands of these compounds in the human environment. Known bitter-tasting substances include fatty acids, peptides, amino acids, amines, azacycloalkanes, N-heterocyclic compounds, amides, ureas, thioureas, esters, lactones, carbonyl compounds, phenols, crown ethers, alkaloids, and metal ions. In mammals, these compounds are recognized by approximately 30 G-protein-coupled receptors belonging to the TAS2R gene family (27). These comparably few receptors face the enormous challenge to sense the numerous synthetic and natural bitter substances. One of the central questions in bitter taste research is how these few receptors enable the detection of so many different bitter-tasting compounds. Detailed knowledge about the molecular basis of TAS2R-tastant interaction is required to answer this question. As a first step toward a better understanding, the identification of as many bitter receptor-bitter agonist pairs as possible is necessary to establish a foundation for detailed structure-function analyses. The successful development of a variety of functional expression assays led to an enormous boost of deorphanizations of bitter taste receptors during the last few years. One of the difficulties in setting up efficient screening assays is the insufficient cell-surface targeting properties of
bitter taste receptor proteins (28), which have been observed for other chemosensory receptor gene families, such as odorant receptors (29) and pheromone receptors of the V2R type (30). These problems are circumvented commonly by the amino terminal extensions of taste receptors with amino termini of other GPCRs such as bovine rhodopsin (28) or rat somatostatin receptor 3 (31). The physiologic cell-surface targeting properties of hTAS2Rs seem to be individual and may depend on various cofactors (32). With such assays in place, nine human (see also Reference 33), two mouse, and one rat TAS2Rs have been de-orphanized (see Table 1) by various laboratories and different experimental approaches to date.

The first mammalian receptor-bitter agonist combinations identified was human TAS2R4 and mouse T2R5 (28). Both receptors responded selectively to one compound of a large panel of known bitter substances best. Whereas mT2R5 was only activated by cycloheximide, hTAS2R4 responded to denatonium benzoate and high concentrations of 6-n-propyl-thiouracil. This study indicated that mammalian TAS2Rs exhibit a strong selectivity for agonists, although, in the case of hTAS2R4, limited promiscuity might occur at high concentrations. The first hTAS2R identified to be activated by natural bitter substances was hTAS2R16 (31). Systematic testing of substances demonstrated that this receptor responded selectively to an entire group of chemically related compounds, the β-glucopyranosides. Thus, hTAS2R16 combines selectivity, even stereoselectivity for the β-p-glucopyranosides, with high concentrations of chemicals and that this information is not modified robustly by neuronal computation. This conclusion is supported strongly by experiments in transgenic mice. Mice are indifferent to β-glucopyranosides, such as salicin, but they taste this compound with similar sensitivity as humans do when they express the human cognate bitter taste receptor hTAS2R16 as transgene.

The more detailed structure-function analyses of several TAS2Rs together with their agonists will provide detailed insight into bitter receptor-agonist interactions required to understand how such few TAS2Rs can recognize so many bitter compounds and might pave the way for the development of bitter antagonists. The recent availability of computer modeling studies of bitter taste receptors with identified agonists docked into these structures increases our knowledge about structure-function relations (40, 41) and may help to guide future mutagenesis analyses.

Salt Taste

Two pathways for salt taste have been reported. Nerve recordings performed in rodents showed that the chorda tympani nerve, which innervates the fungiform papillae of the anterior tongue, responded strongly to stimulation with NaCl and that this effect was highly sensitive to amiloride (42, 43). The amiloride-sensitive response was selective for Na⁺ ions. Based on these observations, the non-voltage-gated, sodium-permeable, heteromeric (αβγ) epithelial sodium channel (ENaC) has been suggested to be a good candidate. In rodents, ENaC subunits are expressed in a specific subset of fungiform taste receptor cells. Whole-cell patch clamp analysis of isolated fungiform taste receptor cells demonstrated that amiloride inhibited Na⁺-induced currents in micromolar concentrations as expected for ENaC-mediated currents (44, 45). Furthermore, behavioral studies in mice and in rats showed that sodium taste is impaired partly by adding amiloride to a sodium salt solution, without affecting responses to other taste modalities (46, 47). In humans, the situation is less clear. Psychophysical analysis revealed only a limited reduction of salt taste by amiloride, and this seemed to be restricted to few subjects (48, 49).
<table>
<thead>
<tr>
<th>TAS2R</th>
<th>Agonists</th>
<th>Agonist structure</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>hTAS2R4</td>
<td>denatonium benzoate, 6-n-propyl-2-thiouracil (PROP)</td>
<td><img src="image" alt="denatonium benzoate" /></td>
<td>28</td>
</tr>
<tr>
<td>hTAS2R7</td>
<td>chloroquine, papaverine, quinacrine, strychnine</td>
<td><img src="image" alt="chloroquine benzoate" /></td>
<td>37</td>
</tr>
<tr>
<td>hTAS2R10</td>
<td>strychnine</td>
<td><img src="image" alt="strychnine benzoate" /></td>
<td>31</td>
</tr>
<tr>
<td>hTAS2R14</td>
<td>aristolochic acid, 1,8-naphthalaldehydic acid, 1-naphtholic acid, 1-nitronaphththalene, picrotin, picrotoxinin, piperonylic acid, sodium benzoate, (−)-α-thujone</td>
<td><img src="image" alt="aristolochic acid" /></td>
<td>36, 37</td>
</tr>
<tr>
<td>hTAS2R16</td>
<td>phenyl-β-glucopyranoside, salicin, helcin, arbutin, 2-nitrophenyl-β-glucopyranoside, naphtyl-β-glucopyranoside, methyl-β-glucopyranoside, amygdalin, esculin</td>
<td><img src="image" alt="phenyl-β-glucopyranoside" /></td>
<td>31</td>
</tr>
<tr>
<td>hTAS2R38</td>
<td>Phenylthiourea (PTC), diphenylthiourea, acetylthiourea, propylthiourea (PR09), methylthiourea</td>
<td><img src="image" alt="Phenylthiourea" /></td>
<td>34, 35</td>
</tr>
<tr>
<td>hTAS2R43</td>
<td>Acesulfame K, aristolochic acid, saccharin, 6-nitrosaccharin, N-isopropyl-2-methyl-5-nitrobenzene sulfonamide (IM NB)</td>
<td><img src="image" alt="Acesulfame K" /></td>
<td>38, 39</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>TAS2R</th>
<th>Agonists</th>
<th>Agonist structure</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>hTAS2R44</td>
<td>Acesulfame K, aristolochic acid, saccharin</td>
<td><img src="image1" alt="chemical structure" /></td>
<td>38</td>
</tr>
<tr>
<td>hTAS2R47</td>
<td>Denatonium, 6-nitrosaccharin</td>
<td><img src="image2" alt="chemical structure" /></td>
<td>39</td>
</tr>
<tr>
<td>mT2R5</td>
<td>cycloheximide</td>
<td><img src="image3" alt="chemical structure" /></td>
<td>28</td>
</tr>
<tr>
<td>rT2R9</td>
<td>cycloheximide</td>
<td><img src="image4" alt="chemical structure" /></td>
<td>31</td>
</tr>
</tbody>
</table>

The depicted chemical structures correspond to the agonists printed in bold.

In vallate and foliate papillae of rodents only α-ENaC could be easily detected, whereas β- and γ-ENaC are less abundant (50–52), which raises questions about the identity of the salt taste receptor of the posterior tongue. Moreover, NaCl-induced responses of the glossopharyngeal nerve that innervates the vallate and foliate papillae of the posterior tongue seemed to be almost insensitive to amiloride (43, 53). The amiloride-insensitive salt taste receptor is a constitutively active ion channel that is blocked by cetylpyridinium chloride. It is not selective for sodium ions but mediates H^+ and K^+ currents (54). Based on its sensitivity to the TRPV1 antagonists SB-366791, it has been suggested that the amiloride-insensitive salt taste receptor is a variant of the vanilloid receptor 1, TrpV1t (55). But because not all properties of amiloride-insensitive salt taste receptor are replicated by TRPV1 and TRPV1 gene-targeted mice preferred NaCl over water at concentrations avoided by wild types and salt taste in these animals was less blocked by amiloride (55, 56), the role of TrpV1t in salt taste remains questionable. Taken together, the molecular identity of the salt taste receptor or receptors cannot be taken for granted.

Sour Taste

Sour taste detects acids, i.e., protons. Several different sour taste receptor candidates such as acid-sensing ion channels (ASICs) (57), hyperpolarization-activated cyclic nucleotide-gated channels (HCNs) (58), and two pore domain potassium channels (K2Ps) (59, 60) have been described in the past. In addition, recent research identified two members of the polycystic kidney disease (PKD) family of the transient receptor potential superfamily (TRP) as strong sour taste “receptor” candidates or as part thereof. Immunohistochemistry and in situ hybridization revealed the presence of the polycystic-kidney-disease-like ion channel PKD1L1 in subsets of taste receptor cells of mouse fungiform, vallate, and foliate papillae. These cells differ from...
those expressing receptors for bitter, sweet, and umami taste (61, 62). Moreover, genetic ablation in mice of the cells expressing PKD1L2 was associated with a loss of response to acidic stimuli in both electrophysiologic recordings and behavioral experiments, whereas other taste modalities remained unaffected, which suggests that PKD1L2 cells are necessary for sour taste perception (62). Previous studies demonstrated that PKD2 polypeptides need to interact with PKD1 polypeptides for proper cell-surface expression (63, 64). Search for interaction partners for PKD1L2 in taste cells identified PKD1L3 in mouse valiate and foliulate papillae being coexpressed with PKD2L1 in the same cells (62, 65). PKD1L3 enhanced significantly the expression of PKD2L1 at the cell surface in vitro (61). In fungiform papillae and in the palate, expression of PKD1L3 could not be detected, which suggests that another polypeptide interacts with PKD2L1 in these structures. In accordance with the proposed involvement of PKD2L1/PKD1L3 in sour taste transduction, cells that express the two polypeptides responded to stimulation with acids (61). The data were consistent with the observation that sour taste in humans and rodents is stronger for weak acids, such as citric acid, than for strong acids, such as HCl (66, 67). However, subtle, important differences between the heterologously expressed PKD channel and the sour taste responses in native taste cells (68) require additional work to determine the precise role of the PKDs in sour taste transduction.

Molecular Evolution of Taste Receptor Genes

Taste reception varies enormously across vertebrates. Because the taste perception of a species is related essentially to its diet and environment, the studies of the variation of genes, which control taste reception among vertebrates will contribute greatly to our understanding of the relationship between the adaptation of organisms and the diversity of environmental chemicals.

Evolution dynamics of TAS2R gene family

To date, the complete repertoire of the TAS2R gene family has been reported in human, mouse, rat, dog, cow, opossum, chicken, frog, and several fish (69-72), in addition to a small number of TAS2R genes described in several primates (73-76). Comparisons of these gene repertoires revealed extremely high variation in the sizes of the TAS2R repertoire among species, ranging from 3 genes in chicken to 20-50 genes in mammals and amphibians (71, 72). This finding is consistent with the fact that bitter taste perception, as a warning sensor for bitter toxins, may suggest that detecting poisons in diet is not as important in ruminants as in other animals because of the high detoxification capacity of cow’s rumen microbes (72).

In terms of long-term evolution of the TAS2R gene family, the phylogenetic analysis showed several interesting evolutionary patterns (Fig. 2). First, the TA2SR gene family evolves after the birth-and-death process (71), which is characterized by frequent gene duplication and gene deactivation, similar to that found in the olfactory receptor gene family (77). Second, there might be multiple TAS2R genes in the common ancestor of tetrapods and teleosts (72). Third, the TAS2R repertoire expanded considerably in the common ancestor of tetrapods, followed by additional independent expansions in frogs and mammals, and contractions of TAS2R repertoire occurred in chicken (71, 72). Last but not least, based on a comparison of human and mouse TAS2R gene repertoires, some TA2SR genes exhibit one-to-one orthologous pairing, whereas other genes are part of lineage-specific (species-specific) clusters, in which the genes from the same species cluster together in the phylogenetic tree (78). These species-specific genes are also located closer to each other on chromosomes, which indicates that these newly duplicated genes resulted from tandem gene duplications (70, 78). Also, these genes seem to be under positive selection, which suggests that they are used for species-specific bitter tastants (78). On the other hand, one-to-one orthologous genes are subject to more selective constraints than lineage-specific (species-specific) genes, which indicates that each of the one-to-one orthologous genes possibly is detecting one or several distinct bitter tastants that are encountered by a wide range of animals (78). Although two recent evolutionary studies (71, 72), which extended the study of TAS2Rs outside of human and mouse to an additional nine vertebrate species, supports this hypothesis, it still awaits to be scrutinized additionally by functional research.

The comparative analysis of the TAS2R gene family between several primate species with rodents revealed that primate genes were under less selective pressure than rodent genes (73-76). First, the comparison of the gene birth/death rate between primates and mice shows that the proportion of pseudogenes in the TA2SR repertoire is lower in mice (15%) than in apes (21%-28%), which is in turn lower than that in humans (31%) (73, 74). Moreover, the prevalence of lineage-specific pseudo-genes in primates supports this conclusion (74). Second, based on the equal levels of nonsynonymous/synonymous substitution rate ratios for the TAS2R genes in primates, the functional constraints were more relaxed in the primate lineage than in the mouse lineage (73, 74, 76). This evolutionary pattern could be caused by the reduced effective population sizes in primates, which might cause less-effective purifying selection (73). The alternative explanation is that the reduced functional constraints in primates might be caused by reduced bitter taste needs because of a change of the environment and the diet (74). In fact, some ecologic studies support this explanation. For example, meat accounts for 2-13% of diet in chimpanzees, whereas it has never been found in other apes’ diet (76). Furthermore, this explanation has been strengthened by the findings that there were significant changes in human diet, such as increasing food from animal sources while decreasing food from plant sources, and the controlled use of fire to detoxify the food (76). Both factors may have caused a reduction in the importance of bitter taste and consequently triggered a functional relaxation in humans.

Evidence for the relaxation of selective constraints on TAS2R genes in apes and humans does not preclude the possibility
positive selection occurred on a few specific genes. Positive selection has been found in the gene for the human TAS2R16, the beta-glucopyranoside receptor (31). By analyzing the sequences from 60 human populations, Soranzo et al. detected signatures of positive selection on a more sensitive derived allele, which was found in all human populations except for African populations (79). This result might reflect the increased sensitivity of the derived TAS2R16 allele under the positive selection through an increased protection against harmful cyanogenic plant foods and natural toxins (79). In addition to the selective relaxation and positive selection, the most complex scenario of evolutionary forces has been observed in TAS2R38 gene, which is responsible largely for the human polymorphism in tasting phenylthiocarbamide (PTC) (34). Interestingly, chimpanzees are also known to have tasters and nontasters of PTC (80). Although humans and chimpanzees shared phenotypic polymorphism, they did not share the same evolutionary forces of maintaining non-taster alleles (80). By contrast, the nontaster allele was lost in chimpanzees, which favors the selective relaxation hypothesis in this lineage (80). As more human TAS2R genes are being studied, the understanding of the evolutionary forces behind each TAS2R will increase considerably.

Evolution dynamics of TAS1R gene family

In contrast to the TAS2R gene family, the TAS1R family is remarkably well conserved during evolution both in gene family size and in sequence divergence (72). In terms of gene family size, the number of TAS1R gene repertoire changes rarely in mammals, which might reflect the necessity of both sweet and umami tastes among mammals (72). But the number of TAS1R gene repertoire varies in some nonmammalian vertebrates, both with a few gene duplications observed in pufferfish and fugu and with gene loss events found in western clawed frog and chicken (72). In addition to the western clawed frog, which does not have any TAS1R genes, a loss of the TAS1R2 gene was identified in the chicken genome (72). In addition, cats and closely related carnivores are also known to lack the TAS1R2 genes (82), which might reflect the insensitivity to sweet taste stimuli in these species (72, 82). Thus, pseudogenization of TAS1R2 occurred multiple times independently in evolution.

In sequence divergence level, TAS1R genes evolve more slowly than TAS2R genes whatever the comparison between
species or within species. For interspecies comparison, the sequence divergence distance of orthologous pairs among human, mouse, rat, and opossum is significantly lower for TAS1R genes than TAS2R genes (72). Similarly in human populations, the mean pairwise differences per nucleotide between sequences of TAS2R are greater than those of TAS1R, which reveals the lower levels of nucleotide diversity in TAS1R family (83, 84). The positive selection has been suggested to operate separately on paralogous TAS1R genes and different alleles (72, 84), which is also the case in TAS2R genes, as mentioned above.

**Extragustatory Expression of Taste Receptors**

The role of taste perception in the oral cavity is to analyze the composition of food for its nutritional value and for the presence of potentially harmful substances prior to ingestion. However, evidence is accumulating that elements of the taste perception machinery, including taste receptor proteins, are expressed at several extragustatory sites as well. The anatomical organization of chemosensory structures is variable, apparently becoming less complex with growing distance from the primary gustatory areas. Taste buds located in gustatory papillae on the surface of the tongue, on the soft palate, and on the oro-pharynx are well-organized groups of about 60 to 100 cells (1). Laryngeal areas. Taste buds located in gustatory papillae on the surface of the tongue, on the soft palate, and on the oro-pharynx are well-organized groups of about 60 to 100 cells (1). Laryngeal taste buds, however, are smaller than lingual buds. Sbarbari et al. (86) observed that the sizes and shapes in rats changed from the most rostral part of the laryngeal inlet, where mostly buds were found, to structures called “chemosensory clusters” distally. These chemosensory clusters contained only 2-3 cells staining positive for PLC-α2, which is a molecule involved critically in sweet, umami, and bitter taste transduction (85). Rarely, solitary chemosensory cells were found in this part of the larynx but they become more numerous distally (86) and extend into the airway epithelium (87). Solitary chemosensory cells are also present in respiratory nasal epithelium and the vomeronasal epithelium (88, 89). A notable type of cells that express components of taste signal transduction are the brush cells, which line the stomach, the duodenum (90), and the pancreatic duct system (91). It should be noted here that, by anatomical criteria, brush cells might be related to, but they are not solitary chemosensory cells (92). Recently, secretory cells of the airway (93) and spermatozoa (94) have been identified to express taste signaling components as well.

Within the gastrointestinal tract of rodents, brush cells of the stomach, duodenum (90), and the pancreatic duct system (91) express the G-protein subunit α-gustducin, which has been demonstrated to be important for bitter, sweet, and umami taste transduction (95). By RT-PCR analyses of gastrointestinal tissues of rats and mice, a considerable number of TAS2R genes have been detected, although the cellular origin of the detected mRNAs was not directly addressed (96). A recent study identified TAS1R1, TAS1R2, TAS1R3, PLC-α2, and TRPM5 along with α-gustducin. In case of TAS1R2, expression was only weak and more restricted, which suggests that most components of the canonical sweet and umami taste transduction are present in the GI tract. However, using the transgenic expression of GFP under the control of the TRPM5 5′-flanking region for colocalization, experiments with the other taste transduction molecules revealed a less clear picture. Induction of TRPM5-driven GFP expression with PLC-α2 is limited, both molecules are indispensable for bitter, sweet, and umami taste transduction as shown by knockout mouse models, which suggests that at least in part a variant transduction mechanism acts in these cells. Moreover, this study identified that, in addition to brush cells, enterendocrine cells are positive for TRPM5-driven GFP expression (97). With respect to bitter taste receptor-mediated signal transduction, however, the mouse intestinal cell line STC-1, which expresses bitter taste receptor genes along with α-gustducin and β-transducin, reacts selectively to various bitter stimuli with transient calcium signals, which indicates the canonical transduction mechanism (96). As chemosensory cells of the gut are not innervated directly (90), and do not express the presynaptic marker SNAP25 (97), it is speculated that they might communicate via NO (98) or a currently unidentified messenger with neighboring cells/nerve fibers to control appetitive behavior or food passage through the gastrointestinal tract.

Morphologically, the α-gustducin expressing cells found in the nasal and vomeronasal epithelia resemble solitary chemosensory cells (SCCs) described in nonmammalian vertebrates (88). In addition to α-gustducin, some bitter taste receptor genes are expressed in SCCs, whereas the TAS1R1 and TAS1R2 subunits, which specifically constitute the sweet and umami receptor heteromers TA-S1R2/TAS1R3 and TAS1R1/TAS1R3, respectively, are absent from nasal chemosensory cells. It indicated already that, analogous to the warning function of bitter taste receptors, cells of the oral cavity, the function of TAS2R expression in nasal chemosensory cells in nasal respiratory epithelium might protect the animal from the aspiration of noxious substances. Indeed, intranasal irritation with bitter compounds not only elicited trigeminal responses but also resulted in pronounced respiratory effects (88). Another recent example for extragustatory expression and multiple functions of taste receptor molecules aside from pure gustation is PKD2L1, a mammalian sour taste sensor, which is also expressed in a discrete population of neurons surrounding the central canal of the spinal cord, perhaps involved in monitoring the pH of the cerebrospinal fluid (92).

In perspective, the growing number of reports on the extragustatory expression of components of the taste transduction cascade requires the careful analyses of taste-specific knock out models for nongustatory deficits.

**Perspectives**

During the last few years, the field of taste research has been progressing with enormous speed. Current research addresses structural details on receptor agonist interactions for bitter and sweet taste receptors. For the other taste modalities, including salt taste, where definitive proof for the proposed roles of candidate sensor molecules is still missing, more questions still need to be answered. A next big step in taste research will be to understand how taste information is processed along its
pathway from the periphery into higher order brain centers. In view of emerging extracellular functions of taste transduction molecules, it will be critical to understand how the different functions and cellular environments have modified the involved proteins and signaling mechanisms to serve multiple functions. More practically, the cloning of the taste receptors enabled recombinant receptor assays to be established. With these tools, the flavor industry will be able to design high throughput screens of complex compound libraries in order to identify new taste-active compounds. Sweet taste enhancers, glutamate substitutes and salt replacers, and specific bitter blockers might be the primary targets.

References
(continued from previous page)
85. Zhang Y, Hoon MA, Chandrasekhar J, Mueller KL, Cook B, Wu D, Zuker CS, Ryba NJ. Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. Cell 2003;112:293–301.
Bacterial Resistance to Antibiotics

Gerard D Wright, Antimicrobial Research Centre, Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada
doi: 10.1002/9780470048672.wecb034

Microbes have evolved a myriad of mechanisms to overcome the toxic effects of antibiotics, which include the production of enzymes that modify or degrade antibiotics, complex membrane-associated efflux systems that can pump antibiotics out of the cell, modification of antibiotic targets, and the production of immunity proteins. The biochemical logic of resistance is often intimately linked with the mode of action of the antibiotics. As a result, chemical biology approaches for understanding resistance not only have application in our understanding of this phenomenon, but also they can guide the generation and deployment of new antibiotics. In this review, a survey of the chemical strategies employed by bacteria to resist antibiotics is presented with an emphasis on the molecular mechanisms of resistance enzymes and proteins.

Modern anti-infective chemotherapy is founded on the deployment of an arsenal of potent and safe small-molecule inhibitors of microbial growth. As a result, it is not an understatement to say that antibiotics and antibacterial agents have revolutionized health care during the last century. These compounds are, respectively, natural product and synthetic small molecules that inhibit bacterial growth. For simplicity, in this review, the term “antibiotic” will be used to refer to both natural products and synthetic molecules.

Since the first articulation of the concept of the “magic bullet” by Paul Ehrlich to describe the discovery and use of small-molecule inhibitors of microbial growth (1), there has been a concerted effort to identify new antibiotics in chemical libraries and from natural sources. Following Ehrlich’s efforts, the synthetic sulfa drugs that inhibit p-aminobenzoic acid biosynthesis dominated anti-infective therapy in the 1930s and early 1940s. The discovery of the highly potent fungal natural product penicillin, followed by peptide antibiotics such as gramicidin from soil bacteria (2), ushered in a period of roughly 15 years of intense natural product drug discovery that revealed almost all the chemical classes of antibiotics in current clinical use.

As has been well documented elsewhere (3, 4), there was a roughly a 25-year gap between the introduction of the quinolone antibiotics and the next new chemical class of antibiotics, the oxazolidinones. During this interval, innovation in antibiotic drug design and discovery was focused on the semisynthetic tailoring of natural product antibiotic scaffolds to improve pharmacologic properties and, most importantly, to overcome resistance to existing antibiotics.

Since the first use of antibiotics, bacterial insensitivity to these cytotoxic agents, both intrinsic and acquired, has been observed. Antibiotic resistance can be the result of an inability of the compound to enter the cell, active efflux from the cytosol, mutation or alteration of the primary molecular target, sequestration of the antibiotic, and enzymatic destruction or chemical modification of the compound (Table 1). Each of these mechanisms requires unique chemical strategies to achieve the same biological outcome: continued cell growth in the face of toxic compounds. This review will discuss the principal mechanisms of antibiotic resistance emphasizing the biochemical logic of the strategy to overcome the cytotoxic effects of antibiotics.

Biological Context of Antibiotic Resistance

Antibiotic resistance is a major clinical problem with great impact on the successful treatment of infectious disease (5). Resistance increases mortality and morbidity, and it lengthens hospital stays. Health-care costs associated with antibiotic-resistant infections range between $6000 and $30,000 USD per patient (6). Furthermore, the rate of resistance continues to rise in all important human pathogens year after year (7). Limits on the use of certain antibiotics and drug cycling can help in some cases to mitigate resistance; however, even when restricted-use protocols are in place, resistance continues to be a problem (8, 9). Therefore, once resistance emerges, it is a continuing clinical difficulty that must be managed.
Bacterial Resistance to Antibiotics

Table 1 Molecular mechanisms of antibiotic resistance

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Example antibiotics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzymatic</td>
<td>β-Lactams, Fosfomycin</td>
</tr>
<tr>
<td>Inactivation (e.g., Hydrolysis)</td>
<td></td>
</tr>
<tr>
<td>Covalent Modification</td>
<td>Aminoglycosides, Chloramphenicol</td>
</tr>
<tr>
<td>Oxidation/Reduction</td>
<td>Tetracycline</td>
</tr>
<tr>
<td>Altered transport</td>
<td>Tetracycline, Fluoroquinolones</td>
</tr>
<tr>
<td>Efflux</td>
<td></td>
</tr>
<tr>
<td>Reduced Uptake</td>
<td>Aminoglycosides, Chloramphenicol</td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
<tr>
<td>Sequestration</td>
<td>β-Lactams, Enediynes</td>
</tr>
<tr>
<td>Immunity Proteins</td>
<td>Bleomycin, Tetracycline</td>
</tr>
<tr>
<td>Target Modification</td>
<td>Fluoroquinolones, Macrolides</td>
</tr>
<tr>
<td>Altered Metabolic Pathway</td>
<td>Vancomycin</td>
</tr>
<tr>
<td>Target Overexpression</td>
<td>Trimethoprim</td>
</tr>
</tbody>
</table>

Intrinsic antibiotic resistance

Paradoxically, it is the remarkable potency of these drugs that is their Achilles heel. Resistant bacteria are selected by exposure to antibiotics, which nondiscriminately kill off susceptible organisms. The resultant and powerful “adapt or die” evolutionary pressure selects for microbes with the ability to evade the toxic effects of these agents. This antibiotic insensitivity can be intrinsic, i.e., a consequence of the genetic or physiologic makeup of microorganisms, or it can be acquired from other sources, generally through the aegis of mobile genetic elements.

Intrinsic antibiotic resistance

Microbial metabolism is largely contained within the cell envelope and the cell interior. The intracellular physiology and biochemistry (ribosomes and protein translation, nucleic acid replication and transcription machinery, metabolic pathways, etc.) are, for the most part, conserved throughout the bacterial kingdom. On the other hand, the bacterial cell envelope is more structurally diverse. In particular, the presence of a relatively impermeable outer membrane in Gram-negative bacteria is a barrier to many classes of antibiotics. The bacterial outer membrane comprises an asymmetric bilayer with a phospholipid interior and a lipopolysaccharide (LPS) exterior. This anionic LPS layer is bridged through cationic metals providing an additional physical diffusion barrier for small and large molecules. Access to the interior of the cell is provided by a series of membrane proteins (e.g., porins) that recognize and allow the transport of nutrients and other metabolites. Antibiotics such as the glycopeptides, macrolides, and others have difficulty penetrating the outer membrane. As a result, Gram-negative bacteria are consequently intrinsically resistant to these drugs.

The absence of a sensitive microbial target is also a form of genetic intrinsic resistance. The oxirene antibiotic fosfomycin covalently modifies the target protein MurA on a sensitive and catalytically relevant Cys residue (Fig. 1) (10). MurA catalyzes the formation of N-acetylmuramic acid via enoyl-pyruvyl transfer of phosphoenol pyruvate onto the acceptor sugar N-acetylglucosamine, providing the necessary carboxylate anchor for attachment of the pentapeptide required for bacterial peptidoglycan (cell wall) formation. Bacteria that encode a MurA orthologue where the susceptible Cys is replaced by an Asp (e.g., Mycobacteria) are intrinsically genetically resistant to fosfomycin (Fig. 1) (11, 12).

Another confounding physiologic barrier to antibiotic action is the bacterial growth state. Most bacteria can grow in planktonic form where cells are free living and are uniformly exposed to antibiotics. Most antibiotic susceptibility tests are in fact performed on this cell state to determine the minimal inhibitory concentration (MIC) of antibiotic required to arrest cell growth. However, many important pathogens such as Staphylococcus aureus and Pseudomonas aeruginosa can form biofilms on surfaces (e.g., on catheters or mucosal layers of the lung) where most cells are metabolically quiescent and often highly resistant to antibiotics. Furthermore, other organisms can enter eukaryotic cells and thus evade host defense systems. For pathogens such as Mycobacterium tuberculosis and Chlamydia, it is essential for pathogenesis, whereas for others such as the enteric pathogens Salmonella and Escherichia, it exacerbates infection. Once inside the cell, these organisms often shift metabolic activities and can enter a relatively inactive state, sometimes for long periods of time. Successful antibiotic therapy, therefore, requires not only penetration of the eukaryotic cell without associated toxicity, but also sufficient antimicrobial activity to kill the often slow-growing bacterial cells in this environment.

These passive obstacles to antibiotic action complicate treatment of infectious disease in the absence of highly accurate diagnostic tools to identify the infectious pathogen. Intrinsic resistance is problematic but is predictable and easily identified in the preclinical drug discovery stages. As a result, the spectrum of susceptible microbial species is well known before the compound enters the clinic.

Acquired antibiotic resistance

In contrast to intrinsic resistance, which is genetically and physiologically “hard wired,” acquired antibiotic resistance occurs...
Bacterial Resistance to Antibiotics

Figure 1 The action of the antibiotic fosfomycin requires an active-site Cys residue in the target MurA. Substitution of Asp for Cys in some bacterial species results in intrinsic fosfomycin resistance.

Acquired resistance often requires the presence of the antibiotic for selection of resistant species within a susceptible genetic background. Resistance can emerge as a result of mutation of target genes on the chromosome (e.g., point mutations in DNA gyrase that confer fluoroquinone antibiotic resistance (13)) or by the presence of resistance genes that are captured on mobile and associated genetic elements such as plasmids, transposons, and integrons (14). The sequencing of entire bacterial genomes has revealed that these resistance-associated genetic elements can sometimes be components of the normal genetic makeup of a given bacterial species (especially transposons and integrons), but more often they are acquired as a result of selection in the face of antibiotic exposure. Conventional wisdom suggests that acquired resistance results in a genetic and physiologic burden that makes these organisms less fit in the absence of antibiotic selective pressure, although second site mutations can compensate (15).

One of the confounding issues with acquired resistance by way of mobile genetic elements is linkage of multiple resistance genes on a single element. R-Plasmids, integrons, and transposons often carry not only one resistance gene, but several (Fig. 2). As a result, selection of resistance to one antibiotic can inadvertently select for resistance to others, which can be especially problematic in clinical settings, can contribute to failure of antibiotic cycling countermeasure strategies within these institutions, and can add to the difficulty in eliminating antibiotic resistance once it has emerged.

Chemical Biology of Antibiotic Resistance

Antibiotic resistance can be the result of several molecular mechanisms (Table 1). Some of the most important of these mechanisms include enzyme-catalyzed antibiotic inactivation or modification, altered transport such as efflux, and others such as metabolic bypass and sequestration. Each of these mechanisms requires the synthesis of associated proteins to mediate resistance. These are often highly specialized and efficient, and frequently the corresponding genes can be acquired on mobile genetic elements. In this fashion, the antibiotic resistome (16), which is the collection of all resistance genes within the bacterial kingdom, has the potential to be shared and can cross species and genus boundaries.

In a simplistic but useful view, antibiotic molecules comprised a target-specific “warhead” and a bioactive delivery vehicle or scaffold (Fig. 3). The warhead interacts specifically with the target (protein, RNA, DNA, membrane) and forms key molecular interactions with it. These interactions can be covalent in the case of electrophilic chemical fragments such as β-lactams, β-lactones, and oxiranes, or noncovalent through hydrogen bonds, electrostatics, and so on. Consequently, alteration of the antibiotic at the warhead region or of the reciprocally vital antibiotic binding site on the target results in resistance. On the other hand, the delivery vehicle/scaffold portion of the antibiotic generally provides few essential interactions with the molecular target but instead contributes physical properties that enhance bioavailability (e.g., membrane permeability) or imparts structurally important features such as rigidity that ensure appropriate display and structure of the warhead. As a result, the scaffold portion of the antibiotic molecule tends to be more tolerant of chemical alteration. In fact, chemical modification of the molecular scaffold while preserving the antibiotic warhead has been the principal occupation of medicinal chemists since the first antibiotics were isolated from natural sources. The goal of these modifications is to improve pharmacology, evade resistance, and circumvent proprietary structure limitations. As a result, this region of the antibiotic molecule is generally not a principal target for resistance mechanisms; nonetheless, modification of key scaffold regions such as cyclizing bonds can result in resistance.

Next, various biochemical strategies employed by microorganisms to evade antibiotics are discussed with an emphasis on describing the chemical logic of the approach.
Bacterial Resistance to Antibiotics

Use of antibiotic A co-selects for resistance to antibiotics B-E

Figure 2 Integrons, transposons, and R-plasmids often collect multiple antibiotic resistance genes. As a result, selection of one resistance mechanism co-selects for resistance to additional resistance genes. The scheme shows a typical arrangement for an integron with associated integrase-encoding gene (int), a promoter to drive gene transcription (P), and antibiotic resistance elements (A-E).

Figure 3 Antibiotics consist of an active warhead and associated chemical scaffold.

Antibiotic Resistance by Enzymatic Mechanisms

One of the most effective and potent strategies in which microbes have evolved to evade antibiotics is the chemical modification of these cytotoxic compounds. This approach is generally very selective to the specific antibiotic or class, thereby avoiding undesirable side reactions and minimally perturbing normal bacterial metabolism. The general strategy is to destroy or otherwise mask the warhead, which results in resistance. The biochemical logic of several examples is discussed below.

Antibiotic destruction

One of the most potent and physiologically irreversible mechanisms of antibiotic resistance is via chemical destruction of the antibiotic warhead or scaffold. This destruction can occur through several mechanisms, but hydrolytic approaches predominate (Fig. 4). Resistance to the highly important β-lactam antibiotics by β-lactamases is the archetype of this class. The β-lactams include the natural product penicillin and cephalosporin scaffolds as well as the synthetic mono-bactams such as azetronam. The highly strained β-lactam ring is the antibiotic warhead, covalently modifying target transpeptidase and carboxypeptidase enzymes required for the synthesis and maturation of peptidoglycan (Fig. 5). This molecular strategy is analogous to the modification of MurA by fosfomycin as discussed above (Fig. 1), but in this case, covalent enzyme inactivation affects late-stage peptidoglycan biosynthesis on the outside of the cell rather than early-stage intracellular steps. The β-lactamases hydrolyze the thermodynamically activated β-lactam ring irreversibly incapacitating the antibiotics. Numerous β-lactamases have been cloned and characterized (recently reviewed in Reference 17). These enzymes are often secreted into the periplasmic region (the immediate extracellular space of the plasma membrane, bounded by the outer membrane in the case of Gram-negative bacteria) setting up a perimeter defense to intercept the antibiotics before they reach their molecular targets. The associated genes very frequently are on mobile genetic elements, but many bacteria also harbor orthologues within their genome (18).
Bacterial Resistance to Antibiotics

Figure 4 Hydrolytically sensitive antibiotics include the macrolides, oxirane, and β-lactams. The hydrolytically sensitive region is circled in each structural class.

Figure 5 Action of β-lactam antibiotics. Cell wall biosynthetic transpeptidases activate peptidoglycan peptides for cross-linking by formation of a covalent enzyme intermediate (A). The reactivity of the active-site Ser nucleophile is exploited by β-lactam antibiotics such as penicillin G (shown) to form a hydrolytically stable acylenzyme (B). The β-lactamases have mirrored the chemical strategy of proteases to cleave the β-lactam bond. The main β-lactamase families emulate Ser or metallo-protease mechanisms to perform ring-opening hydrolysis of the β-lactam ring, destroying the antibiotic warhead in the process (Fig. 6). The Ser enzymes mimic the first step in peptidoglycan transpeptidase reaction by formation of an acyl-enzyme intermediate. However, unlike the transpeptidase, this intermediate is short lived and subsequent hydrolysis releases the inactive antibiotic (Fig. 6).

Similarly, metallo-β-lactamases share with metallo-proteases activation of a hydrolytic water molecule by interaction with an active-site Zn²⁺ ion. These lactamases have gained clinical prominence in the past few years as a result of their association with carbapenem resistance. The carbapenems, such as meropenem and imipenem, are β-lactam antibiotics that have trans stereochemistry across the 6-5 bond rather than the cis geometry found in most other β-lactams (Fig. 7).
Bacterial Resistance to Antibiotics

β-Lactamases hydrolytically cleave the susceptible β-lactam ring using Ser and metallo-protease mode chemistry.

Comparison of the structures of amoxicillin and imipenem. The change in configuration across the 5-6 bond in imipenem maintains the ability to bind to penicillin binding proteins but provides insensitivity to many Ser-β-lactamases. As a result, these are drugs of choice for treatment of infections caused by β-lactam-resistant bacteria. However, carbapenems are susceptible to metallo-β-lactamases such as IMP and VIM that are becoming widely distributed among Gram-negative bacteria and are a growing clinical concern (19).

Hydrolitic enzymes are also known to inactivate the oxirane antibiotic fosfomycin and macrolides antibiotics such as erythromycin by ring-opening epoxidase and esterase activity, respectively (Fig. 4). The former mechanism is exemplified by the Mn$^{2+}$-dependent enzyme FosX from the soil bacterium Mesorhizobium loti (20). The bulk of clinically important macrolide antibiotic resistance occurs through ribosomal RNA methylation or efflux mechanisms. As a result, macrolide esterases are presently not a major clinical problem. However, the genes encoding this mode of resistance have been found on mobile genetic elements (21–23) and, because they confer high level resistance to this important class of antibiotic, there is cause for concern that this mechanism could increase in frequency.

Conceptually similar, although mechanistically distinct, is Vgb, a class B streptogramin inactivating enzyme. Streptogramins consist of two distinct structural classes of antibiotics: polyketide-peptide hybrids (type A streptogramins) and cyclic hexa or hepta depsipeptides (type B streptogramins). The type B streptogramins are cyclized through an ester linkage between the C-terminal carboxyl group and the secondary hydroxyl group of an invariant Thr residue at position 2. Both classes of streptogramins bind to the large subunit of the bacterial ribosome blocking the peptide transfer center (type A) and the peptide exit tunnel (type B). The warhead of the type B streptogramins consists primarily of the 3-hydroxyxypicolinic acid residue and the peptide backbone that blocks the entrance to the tunnel. Resistance to the type B streptogramins can occur through target modification or efflux, but also by the enzyme Vgb and orthologues (24). This enzyme linearizes the peptide backbone at the antibiotic’s thermodynamic weakest point, the ester linkage; however, unlike macrolide esterases, Vgb is not a hydrolase but a C-O lyase (25). The catalytic strategy likely involves acidification of the Thrα-proton by generation of the aci form of the amino acid followed by elimination of the carboxylate resulting in C-O bond cleavage and formation of the 2-amino-but-2-enic acid amide product (Fig. 8).

Antibiotic modification

Chemical modification of key groups on the antibiotic can also result in resistance. Modification generally occurs on the warhead region of the molecule, blocking effective interaction with...
Bacterial Resistance to Antibiotics

Figure 8 Inactivation of type B streptogramin antibiotics by Vgb lyase. The type B streptogramins such as pristinamycin Ia (shown) are cyclic depsipeptides. Vgb lyase catalyzes the cleavage of the C—O depsipeptide bond.

Typical modifications include acylation, phosphorylation, and adenylylation. These add steric bulk to the antibiotic, obscure key target binding elements, and can alter charge of the compound. As the reactions require cosubstrates such as ATP and acetylCoA, antibiotic modifying enzymes are generally located within the cell interior where access to these high energy substrates is secure. This strategy impacts several classes of antibiotics and, in some cases, there is significant conservation of general protein fold among enzymes that modify chemically diverse antibiotics.

Acetylation of antibiotics is a common mechanism of modification. The acyltransferases generally use the metabolically abundant acetyl-CoA as the preferred acyl-donor. The amino-glycoside antibiotic acetyltransferases (AACs), for example, modify important amines that make contact with the target 16S rRNA (Fig. 9), which results in a loss of positive charge and increased steric bulk that contribute to lowering the affinity of the antibiotics for the rRNA target by 600-fold (26). The AACs are members of the GCN5 superfamily of acetyltransferases that include such members as the eukaryotic histone acetyltransferases and serotonin acetyltransferase (27).

Strategically similar, but structurally distinct, are the chloramphenicol acetyltransferases (CATs). Chloramphenicol binds to the large ribosomal subunit at the nexus of the peptidyltransfer center and the peptide exit tunnel (28, 29). CATs are trimeric enzymes with active sites at the interface of the monomers that catalyze acetylation of the hydroxyl group on position 3 (31), which is essential for interaction with the ribosome (Fig. 10). The well-characterized type I–III CATs are robust and efficient modifiers of the antibiotic with $k_{cat}/K_m$ values that approach diffusion-controlled reactions (32). These CATs are structurally distinct from a group of chloramphenicol acetyltransferases that have been classified as xenobiotic acetyltransferases (XAT) in view of their broader substrate specificity and relatively inferior affinity for chloramphenicol (31). The XATs belong to the large hexapeptide repeat family of proteins that incorporate a left-handed $\beta$-helix domain (33). Although they show no primary sequence or tertiary structure homology, both CATs and XATs are trimers with active-site residues at the interface of the monomer subunits and both use a catalytically important His residue to assist in deprotonation of the antibiotic hydroxyl group undergoing acetylation. The streptogramin acetyltransferases (VATs) are also members of the XAT family (34) and are responsible for resistance to the type B streptogramins, notably the clinically used drug dalfopristin (Fig. 11) (35).

Phosphorylation is a common mechanism resulting in resistance to the aminoglycoside antibiotics. This chemical strategy also has been associated with resistance to the macrolides such as erythromycin, the tubactinomycins such as viomycin, and chloramphenicol. The aminoglycoside kinases share 3D structural similarity with the Ser/Thr/Tyr protein kinase family (36), and the conservation of kinase signature sequences in macrolide
Bacterial Resistance to Antibiotics

(a) Aminoglycoside antibiotic modifying enzymes. The aminoglycoside antibiotics such as kanamycin B (shown) bind to the 16S rRNA of the bacterial ribosome impairing cognate codon–anticodon discrimination (A). Resistance occurs via acetylation (AAC), phosphorylation (APH), or adenylylation (ANT) of the antibiotic (B). A wide variety of enzymes are known with different regiospecificities of chemical modification, and the sites of some clinically important enzymes are shown in panel C.

and tuberactinomycins kinases suggests that they will also share the family protein kinase 3D structure (37). Mirroring the effect of acetyl transfer, covalent modification of aminoglycosides alters charge and size, resulting in up to a 1300-fold decrease in affinity of the modified antibiotic for 16S rRNA (Fig. 9) [26].

(b) Like the aminoglycosides, the binding site of the macrolide antibiotics with the large ribosomal subunit has also been determined to atomic resolution by X-ray crystallography (29, 38, 39). Key interactions between the antibiotic and the 23S rRNA occur and are mediated through the essential desosamine sugar.
Bacterial Resistance to Antibiotics

Chloramphenicol modifying enzymes. Chloramphenicol binds to the 23S rRNA of the large ribosomal subunit (A). Chemical modification of the essential hydroxyl groups by acetyltransferase or kinase enzymes confers resistance (B).

Chemical modification of Type B streptogramins. These antibiotics bind to the large ribosome adjacent to the chloramphenicol binding site (A). A key interaction with the 23S rRNA is blocked by the action of VAT-dependent acetylation (B).

Macrolide kinases (MPHs) phosphorylate position 2' of the desosamine sugar, impairing binding with the target. Genes encoding these enzymes have been identified in mobile genetic elements in various bacterial pathogens (40-43). Using an analogous evasive strategy, macrolide producing and other soil bacteria encode macrolide glycosyltransferases that catalyze glucosylation of the same essential 2'-position on the desosamine by UDP-glucose (44-46) (Fig. 12). The net result is the same with either approach: steric blockade of the appropriate target-antibiotic interaction resulting in resistance.
Bacterial Resistance to Antibiotics

Figure 12 Macrolide modifying enzymes. Macrolide antibiotics such as erythromycin (shown) bind to the large ribosomal subunit through interactions with the 23S rRNA (A). Chemical modification of the essential desosamine sugar blocks ribosome binding (B).

Chloramphenicol phosphotransferase from the producing bacterium Streptomyces venezuelae (47) is unrelated to the protein kinase family but rather shows more similarity to small-molecule kinases such as shikimate kinase (48). Analogous to the CAT strategy, phosphorylation occurs at the hydroxyl position 3, blocking this essential group from interacting with the ribosome (Fig. 10).

Aminoglycosides and lincosamides such as clarithromycin can be modified by O-adenylylation in an ATP-dependent fashion resulting in resistance. Position 2′ of the aminoglycosides is the target for the clinically important ANT(2′) enzymes that confer resistance to antibiotics such as tobramycin and gentamicin. In analogy to other aminoglycoside modifications that result in resistance discussed above, the addition of the bulky and negatively charged adenosine diphosphate to the antibiotic attenuates ribosome binding (Fig. 9). The lincosamide antibiotics are frequently modified by AMP at position 3 in a reaction catalyzed by the Lin proteins (49, 50). This modification blocks a key interaction of the antibiotic with the 23S rRNA of the bacterial large subunit (Fig. 13) (29).

Antibiotic Modification by Oxidation/Reduction Mechanisms

Antibiotic resistance via redox reactions is not as common as chemical modification; however, there are some examples of this strategy in nonpathogenic bacteria. The streptogramin antibiotic producer Streptomyces virginiae can reduce a vital ketone group of group A streptogramins in an NADH-dependent manner resulting in formation of the vicinal diol, which lacks antibiotic activity (51). Redox-dependent inactivation of rifampin has also been described by a gene product from Rhodococcus equi (52).

Tetracycline resistance occurs primarily by efflux and target protection mechanisms (53); however, inactivation by the redox enzyme TetX has been reported (54). TetX is a monooxygenase that selectively hydroxylates tetracycline antibiotics (55), including the latest generation of antibiotics of the class the glycylcyclines (56). Tetracycline binds to the A-site of bacterial ribosomes interfering with translation. Binding of the invariant β-diketone functional group characteristic of the class to the ribosome occurs through the mediation of a divalent cation (likely Mg^{2+}) (Fig. 14). TetX catalyzes the mono-hydroxylation of the antibiotics at position 11a, effectively disrupting the Mg^{2+} binding site, resulting in resistance (Fig. 14). Furthermore, some tetracyclines undergo nonenzymatic degradation after mono-hydroxylation, resulting in irreversible destruction of the antibiotic.
Antibiotic Resistance Through Altered Transport

To be effective, antibiotics must accumulate to critical levels governed by their affinity for their molecular targets. Furthermore, these targets are frequently intracellular. As a result, antibiotics must efficiently permeate the bacterial cell envelope and achieve sufficiently effective local concentrations. In Gram-positive bacteria, the cell envelope consists of the cell wall and the phospholipid containing cell membrane. In contrast, Gram-negative bacteria deploy an additional outer membrane consisting of an asymmetric LPS/phospholipid bilayer to complement the peptidoglycan and phospholipid cell membrane components of the cell envelope. These architectures provide a substantial barrier to antibiotic access to the cell interior.

Transport-linked resistance to antibiotics can emerge as a result of poor penetration of the bacterial cell envelope or from the energy-dependent purging of antibiotic molecules that have managed to penetrate the envelope through the aegis of efflux proteins. Often, these two mechanisms, entry and efflux, combine to achieve high level resistance to many antibiotics. Like enzyme-based resistance, several of these transport-based resistance genes are encoded on mobile genetic elements; however, all bacteria encode an elaborate collection of transport proteins in their chromosomes, some of which have the potential to be recruited for antibiotic resistance. In particular, Gram-negative bacteria, such as those from the genera Pseudomonas and Burkholderia, are particularly adapted for transport-mediated resistance, and this mechanism contributes significantly to the high level of broad spectrum antibiotic resistance in these organisms.

Antibiotic efflux

Bacterial cells can express a variety of membrane proteins that mediate the energy-dependent efflux of toxic compounds (Table 2). Five structurally distinct classes of efflux proteins are recognized: ABC transporters, resistance-nodulation-division (RND) proteins, multidrug and toxin extrusion (MATE) proteins, small multidrug resistance (SMR) proteins, and major facilitator superfamily proteins (MFS). The latter four of these proteins promote exchange of a cation (H\(^+\) or Na\(^+\)) for the antibiotic, whereas the ABC transporters use ATP hydrolysis to drive ejection of the compound from the cell. All of the protein classes have been associated with antibiotic resistance in various bacterial strains; nonetheless, some general rules have emerged from the past two decades of efflux research. First, with a few exceptions, plasmid- or integron-mediated resistance usually is associated with MFS and SMR proteins that have broad substrate specificity. Second, ABC transporters are analogous to the P-glycoprotein class of drug efflux systems in eukaryotes. They are not as prevalently associated with antibiotic resistance in pathogenic bacteria as the RND class or in Gram-negative and Gram-positive bacteria, whereas the RND class is generally found in Gram-negatives where they are chromosomally encoded. Fourth, sequencing of bacterial genomes has revealed...
that all microbial genomes have multiple predicted antibiotic efflux proteins (e.g., 25 in Staphylococcus aureus COL, 102 in Bacillus anthracis Ames). Several excellent and comprehensive recent reviews cover this field in depth (57–61); therefore, the following description will briefly emphasize the proposed mechanisms rather than the comprehensive details of efflux systems and their individual components.

The MFS proteins are a large superfamily of transporters that are essential to the shuttling of numerous metabolites (sugars, amino acids, and anions) as well as cytotoxic agents such as antibiotics in exchange for cations. They consist of 12-14 membrane-spanning helices and can be found encoded on mobile genetic elements (e.g., QacA, which is clustered in many integrons) or on bacterial chromosomes (e.g., NorA, which is linked to fluoroquinolone resistance in S. aureus). Three-dimensional structures of members of the MFS superfamily have been reported, including the sugar transporters LacY (62) and GlpT (63) and, more recently, the multidrug transporter EmrD from E. coli (64). These structures reveal a collection of helices that form a pore-like structure to enable transport across the membrane. In EmrD, the interior of the pore is mostly hydrophobic, consistent with the substrate specificity of the protein that includes detergents such as sodium dodecyl sulfate. This structure suggests a “rocker-switch” mechanism where the substrate enters the open pore from the cytoplasm or the inner leaflet of the cell membrane, the open pore then undergoes a conformational change to close around the substrate coupled with H⁺ transport, and finally the pore is reopened facing the exterior of the cell where the substrate can diffuse away (Fig. 15).

The RND proteins have emerged as important mediators of multidrug resistance in Gram-negative bacteria, especially in the opportunistic pathogens P. aeruginosa, Stenotrophomonas maltophilia, Acinetobacter baumannii, and Burkholderia cepacia. These resistance modules combine with reduced uptake and a cadre of chromosomally encoded resistance enzymes to make these bacteria virtually impermeable to antibiotics. The groups of Nikaido and Poole have studied the RND efflux systems in detail from E. coli and P. aeruginosa, respectively, providing great insight into the components of individual RND transports, their mechanisms of action, and substrate specificity. Consequently, the AcrAB/TolC system in E. coli and the MexAB/OprM triad in P. aeruginosa are the paradigms of the class. They are tripartite efflux systems consisting of an inner cell membrane transporter (AcrB, MexB), a periplasmic membrane fusion protein (AcrA, MexA), and an outer membrane pore (TolC, OprM) to seamlessly connect the inside of the cell across two membrane barriers to the external environment (Fig. 16). This structure provides an unimpeded conduit for toxic molecules from the cell interior to the exterior. Additionally, antibiotics such as β-lactams and other molecules are able to access the efflux system within the periplasmic space without entering the cell, thereby expanding the spectrum of the efflux system. The crystal structures of AcrB (65, 66), MexA (67), a fragment of AcrA (68), and TolC (69) have been determined. These structures confirm the predicted membrane spanning arrangement of the efflux system and serve to rationalize the broad substrate specificity of the complex as well as illustrate the challenge of overcoming efflux-mediated resistance.

Figure 14  Enzymatic modification of tetracycline. Tetracyclines bind to the ribosome in several areas, but the primary region is in the A site of the small subunit (shown) (A). The antibiotic binds divalent cations such as Mg²⁺, which is essential for ribosome binding. The enzyme TetX catalyzes the mono-hydroxylation of tetracycline at position 11a disrupting the Mg²⁺ binding site resulting in resistance.

Figure 15  Three-dimensional structure of EmrD from E. coli.

Figure 16  Structure of the AcrAB/TolC system in E. coli.
## Table 2 Selected antibiotic resistance efflux systems

<table>
<thead>
<tr>
<th>Protein</th>
<th>Class</th>
<th>Found in...</th>
<th>Organism</th>
<th>Substrates*</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuacA</td>
<td>MFS</td>
<td>✓</td>
<td>many</td>
<td>Quaternary ammonium compounds, dyes</td>
</tr>
<tr>
<td>NorA</td>
<td>MFS</td>
<td>✓</td>
<td><em>Staphylococcus aureus</em></td>
<td>Dyes, fluoroquinolones</td>
</tr>
<tr>
<td>Bmr</td>
<td>MFS</td>
<td>✓</td>
<td><em>Bacillus subtilis</em></td>
<td>Dyes, chloramphenicol, fluoroquinolones</td>
</tr>
<tr>
<td>Smr (QuacC,D)</td>
<td>SMR</td>
<td>✓</td>
<td>many</td>
<td>Monovalent cations, fluoroquinolones</td>
</tr>
<tr>
<td>EmrE</td>
<td>SMR</td>
<td>✓</td>
<td><em>Escherichia coli</em></td>
<td>Monovalent cations, tetracycline</td>
</tr>
<tr>
<td>MefE</td>
<td>MSF</td>
<td>✓</td>
<td><em>Streptococcus pneumoniae</em></td>
<td>Macrolides</td>
</tr>
<tr>
<td>TetA</td>
<td>MFS</td>
<td>✓</td>
<td><em>Enterococcus faecalis</em></td>
<td>Tetracycline</td>
</tr>
<tr>
<td>Lsa</td>
<td>ABC</td>
<td>✓</td>
<td></td>
<td>Type A streptogramins, lincosamides</td>
</tr>
<tr>
<td>AcrA/AcrB/TolC</td>
<td>RND</td>
<td>✓</td>
<td><em>Escherichia coli</em></td>
<td>Dyes, bile salts, chloramphenicol, β-lactams</td>
</tr>
<tr>
<td>MexA/MexB/OprM</td>
<td>RND</td>
<td>✓</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>Dyes, chloramphenicol, β-lactams, aminoglycosides, novobiocin</td>
</tr>
<tr>
<td>MexC/MexD/OprJ</td>
<td>RND</td>
<td>✓</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>Chloramphenicol, fluoroquinolones, tetracycline, trimethoprim, triclosan</td>
</tr>
<tr>
<td>AdeA/AdeB/AdeC</td>
<td>RND</td>
<td>✓</td>
<td><em>Acinetobacter baumannii</em></td>
<td>Aminoglycosides, chloramphenicol, fluoroquinolones, tetracycline, trimethoprim</td>
</tr>
</tbody>
</table>

*Partial list.

*Includes plasmids, transposons, integrons.

---

Figure 15 Proposed rocker-switch mechanism for MFS-mediated drug efflux. Antibiotic (circle) enters the open pore from the cytosolic face or inner membrane leaflet and is antiported to the cell exterior with an H⁺ ion coupled with conformational change within the protein.

**Reduced uptake of antibiotics**

Although efflux is a daunting obstacle for many antibiotics, it is the balance between influx and efflux that determines whether cells will be sensitive or resistant in the face of an antibiotic challenge. As a result, reduced uptake of antibiotics coupled with activation of efflux pumps is a potent combination that can overcome sensitivity to drugs in many pathogens. Many metabolites and compounds such as antibiotics penetrate cells via transport proteins and, as a result, mutation in these proteins or decrease in their expression can decrease antibiotic uptake. In Gram-negative bacteria, the outer membrane porins are required for entry of many antibiotics such as β-lactams and aminoglycosides, and antibiotic resistance can occur through modulation of these proteins. For example, imipenem resistance...
Bacterial Resistance to Antibiotics

In P. aeruginosa, resistance to antibiotics is frequently associated with mutation in the OprD porin (70).

Other Mechanisms of Antibiotic Resistance

In addition to enzymatic modification or destruction and transport, bacteria can evade antibiotics through a variety of ways, including metabolic bypass, target overexpression or modification, and the production of various protective proteins.

Metabolic bypass: vancomycin resistance

The classic example of metabolic bypass resulting in antibiotic resistance can be found in the glycopeptide antibiotics. Glycopeptides such as vancomycin and teicoplanin are natural product antibiotics that bind to the ubiquitous acyl-D-Ala-D-Ala dipeptide terminus of peptidoglycan (71), which results in sequestration of the substrate for cell wall biosynthetic enzymes including the transpeptidases and glycosyltransferases, resulting in impairment of cell wall synthesis and bacterial growth. Therefore, vancomycin binds to an essential small-molecule metabolite, which is the enzyme substrate, rather than to the protein itself. As a result, simple mutation of a single gene to overcome resistance is not possible.

Vancomycin and other glycopeptides bind to acyl-D-Ala-D-Ala through a series of five hydrogen bonds that includes a key interaction with the amide hydrogen of the D-Ala-D-Ala peptide bond (Fig. 17). The acyl-D-Ala-D-Ala glycopeptide-binding unit is synthesized in stepwise fashion through a series of metabolic steps involving first synthesis of D-Ala, ATP-dependent ligation of two D-Ala molecules to generate the D-Ala-D-Ala dipeptide, and a second ligation where the dipeptide is grafted onto the N-acetyl-muramic acid tripeptide acceptor for incorporation into the growing cell wall (Fig. 18). Vancomycin-resistant bacteria overcome the antibiotic by bypassing the normal cell wall biosynthesis. Instead of producing a cell wall terminating in the D-Ala-D-Ala dipeptide, a peptidoglycan is produced that terminates with the depsipeptide (ester) D-Ala-D-lactate (Fig. 17). This change of NH for O results in loss of a critical H-bond and electronic repulsion between the ester O and the carbonyl of vancomycin. The net effect is a 1000-fold decrease in $K_D$, resulting in vancomycin resistance (72). Incorporation of the D-Ala-D-lactate ester requires three enzymes: VanH, VanA, and VanX. VanH is a D-lactate dehydrogenase generating D-lactate from pyruvate. VanA is an ATP-dependent depsipeptide ligase that generates D-Ala-D-lactate. Finally, VanX is a highly specific D-Ala-D-Ala dipeptidase that clears the cell of D-Ala-D-Ala that continues to be synthesized by the chromosomally encoded D-Ala-D-Ala ligases. These three proteins work in concert to bypass the normal cell wall biosynthetic metabolism and are each essential for resistance. Not surprisingly, they are colocated on resistance islands (73).
Bacterial Resistance to Antibiotics

**Figure 18** Synthesis of the D-Ala-D-Ala peptidoglycan terminus. D-Ala-D-Ala ligases (Ddl) synthesize the dipeptide, which is incorporated into the UDP-MurNAc-tripeptide by the D-Ala-D-Ala adding enzyme (MurF in *E. coli*).

plasmids and transposons that have emerged in strains of enterococci and more recently in *S. aureus*.

In a variant of this strategy, vancomycin resistance can also emerge by incorporation of peptidoglycan terminating in D-Ala-D-Ser. Substitution of D-Ala for D-Ser has steric implications for tight vancomycin-peptidoglycan binding, which also can result in resistance.

**Target modification**

One of the most straightforward mechanisms of antibiotic resistance is mutation of the target to a form that has less affinity for the antibiotic. Spontaneous mutations that provide some benefit to the organism occur roughly $2 \times 10^{-9}$ per replication (reviewed in Reference 73). Exposure to certain classes of antibiotics such as the rifamycins and fluoroquinolones can induce mutation, increasing the opportunity of developing resistance (74). Point mutations are associated with resistance to virtually all antibiotics, but this form of resistance is particularly important in the clinic for the fluoroquinolones, rifamycins such as rifampin, trimethoprim, and the sulfonamides.

Enzyme-mediated target modification can also be an important mechanism of antibiotic resistance. The Erm proteins are 23S rRNA S-adenosylmethionine-dependent methyltransferases that immunize the bacterial ribosome against the macrolide antibiotics such as erythromycin, the lincosamides, and type B streptogramins (75). The Erm proteins, specifically methylate A2058, provide a steric block for all three classes of antibiotics: a remarkable example of potent and strategic resistance parsimony. The *erm* genes are widespread in clinical isolates and can be constitutively expressed or strategically deployed through a complex induction mechanism (76).

By means of similar logic, methylation of the 16S rRNA at position G1405 by Arm and related proteins results in high level aminoglycoside resistance (77).

**Protective proteins**

Expression of various proteins, which can collectively be called protective, can result in antibiotic resistance. These immunity proteins often work in stoichiometric fashion to sequester the compounds, or otherwise protect the cell against the toxic activity of the antibiotics.

Resistance to the peptide antibiotic/anticancer agent bleomycin can occur through the expression of the Blm family of proteins that sequester the antibiotic. Bleomycin includes...
Bacterial Resistance to Antibiotics

Resistance to antibiotics is manifested through a number of different chemical and biophysical mechanisms, each with its own associated biochemical logic. A fundamental component of this logic is the sensitivity of the antibiotic warhead to chemical alteration. Although resistance is inevitable, understanding how resistance can occur, what the sensitive regions of the molecule are, and biochemical strategies that could overcome it such as the identification of specific inhibitors of resistance are essential to delay the emergence of resistance. Recently, it has been shown that penetration of antibiotic resistance in the microbial community at large is much deeper than previously appreciated (16). Understanding the origins and evolution of resistance will also provide drug discoverers with valuable information to circumvent established mechanisms. The concept of overcoming the inevitability of antibiotic resistance with small-molecule inhibitors of resistance mechanisms has been proven to be highly effective in the β-lactam class of antibiotics (90), and this approach should be broadly applicable (91). In fact, this area is one of intense interest, especially as applied to efflux mechanisms (92, 93). The routes forward in this important area of medicine require Chemical Biology approaches to understand antibiotic-target interactions and antibiotic resistance mechanisms. Chemical Biology and other multidisciplinary disciplines are therefore essential to prolong the utility of these essential therapeutic agents.

References


Conclusions

The history of antibiotic use over the past 70 years is invariant: initial deployment to the clinic with littleto background resistance followed by emergence and global spread of resistance, a metal-binding region that complexes Fe\(^{2+}\) resulting in a potent generator of oxygen radicals (warhead) and a DNA intercalating portion (scaffold) that targets the complex to DNA. Bleomycin, therefore, is a remarkable molecule with the ability to produce toxic radicals in an unregulated fashion proximal to DNA. The Blm proteins, such as BlmA from the bleomycin producer Streptomyces verticillus and BlmT found on various resistance transposons such as Tn14,000, are small dimeric proteins (~14,000 Da/molecular), and high resolution crystal structures are available for each protein (78, 79). The structure of the BlmA protein in complex with a Cu(II)-bleomycin complex reveals that each monomer binds one antibiotic molecule and that binding of the second molecule of bleomycin is cooperative (80). The precise mechanism of resistance is likely through sequestration of both the warhead and scaffold portions of the antibiotic, thereby preventing optimal activation of the Fe(II) complex.

Similarly, resistance to the anticancer antibiotic mitomycin C in the producing organism Streptomyces lavendulae, occurs via sequestration of the antibiotic by the binding protein MBD (81, 82). This protein interacts with the mitomycin export system, thus additionally facilitating export of the antibiotic from the cell. The structure of MBD reveals similarity to the bleomycin resistance proteins and, in fact, the protein can also bind bleomycin and confer resistance to this antibiotic (83).

By similar means, resistance to the highly toxic enediyne antibiotics, which are active at ng/mL concentrations, occurs through binding proteins in the producing bacteria. These antibiotics include a conserved core warhead consisting of two acetylenic groups bridged by a carbon–carbon double bond. This warhead generates a diradical after thiol activation resulting in DNA cleavage (84). The class is divided into two structural groups, the 9-membered and 10-membered enediyynes, referring to the size of the warhead-containing ring. The 9-membered enediyynes are stabilized by cooperation and titration by core protective binding proteins. The X-ray crystal structure of one of these complexes, C-1027 with the binding protein CagA, has been reported (85). Resistance to the more chemically stable 10-membered enediyynes (e.g., calicheamicin), however, involves the remarkable expression of the protective protein CalC that undergoes calicheamicin-dependent proteolytic self-sacrifice simultaneously inactivating the antibiotic (86).

Another example of protective protein expression is the tetrapeptide antibiotic, which is active at ng/mL concentrations, occurs through binding proteins in the producing bacteria. The antibiotics include a conserved core warhead consisting of two acetylenic groups bridged by a carbon–carbon double bond. This warhead generates a diradical after thiol activation resulting in DNA cleavage (84). The class is divided into two structural groups, the 9-membered and 10-membered enediyynes, referring to the size of the warhead-containing ring. The 9-membered enediyynes are stabilized by cooperation and titration by core protective binding proteins. The X-ray crystal structure of one of these complexes, C-1027 with the binding protein CagA, has been reported (85). Resistance to the more chemically stable 10-membered enediyynes (e.g., calicheamicin), however, involves the remarkable expression of the protective protein CalC that undergoes calicheamicin-dependent proteolytic self-sacrifice simultaneously inactivating the antibiotic (86).

A rather example of protective protein expression is the tetra-cyclic resistance TetO and TetM proteins and their analogues (87). These proteins structurally resemble the prokaryotic GT-Pase elongation factors such as EF-G and EF-Tu that bind to the ribosome and are essential for translation. TetO/M protection proteins do not replace elongation factors, but they do bind to the ribosome and displace bound tetracyclines in a GTP-dependent manner (88, 89).

References


62. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S.
60. Piddock LJ. Clinically relevant chromosomally encoded mul-
59. Piddock LJ. Multidrug-resistance efflux pumps—not just for re-
57. Kumar A, Schweizer HP. Bacterial resistance to antibiotics: active
54. Speer BS, Salyers AA. Novel aerobic tetracycline resistance gene
53. Roberts MC. Update on acquired tetracycline resistance genes.
52. Yu EW, Aires Jr, McDermott G, Nikaido H. A periplasmic
51. Izard T, Ellis J. The crystal structures of chloramphenicol phos-
48. Izard T, Ellis J. The crystal structures of chloramphenicol phos-
47. Bugg TOH, Wright GD, Dutka-Malen S, Arthur M, Courvalin P.
46. Yu EW, McDermott G, Zgurskaya HI, Nikaido H, Koshland DE
45. Yu EW, Aires Jr, McDermott G, Nikaido H. A periplasmic
44(3):11829–11835.
43. Wright GD. TetX is a flavin-dependent monooxygenase conferring
42. Tanaka T, Fukuda-Ishisaka S, Hirama M, Otani T. Solution struc-
40. Sheldon PJ, Johnson DA, August PR, Liu HW, Sherman DH.
39. Martin TW, Dauter Z, Devedjiev Y, Sheffield P, Derewenda ZS, Derewenda U. Molecular basis for vancomycin resistance in
38. Maruyama M, Kumagai T, Matoba Y, Hayashida M, Fuji T, Hata Y,
29. Tanaka T, Fukuda-Ishisaka S, Hirama M, Otani T. Solution struc-
27. Wright GD. TetX is a flavin-dependent monooxygenase conferring
26. Tanaka T, Fukuda-Ishisaka S, Hirama M, Otani T. Solution struc-
24. Martin TW, Daedler Z, Devedjiev Y, Sheffield P, Jelen F, He M, Sherman DH, Otlewski J. Derewenda ZS, Derewenda U. Molecular basis for vancomycin resistance in
18. Wright GD. TetX is a flavin-dependent monooxygenase conferring


Further Reading


See Also

Antibiotics, Mechanisms of
Antibiotics, Biosynthesis of
Enzyme Catalysis, Chemistry of
Secondary Metabolite, Chemical Diversity of
Calcium Signaling: Encoding and Decoding

Marisa Brini and Laura Fedrizzi, Department of Biochemistry and Department of Experimental Veterinary Sciences, University of Padova, Padova, Italy
Ernesto Carafoli, Venetian Institute of Molecular Medicine (VIMM), Padova, Italy
doi: 10.1002/9780470048672.wecb048

Calcium is a universal second messenger that controls many cellular reactions. Considering its pleiotropic role, it seems evident that the cells, to get a specific message in the proper time and manner, need precise and efficient mechanisms to encode and decode Ca\(^{2+}\) signals. Generally, extracellular stimuli are converted in a transient increase in cytosolic Ca\(^{2+}\) concentration, \([\text{Ca}^{2+}]_c\), which, in turn, modulates cell function. In the last two decades, improvements in the development of probes and instrumentation for Ca\(^{2+}\) imaging have led to the discovery that the coordinated action of different players is responsible for a complex spatio-temporal organization of the Ca\(^{2+}\) signal. It is intriguing to observe that cells can encode and discriminate Ca\(^{2+}\) signals not only according to their magnitude but also according to their localization (microdomains) and shape; i.e., cells can discriminate between sustained and oscillatory signals. Even more, in the case of oscillations, messages can be read differently according to the frequency of the oscillatory signals. The mechanisms by which cells decode Ca\(^{2+}\) signals are now explored in numerous laboratories. This article focuses on the autoregulation properties of the Ca\(^{2+}\) signals. It will show that Ca\(^{2+}\) itself is central in the regulation of the Ca\(^{2+}\) signal. It will also show that it can act as a first and second messenger and that it can modulate the activity and the availability of the other players in the signaling operation.

Eukaryotic cells are surrounded by media containing free Ca\(^{2+}\) concentrations that exceed 1 mM, but they manage to maintain a free Ca\(^{2+}\) concentration in the cytoplasm that is four orders of magnitude lower. The very low internal concentration is maintained by the active transport of Ca\(^{2+}\) ions against their concentration gradient by Ca\(^{2+}\) pumps in the plasma membrane, in the endoplasmic/rough endoplasmic reticulum (ER/SR), and in the Golgi membranes. The plasma/membrane Na\(^+/\)Ca\(^{2+}\) exchangers also play a role, particularly in heart and skeletal muscle. The cells invest energy in this process that not only preserves the low \([\text{Ca}^{2+}]_c\), but also generates an intracellular source of Ca\(^{2+}\) within the lumen of intracellular organelles, essentially, the ER/SR and the Golgi apparatus. Generally, extracellular stimuli are converted in a transient increase in the \([\text{Ca}^{2+}]_c\), which, in turn, activates cellular functions. The sources of Ca\(^{2+}\) are both extracellular and intracellular; i.e., Ca\(^{2+}\) channels in the plasma membrane and in the intracellular membranes are critical in the control of cellular Ca\(^{2+}\) homeostasis. To guarantee the specificity of the signal transmission, the cell organizes dynamically the Ca\(^{2+}\) fluctuations in the cytosol by varying the distribution, the type, and the availability of the different Ca\(^{2+}\) transporters, and it increases the spatial and temporal complexity of Ca\(^{2+}\) homeostasis by compartmentalizing the signals into the organelles. The ER was originally considered to be the sole dynamic Ca\(^{2+}\) regulator in the cell, but it has now become clear that the nucleus, the mitochondria, the Golgi apparatus, the endosomes/lysosomes, and the secretory vesicles also play fundamental roles (Fig. 1). Recent methodological developments have revealed that, in living cells, these systems are strictly interconnected. The control of their Ca\(^{2+}\) homeostasis is not only essential in the control of organelle-specific processes, but it is also fundamental in the overall dynamic modulation of the Ca\(^{2+}\) signaling in the cytosol. Many protein components of this signaling cascade have been cloned and characterized: Strikingly, many exist as different isoforms, the number of which is further increased through mechanisms of alternative splicing.
Ca2+ channels (VOCCs) (1), or on ligand binding [in the case of receptor-operated Ca2+ channels (ROCCs) (2)], or on a still poorly understood mechanism linked to the emptying of intracellular Ca2+ stores. VOCCs are found not only in excitable cells such as neurons, skeletal muscle, and heart, but also in nonexcitable cells. Six classes (termed L, N, P, T, R, and Q) have been identified based on physiologic and pharmacologic properties. Structurally, all VOCCs are complexes of five subunits, termed α1, α2, β, γ, and δ, that assemble into large macromolecular complexes and that are encoded by different genes. α1, the largest subunit, contains the conduction pore, the voltage sensor, the gating apparatus, and the sites of channel regulation, e.g., by protein kinases, by drugs, and by toxins. Much of the diversity of Ca2+ channel types originates from the variety of α1 subunits, and at least 10 different genes encoding for different α1 subunits have so far been identified in mammals. ROCCs are activated by the interaction of ligands with plasma membrane receptors. L-glutamate is the most abundant excitatory transmitter in the vertebrate central nervous system. It activates two general classes of receptors, the “ionotropic” receptors, which are ionic channels, and the “metabotropic” receptors, which are coupled to G-proteins, which activate phospholipase C (PLC) and promote intracellular Ca2+ responses. Three forms of ionotropic receptors have been characterized and named after their most widely used agonist: the kainate (KA), the α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA), and the N-methyl D-aspartate (NMDA) receptors. AMPA and NMDA receptors colocalize at the same postsynaptic membrane, and their close functional interdependence has important roles in the processes of memory storage and learning. A heterogeneous group of channels (most of which belong to the so-called transient receptor potential family, TRPs) are activated by a variety of factors, including mechanical stretch, osmolarity, temperature, second messengers, G-proteins, protein–protein interactions, and so on (3). The molecular nature of an additional group of channels, in this case activated by arachidonic acid, is still undetermined (4). Whatever the nature and gating properties of a channel, it is easily envisaged that upon its opening, a high concentration of Ca2+ will be generated at its mouth, on the inner side of the plasma membrane. In general terms, the peak amplitude and duration of Ca2+ microdomains formed at the internal mouth of a Ca2+ channel and its immediate neighborhood will depend on the following: the conductance of the channel itself, its Ca2+ selectivity, the concentration of Ca2+ in the extracellular medium, the membrane potential, and the nature and amount of the intracellular Ca2+ buffers.

Intracellular Ca2+ Release Mechanisms

The release of Ca2+ from internal stores, usually the ER or its muscle equivalent, the SR, is controlled by Ca2+ itself, or by an expanding group of messengers, such as inositol 1,4,5-trisphosphate (InsP3), cyclic ADP ribose (cADPR), nicotinic acid adenine dinucleotide phosphate (NAADP), and sphingosine-1-phosphate (S1P), that either stimulate or modulate the release channels on the internal stores. The Ca2+ release channels represent the molecular component of the Ca2+ handling machinery of the intracellular stores. They belong to two families, the so-called InsP3 receptors (InsP3Rs) and Ryanodine receptors (RyRs) (5). The heterogeneous distribution of these channels within intracellular compartments is one of the key factors in determining the spatial heterogeneity of the Ca2+ signals and thus the formation of Ca2+ microdomains. The two channel types have been conserved during evolution, as a significant degree of homology characterizes the sequences of the domains next to the C-terminus that span the membrane and contribute to the assembly of the channel proper. Both are...
In cardiac and skeletal muscle, a highly structured morphologic architecture allows the generation of Ca$^{2+}$ microdomains at the surface of the SR; these microdomains are a key component of the trigger for firing the Ca$^{2+}$ signals necessary for cell activation according to the mechanism known as Ca$^{2+}$-induced-Ca$^{2+}$-release (CICR) (6). Indeed, in both cell types, the physiologic stimulatory signal leading to contraction is conveyed by action potentials: A plasma membrane depolarization travels via the opening of voltage-dependent Na$^+$ channels and reaches the cell interior through invaginations (the T tubules) in which VOCCs are located; this process causes the influx of Ca$^{2+}$ that is per se insufficient to induce the physiologic response (the sliding of the acto-myosinic contractile apparatus), but it represents the trigger for the release of Ca$^{2+}$ (particularly in the heart) from the large intracellular Ca$^{2+}$ reservoir (the SR), through the RyRs (7). In skeletal muscle, direct coupling between the two molecules (VOCCs and RyR) is believed to cause the opening of RyRs; thus, the influx of Ca$^{2+}$ through VOCCs plays a facilitatory, but not a necessary, role.

The interaction between the two channels is thought to be the necessary event to cause activation of the RyR isoform of skeletal muscle (RyR1), and then Ca$^{2+}$ release through the RyR1 is amplified by CICR. In heart, no direct physical interaction occurs between the two types of channels, and thus a high [Ca$^{2+}$] in the proximity of the RyR2 (the cardiac RyR isomorph) represents the essential activatory signal (8). Therefore, in heart, Ca$^{2+}$ influx through VOCCs is necessary to trigger SR Ca$^{2+}$ release through the RyR2 via the process of CICR.

In this context, the Ca$^{2+}$ sensitivity of the RyRs, as well as the evaluation of the local Ca$^{2+}$ concentration to which they are exposed, is a key factor in determining the excitability of the muscle fiber and the efficiency and duration of contraction. As a result, a great deal of work has concentrated on identifying the “fundamental” Ca$^{2+}$ signaling. Rapid Ca$^{2+}$ imaging of both cardiac and skeletal muscle has revealed the occurrence of transient, local increases in Ca$^{2+}$ concentration, denominated Ca$^{2+}$ puffs (9) that were attributed to the opening of single RyR channels, or more likely, a cluster of RyRs (10, 11).

The occurrence of elementary signals, generated by the opening of a spatially restricted group of channels (12) and denominated “Ca$^{2+}$ puffs” (13), is also shared by the InsP$_3$-dependent signaling system. Ca$^{2+}$ puffs denote the “Ca$^{2+}$ excitability” of the cell, but the induction of a physiologic response requires the coalescence of these elementary events into a larger rise, which may be limited to a portion of a polarized cell or may diffuse to the whole cell body in a truly global Ca$^{2+}$ signal. The transition from elementary to global events, and its regulatory mechanisms, is thus critical in determining the final cellular outcome of the InsP$_3$-dependent Ca$^{2+}$ signal. Numerous stimuli function through the activation of PLC to generate InsP$_3$. Several isozymes of PLC have been described: all of them specific for phosphatidylinositol (PhIns, PI), a phospholipid that is predominantly present in the inner leaflet of the plasma membrane, and its mono (PhIns-4-P, PIP$_1$) and bis-phosphorylated forms (PhIns-4,5-P$_2$, PIP$_2$). Diacylglycerol (DAG) and phosphorylated inositolis are formed, with 1,4,5 InsP$_3$ being the only one among the numerous inositol phosphate isomers that is capable of releasing Ca$^{2+}$ from intracellular stores by interacting with a specific receptor. Three genes encode the InsP$_3$Rs. They are ubiquitously expressed; type 1 is expressed in particularly high levels in Purkinje neurons in the cerebellum. The differential contribution of InsP$_3$R isoforms to Ca$^{2+}$ signaling is now being explored. Studies on the DT40 chicken lymphoma cell line, in which one or more of the InsP$_3$R genes were systematically eliminated by homologous recombination, have shown that the expression of isoform 2 is necessary for the generation of Ca$^{2+}$ oscillations after cell stimulation (14).

The InsP$_3$R is operated by InsP$_3$, but it is also modulated by Ca$^{2+}$. Low luminal Ca$^{2+}$ has been proposed to stimulate the opening of the Ca$^{2+}$-channel, but the matter is still being debated. The effects of cytosolic Ca$^{2+}$ on the receptor are better documented. Even if they may vary, depending on concentration, cell type, or experimental conditions, the effects of Ca$^{2+}$ are biphasic. They are stimulatory at <300 nM and inhibitory at >300 nM (15, 16).

**Intracellular Ca$^{2+}$-Removing Mechanisms**

During the generation of a Ca$^{2+}$ transient, the “on” reactions are counteracted by “off” reactions, during which time various pumps and exchangers remove Ca$^{2+}$ from the cytoplasm. These mechanisms are essential in maintaining the resting level of Ca$^{2+}$ at approximately 100 nM and in ensuring the Ca$^{2+}$ loading of internal stores. As mentioned, three different molecular components are involved in the “off” reactions: the plasma membrane Ca$^{2+}$-ATPase (PMCA), the Na$^+$/Ca$^{2+}$ exchanger (NCX), and the sarcoplasmic reticulum Ca$^{2+}$-ATPase (SERCA). PMCAs represent the main Ca$^{2+}$-extrusion system of all eukaryotic cells and have high Ca$^{2+}$ affinity. In mammals, four separate genes encode distinct PMCA isoforms. PMCAs and PM CA are ubiquitously expressed, whereas PMCA 2 and PMCA 3 are expressed almost exclusively in the central nervous system. Complex patterns of alternative RNA splicing generate additional isoform variability. Of interest are variants in which the alternative splicing occurs at site C, located in the C-terminal tail of the protein, which involves the region containing the calmodulin-binding domain (see below). The calmodulin (CaM)-binding domain, in the absence of CaM, interacts with two sites next to the active site of the PMCA maintaining the pump in an inhibited state. CaM removes its binding domain from its intramolecular regulatory receptors, freeing the pump from autoinhibition. C site splicing occurs in all isoforms and, in general, causes the inclusion of one (or two) additional exons, inducing a premature truncation of the protein, which now has a shorter regulatory domain (termed C1 or a variant) that differs significantly from that of the variant termed C1 or b (17).

The other plasma membrane Ca$^{2+}$-exporting system, the NCX, is particularly active in excitable cells, e.g., heart, which periodically experiences the need to rapidly expel large amounts of calcium. It has low Ca$^{2+}$ affinity but high Ca$^{2+}$-transporting capacity. Three basic exchanger gene products are known, NCX1, NCX2, and NCX3. The first two products are ubiquitously distributed in tissues, whereas NCX3 is restricted to the brain (18). The SERCA pumps are encoded by three genes and are differently distributed in the animal tissues. SERCA1...
Calcium Signaling: Encoding and Decoding

is largely expressed in fast-twitch skeletal muscle. SERCA2 has two splicing variants: isoform 2b, which is ubiquitously distributed, and isoform 2a, which is dominantly expressed in cardiac muscle. SERCA3 is instead expressed in nonmuscle cells. Recently, the tertiary structure of the SERCA1a pump has been solved, and it has confirmed the predicted architecture containing 10 transmembrane domains (39).

The diverse PMCA, SERCA, and NCX isoforms are expressed in a tissue- and development-specific manner, and their apparent redundancy probably enables the cells to select the combination of reactions that exactly meet their specific Ca$^{2+}$-signaling requirements.

Functional Significance of the Different Modes of Ca$^{2+}$ Signal Transmission

Different cell types use distinct Ca$^{2+}$ signals, as suits best their physiology. In particular, the possibility of local and global Ca$^{2+}$ signaling permits the control of separate processes in the same cell.

Emerging evidence reveals that cells use local Ca$^{2+}$ signals, i.e., spatially confined high Ca$^{2+}$ concentration microdomains, to obtain specific and rapid effects such as the release of the contents of synaptic or secretory vesicles, the activation of ion channels, mitochondrial energy metabolism, and the generation of nuclear specific Ca$^{2+}$ oscillations (20, 21). Confined Ca$^{2+}$ signals at the internal mouth of the channels may thus remain localized and transmit the signal to targets in the immediate vicinity or trigger a chain of autocatalytic Ca$^{2+}$-releasing events that results in the generation and spreading of Ca$^{2+}$ waves across the cell. Ca$^{2+}$ waves have been classified based on the speed of their motion (22). Global Ca$^{2+}$ signals, such as waves and oscillations, are preferentially used when the targets to reach are distributed throughout the cell. Many processes require prolonged stimulation to be activated, such as fertilization, axonal growth of cortical neurons, neuronal cell migration, exocytosis, and gene transcription. Ca$^{2+}$ waves can also spread from one cell to the next, thus coordinating the activity of groups of cells within a tissue. For example, in perfused intact liver, hormones linked to the formation of InsP$_3$ evoke Ca$^{2+}$ transients with an oscillatory pattern, the frequency of which depends on the concentration of the agonist. The spikes spread as wave through the cytoplasm, the nucleus, and also through gap junctions to the entire liver lobule to coordinate the metabolic liver function (23).

Information is also encoded within the frequency of Ca$^{2+}$ oscillations that occur in the cytosol. Oscillations can derive either from fluctuations of the entry of external calcium or of the release from intracellular stores. The former occur primarily in excitable cells, after the periodic opening of plasma membrane Ca$^{2+}$ channels, such as induced, for example, by the rhythmic changes of the plasma membrane potential of the heart or by bursts of action potential in neurons. In nonexcitable cells, the predominant mechanism of [Ca$^{2+}$], elevation is from the activation of plasma membrane receptors coupled to G-proteins and to the phospho-inositol pathway. In these cells, the oscillatory pattern of Ca$^{2+}$ increases is not dependent on the periodic opening and closing of plasma membrane Ca$^{2+}$ channels, but on cycles of Ca$^{2+}$ release and uptake from the intracellular compartment sensitive to InsP$_3$. Recent works suggest at least two possible mechanisms for the generation of oscillatory Ca$^{2+}$ signals: either an oscillatory production of InsP$_3$, or an oscillatory inactivation of InsP$_3$Rs. The two mechanisms appear to operate differentially according to the cell type. The fluctuations in InsP$_3$ may be controlled by Ca$^{2+}$ itself through the regulation of membrane PLC that generates InsP$_3$, from PLC$_{eta}$, or through regulatory proteins, which act directly on G-proteins, thus affecting the downstream InsP$_3$ production. Figure 2 summarizes the different shapes of Ca$^{2+}$ signals and some of their biologic targets.

How are the Ca$^{2+}$ Messages Decoded?

The ability to use Ca$^{2+}$ in different modes helps cells to achieve a multitude of signals varying in amplitude, frequency, kinetics, and localization, as well as to avoid the deleterious effects associated with sustained Ca$^{2+}$ increases. Cells avoid death either by using low amplitude Ca$^{2+}$ signals or, more usually, by delivering the signals as brief transients. It is intriguing to understand how the encoded message could regulate such a large number of different processes. The explanation probably lies in the fact that different "forgers" personalize the message.

After cell stimulation, Ca$^{2+}$ flows into the cells and interacts with different Ca$^{2+}$-binding proteins that function either as Ca$^{2+}$ effectors (i.e., decoding the message) or Ca$^{2+}$ buffers. Proteins able to bind Ca$^{2+}$ with the affinity and specificity required for the regulation of its concentration in the intracellular environment generally belong to the so-called EF-hand family (24). The family contains hundreds of members, of which the most intensively studied is the ubiquitous CaM. CaM (and a number of homologous proteins) consists of two independently folded domains connected by a flexible long helical segment. Each domain contains one pair of EF hands. Upon Ca$^{2+}$ binding, structural rearrangements occur and CaM collapses into a hairpin structure around the binding domains of target proteins. Clearly, the decoding of the Ca$^{2+}$ signal by EF-hand proteins is a sophisticated operation, which is not restricted to the relatively simple process of reversibly binding Ca$^{2+}$ for buffering purposes. In fact, the buffers function to fine-tune the spatial and temporal properties of Ca$^{2+}$ signals. They can alter both the amplitude and the recovery time of individual Ca$^{2+}$ transients. These buffers have different properties and expression patterns. For example, calbindin-D28 (CB) and calrectin (CR) are fast buffers, whereas parvalbumin (PV) has much lower binding kinetics and high affinity for Ca$^{2+}$. The physiologic roles of CB, CR, and PV have been particularly studied in neurons. These studies have been facilitated by the recent generation of mouse strains deficient in these proteins. CR is principally expressed in cerebellar granule cells and their parallel fibres, whereas PV and CB are present throughout the asan, soma, dendrites, and spines
Calcium Signaling: Encoding and Decoding

Figure 2  Scheme of the main cellular events modulated by local (microdomains) and global (waves and oscillations) Ca²⁺ signaling.

of Purkinje cells. PV is additionally found in a subpopulation of inhibitory interneurons, the stellate, and basket cells. Studies on deficient mice, together with in vitro work and the discovery of their unique cell type-specific distribution in the cerebellum suggest that these calcium-binding proteins have evolved as functionally distinct, physiologically relevant modulators of intracellular calcium transients. Analysis of different brain regions suggests that these proteins are involved in regulating calcium pools critical for synaptic plasticity (25). Interestingly, and unexpectedly, PV⁻/⁻ fast-twitch muscles are considerably more resistant to fatigue than the wild-type controls. This effect was attributed mainly to the increased fractional volume of mitochondria in PV⁻/⁻ fast-twitch muscle, where the mitochondria are suggested to functionally replace the slow-onset buffer PV based on similar kinetic properties of Ca²⁺ removal (26).

In addition to limiting the diffusion and the magnitude of Ca²⁺ transients by activating the Ca²⁺ extrusion systems and by buffering them using cytosolic Ca²⁺ buffers, it recently became evident that cells also use the spatial distribution of intracellular organelles, such as mitochondria (for a comprehensive review, see Reference (27)). The control of the ion diffusion throughout the cell has been found to control the shaping of cytosolic Ca²⁺ signals in different cell types. In pancreatic acinar cells, mitochondria strategically located beneath the granular region prevent the spreading of a Ca²⁺ wave from the secretory pole toward the basolateral region by accumulating Ca²⁺; a similar role was reported for mitochondria in rat cortical astrocytes. In addition, the clearing of local [Ca²⁺]c in the proximity of Ca²⁺ release channels plays a role in the modulation of their activity. The first example of this effect was reported in Xeropus oocytes where mitochondria buffer microdomains of [Ca²⁺], regulating the open probability of InsP₃ channels, relieving the inhibitory effect of Ca²⁺. As a consequence, the rate of Ca²⁺ efflux from ER and, in turn, the shape of cytoplasmic Ca²⁺ waves, becomes modulated. A similar role for mitochondria has been described also in mammalian cells where mitochondria suppress [Ca²⁺]c positive (or negative) effects on the InsP₃ or on ryanodine channels or plasma membrane channels. By dissolving these local (Ca²⁺) microdomains, mitochondria buffer the Ca²⁺ that enters T cells via store-operated Ca²⁺ channels. They sequester Ca²⁺ during periods of rapid Ca²⁺ entry and release it slowly after Ca²⁺ entry has ceased. The idea that mitochondria can prolong the period of [Ca²⁺], elevation in response to a transient episode of Ca²⁺ influx by slowly releasing stored Ca²⁺ was already proposed as important for Ca²⁺-dependent processes such as exocytosis and synaptic transmission.

We still have little understanding of how cells actually decode the information contained in the different "shapes" of the Ca²⁺ signaling and, in particular, in the frequency of Ca²⁺ signals. The molecular machines that are responsible for decoding frequency-modulated Ca²⁺ signals include CaM kinase II (CaMKII) and protein kinase C (PKC). CaMKII is a multimer, consisting of 6–12 monomers that could be identical or different. It has a broad tissue distribution, being particularly abundant in the forebrain, where it is supposed to be particularly important in decoding the frequencies of synaptic inputs. CaMKII phospho-tylates and regulates multiple cellular targets that contribute to neurotransmission, neuronal plasticity, cell excitability, gene expression, secretion and cell shape, and, especially, memory formation and storage (28). Evidence that CaMKII could act...
as frequency decoder of Ca\(^{2+}\)-oscillations comes from the work of De Koninck and Schulman (29), who had shown that, after the threshold for kinase activation was reached and some subunits of the enzyme had become autophosphorylated, the response of the kinase to low frequency stimuli increased, presumably because CaMKII autophosphorylation is functionally cooperative.

PKC is a member of a family of Ser/Thr phosphotransferases that is involved in numerous cellular signaling pathways. These enzymes possess two regulatory domains, C1 and C2, that are the targets of different second messengers. Conventional and novel PKCs migrate to the plasma membrane in response to increased levels of DAG, resulting in complete activation of the isoenzymes. Activation is produced by the direct binding of DAG to a motif known as protein kinase C homology-1 (C1) domain. The main role of C2 domains in conventional PKCs is to act as the Ca\(^{2+}\)-activated membrane-targeting domain. In the absence of receptor-mediated lipid-hydrolysis, PKC isoenzymes are, in most cases, cytosolic and autoinhibited by the binding of a pseudo substrate motif to the substrate-binding site. Receptor-mediated generation of DAG and elevation in [Ca\(^{2+}\)]\(_{i}\) result in the membrane recruitment of conventional PKC molecules. Once fully membrane-associated, the pseudo substrate motif is released, allowing substrate binding and phosphorylation. The use of the green fluorescent protein (GFP) has shown that conventional PKC isoenzymes undergo oscillatory plasma membrane associations, which are "phase-locked" with the underlying receptor-mediated oscillations in [Ca\(^{2+}\)]\(_{i}\). Each oscillatory membrane association may lead to a burst of PKC activity, which in turn induces transient bursts in substrate phosphorylation. As Ca\(^{2+}\) returns to the basal level between oscillations, PKC regains the autoinhibited conformation. This pause in PKC activity allows the dephosphorylation of the substrate to dominate (30, 31). Such precise control of PKC activity is crucial to decoding the information contained in [Ca\(^{2+}\)]\(_{i}\) oscillations, and it is worthwhile to note that the other Ca\(^{2+}\)-oscillation decoder (CaMKII) does not fully deactivate between oscillations (29).

Recently, it has been proposed that Ca\(^{2+}\) signals could be decoded also through a pathway involving the Ras GT-Pases (32). In neuronal cells, Ca\(^{2+}\)-microdomains generated through Ca\(^{2+}\)-influx through the NMDA receptors and Ca\(^{2+}\)-influx through the AMPA receptors, the activity of which modulates the conversion between Ras-GTP and Ras-GDP in a Ca\(^{2+}\)-dependent manner. This confined coupling is essential in the physiology of neuronal cells, because the Ras/ERK pathway is important in eliciting two forms of synaptic plasticity, long-term depression (LTD) and long-term potentiation (LTP), both processes that have been proposed to form part of the cellular basis of learning and memory. AMPA-Rs trafficking in the postsynaptic membrane, a process that also controls LTP and LTD, appears to be modulated through the Ras/ERK pathway by CaMKII-mediated phosphorylation (33). Efficient Ca\(^{2+}\)-mediated activation of the Ras/ERK cascade appears to be optimized according to the frequency of the Ca\(^{2+}\) oscillations thanks to the action of two Ca\(^{2+}\)-regulated Ras GAPs, CAPRI and RASAL. These two proteins have been recently identified as sensors of distinct temporal aspects of the Ca\(^{2+}\) signals. Whereas CAPRI detects the intensity of the Ca\(^{2+}\) signal and undergoes transient association with the plasma membrane, RASAL senses the frequency by undergoing synchronous and repetitive oscillatory associations with the plasma membrane (34).

Calcium Controls the Expression of Its Own Transporters

Control of gene expression by Ca\(^{2+}\) was identified in 1985 when it was shown that prolactin gene transcription was stimulated up to 200-fold by the elevation of intracellular Ca\(^{2+}\) in cultured CH3 cells (35). After this observation, the transcription of numerous other genes was also found to be stimulated by Ca\(^{2+}\). According to general consensus, Ca\(^{2+}\) acts through three major pathways (Fig. 3). The first pathway, which is probably the principal one, involves changes in the activity of several Ca\(^{2+}\)-dependent kinases and phosphatases, which in turn change the trans-activating properties of several transcription factors. The best-known proteins involved in this pathway are probably CaM-dependent kinases CaMKIV, CaMKII (36), and the CaM-dependent protein phosphatase calcineurin (37). All of these proteins need CaM to process the Ca\(^{2+}\) signal. They stimulate gene transcription by regulating the phosphorylation state of the transcription factor CREB, and, as a result, the extent of CRE (cAMP Response Element)-dependent transcriptional activation, or by dephosphorylating the transcription factor NF-AT, thus promoting its translocation to the nucleus, respectively.

The regulation of synaptic plasticity in neuronal cells is an important consequence of the activation of this pathway of gene expression, after Ca\(^{2+}\) entry from the extracellular environment. An interesting finding that recently emerged was that Ca\(^{2+}\) itself could control the expression of its own transporters, suggesting that the cell can finely adjust the amount, the type, and the distribution of the Ca\(^{2+}\) transporters according to its specific demands. Ca\(^{2+}\) regulation of the plasma membrane Ca\(^{2+}\) pump isoforms at the transcriptional level was firstly reported by Guerini et al. (38) in cerebellar granule neurons. When neurons were cultured under condition of partial membrane depolarization, which is required for their long term in vitro survival, they underwent a chronic and modest increase of the resting free Ca\(^{2+}\) concentration, which promotes the up-regulation of PMCA2, 3, and 1CII (1a) (that are typical of the adult brain), and the down-regulation of PMCA4CII (4a) (which is absent from adult rat cerebellum). Calcinurin regulates the disappearance of PMCA4a but not the up-regulation of the other PMCA isoforms (39).

Interestingly, a similar pattern of remodeling also occurs for the expression of the other Ca\(^{2+}\) extrusion system of the plasma membrane, the NCX, which has a high level of expression in brain. Three separate genes encode for the three different isoforms of NCX, and, as in the case of PMCA, a number of splice variants are generated at least at the transcriptional level. It has been shown that during the maturation of neurons, the number of NCX variants are generated at least at the transcriptional level.
Calcium Signaling: Encoding and Decoding

Figure 3  Signaling pathways that participate in Ca\(^{2+}\)-regulated gene expression. Two important enzymes (calcineurin, CN, and calmodulin-dependent kinases, CaMK), which by regulating the phosphorylated/dephosphorylated state of the transcription factors CREB and NF-AT, translate the Ca\(^{2+}\) signal into nuclear specific information, are indicated. The transcriptional regulator DREAM, which is an EF-hand Ca\(^{2+}\)-binding protein, and the basic-helix-loop-helix (bHLH) transcription factors, which bind calmodulin (CaM) or S-100 proteins, are also shown. NLS, nuclear localization signal. Ca\(^{2+}\) levels, through the indicated pathways, directly regulate the transcription of the genes for the membrane Ca\(^{2+}\) transporters PMCA, NCX, and InsP\(_3\)R.

of splicing variants of NCX1 decreased but the total amount of NCX1 protein did not change significantly. At variance with this result, the amount of NCX2 protein increased dramatically with time if the cultured neurons were kept in low KCl-containing medium, reflecting the behavior of the isoform in the cerebellum during the first week, but became rapidly down regulated after partial depolarization of the plasma membrane (40). The NCX3 gene was also affected by the depolarizing treatment: its transcription became up regulated when Ca\(^{2+}\) influx is promoted by high KCl treatment. Interestingly, the NCX2 transcript up regulation was dependent on the activation of calcineurin, whereas the effects on the NCX1 and NCX3 genes were instead calcineurin-independent. Further reports have indicated that the Ca\(^{2+}\)-mediated increase in the expression of type 1 InsP\(_3\)R in cerebellar granules (41) and in hippocampal neurons (42) is mediated by Ca\(^{2+}\) influx through L-type channels or NMDA-Rs. The expression of type 1 InsP\(_3\)R appears to be regulated through the activity of calcineurin, which dephosphorylates the NF-AT transcription factor, promoting its translocation to the nucleus and transcription activation. These regulation pathways underline the importance of regulating cell Ca\(^{2+}\) with absolute precision: To promote the survival of the neurons, cytosolic Ca\(^{2+}\) must increase, but only to the relatively modest level that is necessary and no more. Evidently, to better control Ca\(^{2+}\) levels, cells adjust the abundance and the type of the different Ca\(^{2+}\) transporters.

Another important recent development in the transcriptional autoregulation of the Ca\(^{2+}\) message is that linked to the second pathway that Ca\(^{2+}\) uses to control gene expression: Ca\(^{2+}\)-ions bind directly to the transcriptional regulator DREAM (DRE Antagonist Modulator) and change its affinity for the DNA, relieving the repression on the transcription of specific target genes (43). DREAM belongs to the family of neuronal calcium sensors. It contains 4 EF-hands and acts as a transcription silencer for a large number of genes by binding to specific DRE (Downstream Regulatory Element) sites in the 5′ UTR region of the gene promoters. When Ca\(^{2+}\) becomes bound to DREAM, presumably as a result of its increase in the intracellular (intranuclear) environment, the DRE sites release DREAM and transcription resumes. The first discovered target gene for DREAM was that for the human prodynorphin, a protein involved in memory acquisition and pain (43); but an increasing number of genes have now been found to be regulated by this Ca\(^{2+}\)-sensitive transcriptional repressor. Very recently, DREAM has been shown to control the transcription of the gene for NCX3, an isoform of NCX that is important in Ca\(^{2+}\) extrusion.
in neurons (44). Overexpression of a DREAM EF-mutant (EFm- 
DREAM) insensitive to Ca2+ in hippocampus and cerebellum of 
transgenic mice significantly reduced NCX3 mRNA and protein 
levels. Cerelinar granaules from EfMDREAM transgenic mice 
displayed increased levels of cytosolic Ca2+, lost the ability to 
efficiently export Ca2+, and were more vulnerable to increased 
Ca2+ influx after partial opening of VOCCs. However, they 
survived better under conditions of reduced Ca2+ influx, sug-
gesting that DREAM plays a role in the autoregulation of the 
Ca2+ signal in neurons.

Lastly, the third mode that Ca2+ uses to control gene ex-
pression must be mentioned, even if no evidence has yet con-
clusively shown that this mode is employed to control the 
transcription of the Ca2+ transporters. This pathway involves 
the basic-helix-loop-helix (bHLH) transcription factors that, af-
ter Ca2+-dependent interaction with CaM or with the S-100 
proteins, modify the ability to bind to DNA and thus to activate 
transcription (45).

Measuring Ca2+ Concentration

Ca2+ probes (also known as indicators, reporters, or sensors) 
are molecules that form selective and reversible complexes 
with Ca2+ ions. The physicochemical characteristics of the 
Ca2+-free and Ca2+-bound form are sufficiently different to 
enable their relative concentrations to be measured. They can be 
divided in two main categories: synthetic Ca2+ indicators and 
protein-based Ca2+ indicators. In the late 1970s, Roger Tsien 
synthesized the first fluorescent Ca2+ probe for intracellular use. 
Its structure was based on the selective Ca2+-chelator EGTA. 
The portion of the molecule that binds Ca2+ is a carboxylic 
backbone perfectly adapted to the dimension of the ion, which 
confers its specificity. A fluorophore group, associated with 
the carboxylic group, endows the molecule with fluorescent 
properties dependent on the binding of Ca2+ to the carboxylic 
cage. Tsien and co-workers modified the original molecule 
by esterification of the charged carboxylic groups, making it 
permeable through the plasma membrane, it is trapped in the cytoplasm 
thanks to the action of cellular extremates that hydrolyse the 
ester, yielding the active form of the indicator. Thanks to 
the simplicity of use, this type of indicators is enormously 
employed by many researches. Big developments have been 
made to improve their fluorescent signal that, together with the 
improvements of the instrumentation, have contributed to image 
a single cell level the changes in Ca2+ concentration induced 
by physiologic stimuli.

In recent years, the wide diffusion of molecular biology tech-
niques has extensively expanded the number of applications of 
protein probes in cell biology. Two types of protein probes 
currently employed derive from bioluminescent organisms. The 
first is the group of chemi-luminescent proteins, which emit light 
usually in response to changes of a physiologic parameter, such 
as concentration of ATP or Ca2+. Among these proteins, the 
photoprotein aequorin (AEQ) has dominated the Ca2+-signaling 
field. The second group of protein probes is that of fluorescent 
proteins, of which GFP from Aequorea victoria is the “protag-
onist”.

AEQ was largely employed before the introduction of syn-
thetic fluorescent probes, as it was extracted and purified from 
jellyfish and microinjected in giant cells to monitor Ca2+. A 
Ca2+-induced conformational change in the apoprotein leads to 
peroxidation of the coenzyme, which results in the release of 
blue light. The rate of the reaction depends on the Ca2+ con-
centration to which the photoprotein is exposed. The cloning 
of the AEQ cDNA in 1985 (47) opened the possibility to ex-
tend AEQ use to a large variety of cells by transfecting them 
with a plasmid that allows recombinant expression of exoge-
nous protein. But the most important incentive to reconsider 
AEQ to monitor Ca2+ was the possibility to target it to a spe-
cific cell compartment by introducing in its sequence specific 
signal sequences. This approach has allowed the construction 
Ca2+ probes, which, in contrast to fluorescent dyes, are exclu-
sively localized in the intracellular district of interest. The use 
of chimeric AEOs allowed substantial advances in our under-
standing of Ca2+ signaling, such as the interplay between the 
ER and mitochondria (48), the presence of subplasma mem-
brane Ca2+ microdomains (49), and the role of Golgi apparatus 
as an important Ca2+ store (50). Despite its undoubted ad-
vantages, aequorin has a big defect: Although the amount of 
photons that are emitted from a cell population is more than 
adequate to measure Ca2+ concentration, the amount of photons 
that are emitted by a single cell is very low and not sufficient to 
guarantee a good space and time resolution. To overcome this 
deficiency, recombinant Ca2+ probes based on GFP were devel-
oped. Essentially, these indicators can be divided in two main 
groups based on their structure: double barrel probes, such as 
cameleon, made of two different coloured mutants of GFP con-
ected by a Ca2+-sensitive linker and single barrel probes, such 
as campanor and pericams, based on a single GFP engineered 
to bear a Ca2+-dependent inserted sequence. Generally, Cam 
is used as a molecular switch, which changes its conformation on 
the binding of Ca2+. The first two GFP-based fluorescent Ca2+ 
indicators were developed in 1997 (51, 52). Both probes are 
based on a similar strategy: the change of Fluorescence Reso-
nance Energy Transfer (FRET) between the two GFP mutants 
that is caused by the interaction between Ca2+-activated CaM 
and the targeted peptide. Although the cameleons were greatly 
improved over the original design, they still displayed insuffi-
cient signal-to-noise ratio when targeted to the organelles. To 
further increase the dynamic range of cameleons, the second 
generation probes (campanor and pericams) have been devel-
oped. In this construct, the GFP β barrel has been cut open and 
the original N- and C-termini linked together to create new ter-
imini located close to each other. The circular permuted (cp) 
GFP retains the ability to form a chromophore, but the permu-
tation renders the chromophore more accessible to changes in 
the pH of the surrounding ambient. As a result, the variation 
around the chromophore changes its ionization state, the 
fluorescence of all cpGFP is highly pH sensitive. This property 
has been exploited to detect Ca2+ by fusing Cam to cpGFP. In 
this way, the binding of Ca2+ mimics alkalization or acidifica-
tion, and so results in the increase or decrease of chromophore 
fluorescence. Further developments have been made to improve
the Ca\(^{2+}\) sensors and, in particular, CaM has been substituted by troponin C (53) or by an artificial Ca\(^{2+}\)-binding module to reduce the interaction with endogenous proteins (54), but the "perfect" Ca\(^{2+}\) probe is still missing.

Conclusion

More than 40 years of investigations of Ca\(^{2+}\) transport and Ca\(^{2+}\) actions have resulted in the well-defined concepts of intracellular Ca\(^{2+}\) homeostasis and Ca\(^{2+}\) signaling. The molecular mechanisms of Ca\(^{2+}\) transport between cells and extracellular space as well as within cells are clearly defined. We also know that Ca\(^{2+}\) signaling is extremely compartmentalized and that local and transient gradients could be responsible for spatial signal encoding. The study of the mechanisms of encoding/decoding the Ca\(^{2+}\) signal is fundamental to understanding the Ca\(^{2+}\) signaling pathways carried out in cultured cells or in tissue slices has clearly contributed to the progression in this field. Now, it will be nice when the recent technological developments drive the analysis toward in vivo cellular imaging, where the effects of physiology and pathologic stimuli could be investigated on the whole living organisms.

References

Calcium Signaling: Encoding and Decoding

Further Reading


Calcium Signaling: Encoding and Decoding

Further Reading

Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication

Arthur Wuster and M. Madan Babu, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

doi: 10.1002/9780470048672.wecb501

In this article we review cell-to-cell communication in three groups of organisms: bacteria, fungi, and the amoeba Dictyostelium. In particular, we concentrate on cell-density-dependent quorum sensing, which seems to have evolved independently on several occasions. The biologic and chemical properties of select systems are outlined, and the way they affect transcription, often via cross-genome interactions, is described for each of the three groups. Furthermore, we give a short overview of the methods that can be used to study quorum sensing. We also outline the general properties of quorum sensing systems, such as positive feedback and its link to starvation-dependent pathways.

Cell-to-cell signaling is a prerequisite for the development of multicellular organisms, such as animals and plants, but has also evolved in groups that would not usually be described as multicellular, such as bacteria and unicellular fungi. In this article we review and compare cell-to-cell signaling in several representative organisms, which are normally unicellular and never or only transiently aggregate to form multicellular bodies.

Quorum sensing is a process of cell-to-cell communication by which individual cells regulate their phenotype in response to the extracellular concentration of pheromones. This regulation is achieved by the secretion of pheromones into the environment that bind sensory proteins and directly or indirectly affect transcription and translation. The binding threshold is assumed to be reached once the growing population, and hence the secreted pheromone, reaches a certain density. In what follows, we will use the term quorum sensing system to mean a cell-to-cell communication system in unicellular organisms, which functions via the secretion of pheromones into the environment and their subsequent binding to sensor proteins. The term pheromone as we use it here refers to any chemical molecule that is excreted by one organism and changes the behavior of another one. It is not necessary that the pheromone synthesis or detection machinery has evolved for this purpose, although this is often the case. Different systems can be distinguished by the different types of pheromones they use, which are normally associated with different types of signal synthesis, import and export, and reception and response machinery.

In this article, we will concentrate on cell-to-cell communication in the form of quorum sensing in two unicellular groups, bacteria and yeast. Additionally, we will consider intercellular signaling in one species that is at the boundary between unicellularity and multicellularity, Dictyostelium discoideum. To illustrate the diverse mechanisms of cell-to-cell communication, we will discuss their biologic function, on the one hand, and the molecular details of pheromone synthesis, signal transduction, and transcription regulation in each of these groups, on the other hand.

The study of cell-to-cell communication and its effects on transcription in unicellular organisms promises a variety of practical applications. One of them is the possibility of interfering with intercellular communication systems in pathogenic microbes, which is a process also referred to as quorum quenching. A short mention of this topic and a more general attempt to integrate cell-to-cell communication in unicellular organisms with other pathways and processes, such as starvation, will be given toward the end of this article.

Although this article can be read by itself, we strongly recommend the reader consult the supplementary material on http://www.mrc-lmb.cam.ac.uk/genomes/awuster/wecb/ for links to additional literature and other relevant information.
Cell-to-Cell Communication in Bacteria

Quorum sensing involves dedicated cellular systems for the production and detection of pheromones, sometimes called quorumones. In bacterial species that employ quorum sensing, each cell secretes a basal amount of pheromones at low cell density. As cell density increases, pheromone concentration also increases, provided that the cells are not too far apart. Pheromones bind to specific receptors once their concentration exceeds a certain threshold, which in turn produces the physiological response. It has been shown that in addition to population density, diffusion barriers can be sensed in the same way. For example, Staphylococcus aureus can induce quorum sensing-dependent genes when confined in a host endosynome (1). It has even been proposed that most pathways attributed to quorum sensing in fact are diffusion-sensing pathways (2), although evidence for this is relatively scarce. In bacteria, quorum sensing-regulated phenotypes include bioluminescence, antibiotic and exoenzyme production, biofilm formation, and growth inhibition (3). Types of pheromones discussed in this section are acyl homoserine lactones (AHLs), AI-2 molecules, and modified oligopeptides. AHLs mostly affect transcription via a one-component signal transduction system, where the pheromone-binding protein domain is fused to a DNA-binding domain. Peptide pheromones and AI-2, on the other hand, often affect transcription via two-component signal transduction systems (TCSs) composed of a histidine kinase and a response regulator domain. Peptide pheromones and AI-2 in some instances, the AI-2 signal transduction cascade can also be composed of more than two components. The structure of the pheromones as well as their cognate system components are listed in Table 1 (columns 1–3).

Acyl homoserine lactones

Quorum sensing via acyl homoserine lactones is the best characterized bacterial cell-to-cell communication system. AHLs are often referred to as autoinducer-1 (AI-1)-type molecules. The term autoinducer has been chosen because synthesis of AHLs is regulated by positive feedback (Fig. 1), as has been described in the bacterium Vibrio harveyi (5). AHLs are composed of a homoserine lactone ring with an attached fatty acid chain, which can vary in length between 4 and 18 carbons and may or may not have a keto-group in position 3 (6, 7). For example, in V. harveyi, the AHL synthase LuxI produces C12 homoserine lactone with a keto-group on the third of six carbons, but in Agrobacterium tumefaciens, the LuxI homologue TraI produces the C12 molecule with a keto-group on the third of eight carbons. In most studied systems, AHLs are synthesized from S-adenosyl methionine (SAM) and fatty acid carrier proteins by LuxI and its homologues. However, alternative AHL synthases that are not homologous to LuxI exist. These synthases include LuxI in V. harveyi and HhIs in Pseudomonas fluorescens. In many species, more than one AHL is synthesized by different LuxI homologues. For example, in Rhizobium leguminosarum, six different AHLs synthesized by four different LuxI homologues have been identified (3). The structure of the LuxI homologue in P. aeruginosa has been resolved, and a detailed mechanism of its function has been proposed (8).

Most AHLs cross membranes by diffusion and bind LuxR-like response regulators. LuxR-like response regulators are sensors and transcription factors that at the same time (9), which makes the system a very simple one (Fig. 2a). Signal transduction systems in which the signal binding domain and the transcription regulating DNA binding domain are fused are referred to as one-component signal transduction systems. They are the most common type of bacterial signal transduction system (4, 10, 11).

The N-terminal signal binding domain of LuxR-like proteins has a α-helical αC domain. This occurrence is particularly interesting as GAF domains are usually found in signaling and sensor proteins (12). The structure of the A. tumefaciens LuxR homologue TraR (3) shows that at binding AHL is deeply embedded in the protein. Contact with some conserved hydrophobic and aromatic residues is established via several hydrogen bonds directly or via water. AHL binding is high affinity, which means that bacteria can sense relatively small pheromone concentrations. The specificity of the LuxR-like protein for AHL is determined by the acyl binding pocket of the LuxR homologue (14), as AHLs differ only in their acyl chains. The C-terminal domain belongs to the DNA and RNA binding helix-turn-helix fold. LuxR-like proteins use this domain to bind their cognate DNA motifs, such as the palindromic Lux box in the case of V. harveyi in which it activates transcription of the luxCDABE operon. Because this operon encodes the AHL synthase LuxI, AHL synthesis is subject to positive feedback (Fig. 1). However, in other species, LuxR protein orthologues with AHL bound can act as transcriptional repressors, such as the case for P. aeruginosa RhlR and LasR, which have been shown by microarray analysis (15, 16) to negatively regulate multiple genes.

Many, but not all, LuxR homologues dimerize at the ligand binding the N-terminal domain. An example of this dimerization is the A. tumefaciens LuxR homologue TraR. Furthermore, it is likely that the conformational changes caused by ligand binding modify the DNA-binding abilities of the protein (10). Modulated DNA affinity of TraR has been observed. It is also possible that ligand binding affects the LuxR protein ability to interact with RNA polymerase. Furthermore, AHL binding to LuxR has been shown to stabilize it, and in its absence, degradation is more likely (6).

Because of the positive feedback effect AHLs have on regulating the transcription of their own synthases, concentrations of AHLs can vary enormously between lower density and higher density cultures, such as biofilms. P. aeruginosa 3OC12 homoserine lactone (with a keto-group on the third of the 12 carbons of the acyl chain) has been measured to have a concentration of 2–10 µM in a standard lab culture and of up to 600 µM in the vicinity of an in vitro biofilm (17). However, experimental sensitivity and issues with AHL stability might have influenced the above results.

Although the basic AHL quorum sensing system only consists of two proteins—LuxI and LuxR—a variety of proteins modulates it. These proteins include the AIA AHL lactonase of Bacillus thuringiensis, which inactivates AHL by hydrolyzing...
Table 1. Components of cell-to-cell communication systems discussed in the text

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Vibrio fischeri</td>
<td>Acyl homoserine lactone (AHL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>Autoinducing peptide (AIP, encoded by agrD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vibrio harveyi</td>
<td>AI-2: furanosyl borate dimer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>DIF-1: chlorinated hexanophenone</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Aromatic alcohols: phenylethanol</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Synthase**
- **Vibrio fischeri**: LuxI, synthesis from S-adenosyl methionine (SAM) and fatty acid carrier protein.
- **Staphylococcus aureus**: Prepeptide modified by AgrB, which adds thiolactone ring.
- **Vibrio harveyi**: LuxS, synthesis from 4,5-dihydroxy-2,3-pentadione (DPD).
- **Dictyostelium discoideum**: Steely enzyme; subsequent chlorination and O-methylation.
- **Saccharomyces cerevisiae**: Aro9p, Aro10p.

**Transporters (exporters and importers)**
- **Vibrio fischeri**: no transporter—AHL can diffuse through membranes.
- **Staphylococcus aureus**: ABC exporter.
- **Vibrio harveyi**: no exporter for AI-2. no importer in *V. harveyi*, but importer is present in *Escherichia coli* and *Salmonella* (Lsr ABC-type transporter).
- **Dictyostelium discoideum**: no transporter—DIF-1 can diffuse through membranes.
- **Saccharomyces cerevisiae**: no transporter—aromatic alcohols can diffuse through membranes.

**Sensor**
- **Vibrio fischeri**: LuxR, AgrC transmembrane sensor kinase binds extracellular AIP, and phosphorylates cytoplasmic AgrA.
- **Staphylococcus aureus**: unknown.
- **Vibrio harveyi**: ABC-type transporter.
- **Dictyostelium discoideum**: LuxO transmembrane sensor kinase binds periplasmic complex LuxP-AI-2, and phosphorylates cytoplasmic LuxU.
- **Saccharomyces cerevisiae**: signal passed on to the protein kinase Tpk2p and subsequently to Flo8p.

**Transcription regulator**
- **Vibrio fischeri**: AgrA.
- **Staphylococcus aureus**: LuxO.
- **Vibrio harveyi**: DimAB.
- **Dictyostelium discoideum**: Flo8p.

For columns 1–3, also see Fig. 1.
**Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication**

The homoserine lactone ring. It is an instance of quorum quenching, more examples of which will be given later. Other LuxR homologues, such as *E. coli* SdiA, recognize more than one AHL (6, 7), which might be because SdiA is probably used to detect AHL produced by other species, as no LuxR homologue has been detected in *E. coli*.

**Autoinducer-2**

An overlap of AHL signaling exists with other quorum sensing systems, such as observed in *V. harveyi*. Apart from an AHL system, *V. harveyi* has a parallel system whose signaling molecules are referred to as autoinducer-2 (AI-2). In *V. harveyi*, AI-2 is a furanosyl borate diester, whose precursor is 4,5-dihydroxy-2,3-pentanedione (DPD). DPD is synthesized by LuxS from S-adenosyl methionine (SAM). Thus, SAM is a precursor in both AHL synthesis and AI-2 synthesis. It is assumed that DPD spontaneously rearranges into AI-2 when borate is available (18). In other species this rearrangement does not occur, and AI-2 has a different structure. The homologues of the DPD synthase LuxS are distributed widely among bacteria, but it is not clear whether they all produce furanosyl borate diesters. For example, LuxS in *S. aureus* was not shown to have any involvement in quorum sensing (19), because LuxS is not devoted solely to AI-2 production but also has a function in the methionine metabolic pathway.

In *V. harveyi*, in the absence of the AI-2 pheromone, the membrane-bound kinase LuxQ undergoes autophosphorylation on a conserved histidine residue (20, 21). The phosphoryl group is then transferred from the histidine of LuxQ to an aspartate of the response regulator LuxU. Phospho-LuxU in turn phosphates LuxO. Phospho-LuxO together with the sigma factor σ^73 then activates transcription of a set of small RNAs (Fig. 3c). These small RNAs, together with the RNA chaperone Hfq, contribute to the degradation of the mRNA of the LuxR transcription factor. LuxR is therefore the ultimate effector of the system in the presence of AI-2 in *V. harveyi*. It is important to distinguish this LuxR from the nonhomologous LuxR protein that binds AI-2 in *V. fisheri* as discussed in the previous section.

In the presence of the AI-2 pheromone, AI-2 is bound by LuxP (Fig. 2c), which is a periplasmic binding protein. LuxP–AI-2 binding to the LuxQ kinase triggers a dephosphorylation cascade by turning the kinases into phosphatases. LuxQ dephosphorylates LuxU, and LuxU dephosphorylates LuxO. The result is that the mRNA of the transcription factor LuxR is no longer degraded by sRNAs, and it can regulate its target genes (22). Interestingly, LuxU can also be phosphorylated by two other mechanisms. The first mechanism involves 3OC_4 homoserine lactone that binds to the LuxS sensor kinase, which in turn dephosphorylates LuxU. The second mechanism acts via the unidentified autoinducer CAI-1. Thus, AI-2, 3OC_4 homoserine lactone, and CAI-1 funnel their signals into one common system. Furthermore, sensing of the AI-2, 3OC_4 homoserine lactone by a membrane-bound kinase in *V. harveyi* shows that AHLs can also be sensed by pathways that are dissimilar to the one-component pathway outlined in the previous section. AI-2s are found only in Gram-negative bacteria, whereas AI-2 regulates phenotypes in genera as diverse as the firmicute *B. subtilis* and the γ-proteobacterium *V. harveyi*. As a result, AI-2 has been proposed to be involved in interspecies communication (23). Homologues of LuxQ have only been found in *Vibrio* species. Therefore, it can be assumed that sensing of AI-2 occurs in a different way in other species. For example, in *E. coli* and *Salmonella typhimurium*, AI-2 is imported into

---

**Figure 1** Three bacterial quorum sensing systems (also see Table 1, columns 1–3). The gray ovals represent cells; the small circles are the quorum sensing molecules (AHL, brown; AP, light blue; furanosyl borate diester, violet). Receiver and signal transduction molecules in blue, DNA binding response regulators in green, transcriptional interactions, green arrows; phosphorylation/dephosphorylation cascades, blue arrows; metabolic pathways, black arrows. Sender and receiver cells are only shown separately to keep the figure simple; in reality most cells in a population will receive and send signals at the same time. (a) The *Vibrio fisheri* AHL system; SAM: S-adenosyl methionine; (b) the *Staphylococcus aureus* AI-2 system; (c) the *Vibrio harveyi* furanosyl borate diester system; SAM: S-adenosyl-L-homocysteine; MTAN: methylthioadenosine/S-adenosylhomocysteine hydrolase; DPD: dihydroxy pentadione. Interaction with other quorum sensing systems (CAI-1, 3OC_4 homoserine lactone; also see main text) not shown.
Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication

Figure 2 Representative domain organization and structure of some proteins involved in cell-to-cell communication. (a) The AHL binding transcription factor TraR; (b) EsaI, a LuxI homologue and involved in AHL synthesis; (c) LuxP, a periplasmic protein binding AI-2; (d) LuxS, a protein involved in AI-2 synthesis; (e) the quorum quenching enzyme AiiA.

the cytoplasm by an ATP binding cassette (ABC) transporter. This transporter recognizes the periplasmic protein LsrB, which has the same periplasmic binding fold as LuxP. A-I-2 is then phosphorylated by the cytoplasmic kinase LsrK. It has been suggested that phospho-AI-2 subsequently interacts with the transcription factor LsrR (24).

Processed oligopeptides

Many Gram-positive bacteria use processed oligopeptides as pheromones. The precursor peptides are typically between 40 and 65 amino acid residues in length. These pre-peptides are processed (cleaved) in all known cases, and the resulting peptide pheromones are typically 5 to 34 residues in length. In many cases the peptides are also modified. The minimum components for peptide communication, apart from the peptide signal itself, are a membrane-bound histidine kinase and a response regulator with an aspartate phosphorylation residue. These components constitute a two-component signal transduction system, as opposed to the one-component system observed in AHL-based communication.

Peptide communication systems include the competence stimulating factor (CSF) of B. subtilis, which is a pentapeptide (sequence: ERGMT) derived from the 5 carboxy-terminal amino acids of the 40-amino-acid peptide encoded by phrC (25). Other systems include the lantibiotic nisin in Lactococcus lactis, which positively regulates its own expression. Staphylococcus epidermidis has a similar system to the lantibiotic epidermin (26). A peptide pheromone discovered in S. aureus is the autoinducing peptide (AIP), which is derived from the precursor AgrD to which a thiolactone ring is added between the carboxyl residue and a Cysteine residue at position 5 (Table 1, column 2) by AgrB. The AIP is sensed by the AgrC sensor kinase, which may dimerize at signal binding. The signal is then passed on to the AgrA response regulator, which activates transcription of the agr operon (6, 27), therefore again leading to autoinduction. Both AgrD and ComX, another peptide pheromone precursor found in B. subtilis, have an amphipathic motif in their N-terminal region, which might serve the purpose of recruiting the peptide to the membrane where it can be processed (28). This recruitment is often done by a dedicated ATP binding cassette (ABC) transporter, which in some cases recognizes a so-called GG leader sequence (consensus sequence LSxxELxxixGG) to the N-terminus of the region that encodes the actual signal (29).

Peptide signals might be more flexible than the small molecules discussed, as they do not require a specialized synthase and can change to adapt to ecological niches by simple codon mutation (30). The observation that some genes encoding peptide signals are more variable than other genes might be an indication that this actually happens. Furthermore, peptides might also be more flexible in the sense that they could be modulated by the external environment of a cell, therefore serving
some kind of sensor function in addition to their function as communication molecules. It can be caused by the external environment (for example, pH, temperature, salinity, the presence of certain chemicals or enzymes including proteases) modifying the communication peptides, therefore changing their ability to change the activity of the sensor histidine kinase and the subsequent signaling cascade (30).

Other systems

Other bacterial quorum sensing systems include the P. aeruginosa quinolone signal (PQS; 2-heptyl-3-hydroxy-4-quinolone) synthesized by PqsH, which is hydrophilic and is exported out of the cell by a vesicular transport system analogous to the ones used by eukaryotes. It seems that PQS is directly facilitating vesicle formation, as pqsH mutants do not produce vesicles. However, the phenotype can be rescued by the addition of exogenous PQS. The advantages of vesicle formation in P. aeruginosa might include the high concentration of vesicle contents delivered, as well as a certain resistance to quorum quenching signal degrading substances (31). The PQS genes have also been found in Burkholderia, although the production of PQS in these species could not be confirmed.

In Streptomyces γ-butyrolactone, a molecule structurally resembling AHLs has been shown to be a quorum sensing signal (32). Interestingly, in the genome of the Rhodococcus strain, RHA1 homologues for protein domains of both the γ-butyrolactone synthase and the receptor can be found, which suggests that γ-butyrolactone might play a role in this organism too. This suggestion is in agreement with close genomic proximity of the synthase genes and the response regulator genes (32).

Fruiting body formation in the bacterium Myxococcus xanthus is facilitated by the diffusible A-signal, which seems to consist of six different amino acids generated by extracellular proteolysis, and the contact-dependent C-signal. The C-signal is a cell-surface protein (32). Contact-dependent communication has also been observed in E. coli, where it is facilitated by CdaA and CdbB and leads to growth inhibition (33). Potential CdaAB homologues have been identified in a wide range of bacteria.

The list discussed here is by no means complete (for a comprehensive list, see Reference (34)), and it is reasonable to expect that even more bacterial quorum sensing systems await discovery, especially when considering the vast amounts of microbial diversity made available by metagenomics projects.

Quorum quenching and cross-genome interactions

In the human pathogen P. aeruginosa it has been shown that disrupting its quorum-sensing system diminishes virulence (35). Quorum quenching often takes the form of enzymes degrading the pheromones. Two different sorts of enzymes doing this for AHLs have been described: AHLases [e.g., AiiA in B. thuringiensis (Fig. 2e)] hydrolyze the lactone ring, resulting in acyl homoserine, but AHL-acylases (e.g., AiiD) break the amide bond, cleaving homoserine lactone from the acyl side chain (36, 37). Quorum signals can also be quenched by organisms that do not produce the signals, presumably to gain an advantage over communicating bacterial species in the same ecological niche. For example, Rhodococcus can degrade AHL signals without having any known ability to produce them (38).

On the one hand, closely related bacterial strains can produce surprisingly diverse quorum sensing molecules. For example, in the AIP system, the gene regions encoding the AIP precursor AgrD are extremely variable between strains (30). The purpose of this is most likely to avoid quorum sensing interference between strains living in the same niche. On the other hand, identical quorum sensing molecules are used by a wide variety of bacterial species, such as seems to be the case for AI-2. In fact, quorum sensing molecules can have effects outside the eubacterial kingdom, interfering with transcription in eukaryotes. One reason why quorum sensing molecules might have an influence on eukaryotic gene expression might be in order to enable the eukaryote to respond to bacterial infection in a more timely fashion by being able to sense bacterial presence sooner than would normally be possible. A further possibility is that interference with eukaryotic transcription is in the interest of the bacteria, causing the host to express genes that create a more favorable environment. This interference is of particular relevance for pathogenic bacteria, such as P. aeruginosa, a pathogen of humans that infects all sorts of organs, including the lungs of cystic fibrosis (CF) patients. In the mouse model for CF, it has been shown that the P. aeruginosa 3OC12 homoserine lactone induces and represses the expression of several mouse genes, including the chemokine interleukin-8 (IL-8). IL-8 causes the migration of neutrophils to the site of infection. If these neutrophils become activated, this can then lead to tissue damage. Although the exact molecular mechanism of how AHLs can regulate eukaryotic transcription is not yet known, at least four possibilities exist: 1) AHLs have been shown to be able to cross mammalian cell membranes, which means that inside the cell they could bind specialized receptor molecules, although no such molecules have been identified yet, 2) The AHLs could regulate the function of enzymes in a nonspecific manner and therefore modulate metabolic pathways, 3) The AHLs could directly act as nonspecific substrates in metabolic pathways, 4) The AHLs could bind membrane-associated receptors without entering the cell and possibly could disrupt host signaling (37).

Another way in which quorum sensing can be disrupted is by pheromone analogs that have similar structures and functions as agonists or antagonists to the actual molecules. In Staphylococcus, where quorum sensing is also mediated by peptides, different strains use slightly different, but homologous, peptides. These peptides tend to interfere with each other and inhibit the quorum sensing cascades of other strains (23). Eukaryotes can use a similar approach by synthesizing molecules that mimic AHLs as has been demonstrated in plants (39). The advantage of this approach might be that in this way pathogenic bacteria are exposed to a high level of pheromones earlier in the infectious process than they would normally be, therefore “confusing” them by suggesting a higher population density than is actually the case. This situation can cause the premature expression of virulence genes, which gives the host immune system more time to respond (37).
Cell-to-Cell Communication in Yeast

In yeasts, several workers report on putative quorum sensing-like mechanisms. In all reported cases, the major phenotype affected by quorum sensing is the transition between the filamentous form and the solitary yeast form. For example, Histo-plasma capsulatum is a parasitic yeast that can exist either in its filamentous form in soil or in its yeast form as a parasite of humans. Once it enters the host, its morphology switches to the yeast form that synthesizes cell-wall polysaccharides, such as α-(1,3)-glucan. It has been shown that the glucan concentration increases in a cell-density-dependent fashion. A culture grown in fresh medium to which filtrate from a dense culture is added will produce glucan, which suggests the existence of a factor that promotes glucan incorporation into the cell wall (see the section “Tools and Techniques” for more details on this approach). Very similar results have been reported for the causative agent of the Dutch Elm disease, Ceratocystis ulmi.

Farnesol

The existence of a couple of different quorum sensing molecules has been reported for the human pathogen Candida albicans. At low densities, the cells develop germ tubes (filamentous protrusions), which are not observed normally at high cell densities, which suggests that the switch between unicellular yeast and filamentous form depends on cell density. A molecule that blocks the formation of these germ tubes at high cell densities has been identified as farnesol, which has only been shown to affect transcription. The product of the aromatic alcohol quorum sensing system of C. albicans is tyrosol (42). As opposed to farnesol, it promotes cell growth and the development of germ tubes at low cell densities. Expression profiling of cultures under conditions of reduced tyrosol concentration showed reduced expression of proteins involved in DNA synthesis and cell cycle regulation (43). Other putative C. albicans quorum sensing molecules include the substance MARS of unknown identity and, with diminuitive effect, farsenoic acid (41).

Aromatic Alcohol Derivatives

A more detailed picture of density dependent cell-to-cell communication has been uncovered for Saccharomyces cerevisiae (44). In this fungus, phenylethanol (a phenylalanine aromatic alcohol derivative) and tryptophol (a tryptophan aromatic alcohol derivative) have been implicated in quorum sensing, and a model of how they act on their target genes has been proposed (Table 1, column 5). As in the other examples of yeast quorum sensing discussed above, these molecules regulate the transition to the filamentous form. They synergistically affect the upregulation of FLO11 via the cAMP-dependent kinase Tpk2p (a PKA subunit) and the transcription factor Flo10p. Flo10p, the product of FLO10, is the GPI-anchored cell-surface flocculin protein and is essential for filamentous growth. S. cerevisiae strains with deletions of either TPK2 or FLO10 do not form filaments in response to aromatic alcohols (44).

As in bacterial quorum sensing, the production of these two aromatic alcohols is cell-density dependent. Another similarity to bacterial quorum sensing is the autocatalytic fashion of phenomone production. Arotol is an aromatic aminotransferase and catalyzes the first step of tryptophan, phenylalanine, and tyrosine catabolism. The expression of the genes ARO9 and ARO10 is induced by tryptophol via the Aro80p transcription factor. Both ARO9 and ARO10 are required to synthesize tryptophol from tryptophan. It has been speculated that Aro10p is also involved in the aromatic amino acid catabolic pathway (45). As tryptophol induces the enzymes required for its own synthesis, it results in a positive feedback loop. Therefore, cells at high densities produce more aromatic alcohols per cell than cells at low density.

Together with C. albicans and many other yeast species, S. cerevisiae switches to the filamentous form when starved of nitrogen. The signal conferring information about nitrogen shortage has not been characterized in detail yet. Its message is transmitted via a MAPK–PKA pathway and results in an increase in the transcription of filamentation genes. Overlaps of tryptophol/phenylethanol communication with the nitrogen sensing pathway exist: Ammonia (nitrogen rich condition) represses filamentous growth as well as the expression of the above-mentioned ARO9 and ARO10 genes. However, elements of the MAPK–PKA pathway do not seem to be affected by aromatic alcohol communication.

Microarray data show that aromatic alcohols upregulate the expression of around 150 genes by twofold or more. These genes have diverse functions, but it is remarkable that 70% of them were shown independently to be upregulated upon entry into stationary phase. In contrast to S. cerevisiae, C. albicans filamentation is not stimulated by tryptophol or phenylethanol, although high concentrations of tyrosol were shown to slightly stimulate biofilm formation. This result shows that the function of the aromatic alcohol quorum sensing system of S. cerevisiae is not conserved in C. albicans (44).

Cell-to-Cell Communication in Dictyostelium

D. discoideum is a protozoon that is closely related to opisthokonts, the group comprising animals and fungi. The Dictyostelium amoebae are interesting from the point of view of cell-to-cell communication: not only starvation-dependent signaling but also a complex network of developmental intercellular signaling is observed (46).
Starvation Response-Related Signaling

It is a well-reviewed fact (for a review see Reference 47) that in Dictyostelium development triggers a behavior that causes aggregation of the single amoeba. The signal controlling this behavior is cyclic AMP (cAMP). Genes involved in cAMP signaling include the receptor for extracellular cAMP (CAR1) and the cAMP synthase, adenylyl cyclase (ACA). Interestingly, expression of these genes increases even before starvation sets in toward the end of the exponential growth phase. The amoeba can discern that genes required for the starvation response should be expressed by sensing their local population density. A high population density means that more nutrients are consumed and that future starvation is more likely. The amoeba can sense population density in a similar way to quorum sensing in bacteria. Glycoproteins, termed prestarvation factors or PSFs, are secreted into the environment at a basal level. Once their concentration exceeds a certain level, the genes required for cAMP signaling are upregulated (48). The exact mode of function of the two PSFs identified so far has yet to be determined (49).

Signaling during the developmental cycle

Once the amoebae have gathered in response to cAMP signaling, they form a multicellular structure surrounded by an extracellular matrix. This structure has many characteristics of development, such as differential cell sorting, pattern formation, or cell-type regulation. Although these characteristics can also be observed in animals, in Dictyostelium, multicellularity seems to have evolved independently (50). To reproduce, the cells of the Dictyostelium slug differentiate into two basic cell types, known as prespore and prestalk cells. The prespore cells are relatively homogenous, whereas there are several prestalk cell subtypes. The prestalk cell subtypes can be distinguished by using the promoters of the extracellular matrix genes ecmA and ecmB. By using these genes to control the expression of reporter genes, it is possible to differentiate between prespore and prestalk cells.

The recent sequencing of the Dictyostelium genome has shown that it shares many genes with higher eukaryotes, such as animals, which are not found in fungi. Because fungi are more closely related to animals than to Dictyostelium, this suggests loss of these genes in fungi. An example of a category of such genes is the transmembrane G-protein-coupled receptors (the cAMP receptor-like family). Three families of G-protein-coupled receptors that were thought to be unique to animals prior to the sequencing of the Dictyostelium genome are the frizzled/smoothed, secretin, and metabotropic glutamate/ GABAB families (50). Another finding resulting from the sequencing of the Dictyostelium genome is the relative abundance of genes encoding polyketide synthases (51). Polyketide synthases are needed for the synthesis of many molecules involved in developmental signaling in Dictyostelium, some of which will be discussed here.

Differentiation inducing factors (DIFs; see Table 1, column 4) are polyketide-derived signaling molecules critical for orchestration of cellular differentiation. DIF-1, a chlorinated hexanaphene (1,3,5-dichloro-2,6-dihydroxy-4-methoxyphenyl hexan-1-one), has been proposed to induce formation of prestalk cells by activating the expression of prestalk specific genes such as ecmA mentioned above. DIF-1 concentration is regulated by negative feedback, wherein DIF-1 activates the DIF-1 dechlorinase, which in turn inactivates it.

However, DIF-1 clearly cannot be the only factor involved in prestalk development as mutants blocked in the final step of DIF-1 biosynthesis, which is catalyzed normally by an O-methyltransferase, still produce pink cells and develop aberrant fruiting bodies. The way in which DIF-1 regulates the expression of its target genes might be via DimA and DimB, which are DNA binding basic leucine zipper proteins that migrate to the nucleus on exposure to DIF-1. The inability of mutants with a dim deletion to express ecmA in response to DIF-1 is consistent with the idea that DimA and DimB are DIF-1-activated transcription factors (46).

DIF biosynthesis has been investigated recently (52), and the protein Steely has been described as a key enzyme. Steely consists of six fatty acid synthase (FAS) catalytic domains and a C-terminal polyketide synthase (PKS). The N-terminal acyl products of steely are transferred to the C-terminal PKS active sites, which then catalyze polyketide extension and cyclisation. The result is a chloropropenophene, which is chlorinated subsequently and finally methylated.

Recent research (51) has shown that two putative DIF-1 precursors, 1,3,5,6-tetrachloro-2,4,6-trihydroxyphenyl hexan-1-one and 1,3-chloro-2,4,6-trihydroxyphenyl hexan-1-one, are also inducers of stalk development. This indicates that the dimethyl and desmethyl-monochloro analogs of DIF-1 are sufficient for stalk formation, which does not exclude the possibility that the methyl forms serve a different purpose. A third stalk-inducing compound identified in the same study was 4-methyl-5-pentylbenzene-1,3-diol.

From current data, it can be assumed that signaling complexity in Dictyostelium development is higher than previously thought (51), which is also reinforced by the insight that development requires highly selective signaling systems (50). The description of the systems given here, including the PSF system and the DIF systems, is therefore far from exhaustive.

Tool and Techniques

Chemical

The identification of cell-to-cell communication molecules is a nontrivial task. At an abstract level, a simple assay that allows for ascertaining if a certain phenotype is influenced by quorum sensing is to grow cells in culture in stationary phase for some time. During that period, potential phenomones can accumulate. The cells are then filtered and/or centrifuged out of the growth medium and the remainder is purified. If the addition of this conditioned medium to fresh exponentially growing cells induces the phenotype in question, one possible explanation is that quorum sensing molecules affecting the phenotype
were present in the filtrate. High-resolution liquid chromatography/mass spectrometry (LC/MS) and nuclear magnetic resonance (NMR) spectroscopy can aid in the identification of the actual quorum sensing molecules by enabling the researcher to compare the spectra of synthetic molecules to molecules purified from the conditioned medium.

Genetic

Genetic techniques to investigate quorum sensing include the disruption of the genes in a pathway that produces quorum sensing molecules, knockout of sensor and response regulator genes, addition of purified or synthetic quorum sensing molecules, or addition of quorum quenchers. The identity of synthetic and actual (purified) communication molecules can be verified by adding synthetic molecules to a mutant culture that cannot produce its own signals. If the phenotypic effects are the same, this indicates that the synthetic and purified molecule share the same structure. It can, however, also happen that chemically analogous molecules elicit similar effects. For example, the quorum sensing homoserine lactone 3OC12 as produced by the bacterium P. aeruginosa has been shown to mimic the effects of farnesol in S. cerevisiae, probably because it is somewhat similar in its side chain structure.

Computational

Quorum sensing can also be studied from a different viewpoint, for example, by computational methods. The structure of many quorum sensing molecules as well as of their cognate synthases and receptor molecules has been solved by X-ray diffraction analysis or NMR. A list of PDB entries associated with quorum sensing is available from the supplementary materials section on http://www.mrc-lmb.cam.ac.uk/genomes/awuster/wecb/. and some structures are shown in Fig. 2. These structures provide invaluable insights in determining the molecular mechanisms of quorum sensing. Similarly gene expression microarray datasets under quorum sensing conditions are available publicly. Integrating structural and gene expression data with the abundantly available genome sequence data will therefore be of great value for the discovery of new quorum sensing systems as well as for the integration of quorum sensing with intracellular signaling pathways.

Discussion and Concluding Remarks

A Link Between Quorum Sensing and Starvation

It is easy to see that in a growing population limited resources will become scarce sooner or later, which suggests that there is a fundamental link among population density, growth, and starvation. One such link has been uncovered in P. aeruginosa, where an AHL induces the transcription of the gene encoding σ53, a sigma factor that is active maximally under conditions of starvation (53). In yeast, the link has been demonstrated in S. cerevisiae with ammonia—a often a limiting resource for growth— inhibiting the expression of ARO9 and ARO10, two enzymes involved in the production of the aromatic alcohols functioning as quorum sensing pheromones. In Dictyostelium, the pre-starvation factors regulate changes in gene expression in a cell-density-dependent manner toward the end of the exponential phase during the vegetative cycle to be prepared for the secretion of cAMP, which is the starvation signal and leads to the aggregation of cells.

Positive Feedback

Positive feedback is another theme common to quorum sensing systems. The purpose of positive feedback during quorum sensing is presumably to speed up the response time, or to give a more pronounced response. In many of the discussed cases, positive feedback takes the form of quorum sensing molecules increasing the expression of their own syntheses, such as in yeast, where the aromatic alcohols, via the transcription factor ARO80p, promote the expression of ARO9 and ARO10. Similarly, in the bacterial LuxI/LuxR system, the AHLs synthesized by LuxI promote their own synthesis when bound to LuxR by positively regulating the expression of luxICDABE operon. One result that can be anticipated from this is that cells at high densities secrete more autoinducer per cell than cells at low densities.

Interest in cell-to-cell communication in unicellular organisms is still increasing, which might also be because of two important realizations, the first one of which is the relevance of quorum sensing to many disease-related processes, for which P. aeruginosa with its multiple and interacting systems has become a model organism. Furthermore, interaction between eukaryotic hosts and microbial pathogens via quorum sensing molecules might prove to be of high importance for furthering our understanding of bacterial infectious diseases. The second realization is that quorum sensing is not limited to a few species but that it seems to be a next-to-universal mechanism in unicellular organisms, with a multitude of quorum sensing systems likely to be awaiting discovery.

Acknowledgments

A W and M MB acknowledge financial support from the Medical Research Council. We would like to thank Sarah Chandra Janga, Subhajyoti De, and Siarhei Maslau for critically reading the manuscript.

Supplementary website

http://www.mrc-lmb.cam.ac.uk/genomes/awuster/wecb/

References

10

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.


Chemical Molecules that Regulate Transcription and Facilitate Cell-to-Cell Communication


Further Reading

Supplementary material available at: http://www.mrc-lmb.cam.ac.uk/genomes/awuster/wecb/.


See Also

Intercellular Signaling
Quorum Sensing
Transcription Factors
Lipid molecules can act as extracellular signals. They form a large and expanding class of influences on cellular and organismal homeostasis, which affect all organ systems and participate in many disease processes. Prominent members of this class include eicosanoids (e.g., leukotrienes or prostanoids like prostaglandins, protacyclins, and thromboxane), lysophospholipids (e.g., lysophosphatidic acid and sphingosine 1-phosphate), endocannabinoids (e.g., anandamide and 2-arachidonoylglycerol), ether lipids (e.g., platelet activating factor), and free fatty acids. These signals are in a dynamic steady state with structural phospholipids that make up the lipid membrane bilayer of all cells. Both distinct and shared enzymologies are involved in lipid signal biosynthesis and degradation. A dominant mechanistic theme for all bona fide extracellular lipid signals is their use of cognate 7-transmembrane domain G protein-coupled receptors. Activation of these receptors causes myriad physiologic and pathophysiologic effects that encompass most aspects of cell biology and physiology. Studies on extracellular lipid signals have led to both mechanistic understanding and successful creation of useful medicines.

Extracellular lipid signals are small fat molecules that produce myriad cell signaling effects upon exposure to the extracellular surface of cells. The dominant mechanism for these effects is the activation of specific cognate cell surface receptors, called G protein-coupled receptors (GPCRs), which are characterized by a predicted 7-transmembrane-spanning structure and which activate most known, intracellular heterotrimeric G proteins. These G proteins in turn activate a wide range of downstream signaling pathways, with the actual physiologic responses dependent on receptor expression patterns and involved cell types. All of these lipid mediators are linked, directly or indirectly, to membrane phospholipids that can be thought of as a dynamic repository of signaling lipid precursors. A diverse, yet often overlapping, enzymatic machinery exists for both the biosynthesis and the degradation of these signals. Compared with peptidergic factors, the lipid mediators are often one tenth or less in mass, and they have brief half-lives, although binding to carrier proteins can substantially increase their stability. Historically, the understanding of these factors emerged in two phases. The first phase, which began about 70 years ago and continues today, was the important biochemical identification of lipids that showed bioactive properties in animals, and could be chemically isolated and structurally analyzed. This phase was marked by many mechanistic hypotheses to explain the observed bioactivities, including receptor hypotheses. Tools for the identification of lipid components include classic thin layer chromatography, liquid chromatography, mass spectroscopy, as well as the use of bioassays and isotope labeling.

The second phase began with the advent of molecular cloning of cell-surface lipid receptors that commenced in the late 1980s/early 1990s and that allowed rigorous assessment of the existence, pharmacology, and functional roles for extracellular lipid signaling. It included the use of genetics to create mouse mutants that allowed additional analyses of both receptor identity and essential physiologic or pathophysiologic roles. Most receptor-ligand interactions occur at nanomolar affinities, with the exception of free fatty acids that interact at micromolar concentrations. A mainstay of these studies is the use of modern cell and molecular biology leading to the creation of mutant cell lines and animals. A standard in all fields is the use of receptor overexpression or heterologous expression, particularly combined with the use of receptor-null cells and tissues created by targeted deletion of one or more of the receptors in question. Interestingly, deleting individual lipid receptors produces a variety of phenotypes, which demonstrates both the possible redundancy of some receptors as well as the unique functions that cannot be rescued by the large number of lipid GPCRs. These reagents have led to the development of specific chemical agonists and antagonists, including some that have become therapeutic drugs. Intracellular roles for some of these mediators have been documented; however, this aspect of lipid signaling...
will not be considered further in this article. Major extracellular lipid signals are considered next, in alphabetical order. Representing literally tens of thousands of primary references over the past 70 years, the reference list is necessarily incomplete, and the reader is referred to the recent primary literature for more in-depth data and discussion in this rapidly evolving and expanding field.

Eicosanoids

Eicosanoids, also referred to as eicosanoids, are so named because of the 20-carbon constituency that identifies this class of oxygenated lipid molecules. A primary synthetic pathway for these molecules involves the phospholipase-mediated cleavage of a membrane phospholipid to produce arachidonic acid [(all-2, conjugated linoleic acid)] and eicosapentaenoic acid (EPA). From this biologically essential intermediate fatty acid, two major subclasses of eicosanoids can be produced: 1) leukotrienes, via the action of lipooxygenases, and 2) prostanoids, via the action of cyclooxygenases (COX-1 and COX-2). Examples of chemical structures for a leukotriene (Fig. 1a) and three types of prostanoids (Fig. 1b-d) underscore their shared arachidonate origin.

Leukotrienes

Leukotrienes were identified in the 1930s as a bioactive substance released by smooth muscle in the lung after an antigenic challenge. Their current name reflects their well-recognized role in leukocyte (white blood cell) activities and the three prominent adjacent double-bonds in their chemical structure (Fig. 1a). There are multiple leukotriene species, including leukotriene (LT) A4, B4, C4, D4, and E4. The C, D, and E4 species are named cysteinyl leukotrienes because of the presence of the amino acid cysteine in their structure or slow-reacting substance for a leukotriene (Fig. 3a) and three types of prostanoids (Fig. 3b-d) underscore their shared arachidonate origin.

Prostanoids

These lipid factors were first identified in the 1930s as bioactive components of semen. Their name is derived from the original belief that they were secreted from the prostate gland. Their unique chemical structures are characterized by a trans double bond at C13-14, a hydroxyl group at C15, and a cyclic carbon ring (Fig. 3b-d). Three subclasses of prostanoids are defined by their cyclic ring configuration: 1) the well-known prostaglandins, which have a cyclopentane ring (Fig. 3b), 2) prostacyclins that contain a six-residue ring (Fig. 3c), and 3) thromboxanes that contain a six-residue ring (Fig. 3d). Naturally occurring prostaglandins include PGD2, PGE2, PGI2, and PGH2, of which the latter is a common precursor in the synthetic pathways of many prostanoids. PGF2a and TXA2 are the most common naturally occurring prostacyclin and thromboxane forms, respectively. Like leukotrienes, prostanoids are important lipid mediators in the inflammatory response and are not synthesized unless their source cells are activated (3). Because of their short half-life, these lipids are primarily paracrine and autocrine extracellular effectors, which signal through at least nine cognate GPCRs (6). Alternatively spliced isoforms of these receptors also exist, with varying affinities for individual prostanoids. Signaling through prostanoid GPCRs encompass at least three classes of heterotrimetric G proteins: Gs (receptors DP, IP, EP1, EP4), Gr, and Gq (receptor EP3) (7).

Prostanoids are most prominently recognized as proinflammatory factors that are vasoactive, cause blood coagulation, and produce associated symptoms of fever and pain (8). With the development of mouse receptor-null mutants, an increasingly wide range of biologic effects mediated by prostanoids is beginning to be appreciated (8). Complex physiologies associated with the numerous receptors are still being determined and point to kidney, cardiovascular, gastrointestinal, and reproductive system functions (6). Inhibitors of the rate-limiting enzyme in prostanoid synthesis, COX1 and COX2, have been the basis for nonsteroidal anti-inflammatory drugs, most prominently the nonselective COX inhibitors such as aspirin, ibuprofen, and naproxen, as well as the selective COX-2 inhibitors celecoxib (Celebrex; Pfizer, New York, NY) and the recently withdrawn drug, rofecoxib (Viscera; Merck).

Lipoxins

A third class of eicosanoids, the lipoxins (LX), has recently been defined based on its unique anti-inflammatory signaling. LXs derive their name from their generation via lipooxygenase interactions, requiring at least two lipooxygenases in their biosynthesis. Unlike the LTs and prostanoids, which are proinflammatory, LXs reduce inflammation through multiple pathways, including binding LX-specific GPCRs (e.g., LXA4), inhibiting LT-mediated inflammation through competitive inhibition of cysteine leukotriene receptors, and directly interacting with intracellular targets, e.g., transcription factors (10). Recently, the carbon-15 R epimers of LXs, termed aspirin-triggered lipoxins (ATLs), have been identified as a novel component in the anti-inflammatory mechanism of action for aspirin (10). Production of LXs has been observed in many human tissues, and reduced production is implicated in several inflammatory related...
Extracellular Lipid Signals

Figure 1. Chemical structures of representative eicosanoid molecules: (a) leukotriene, (b) prostaglandin, (c) prostacyclin, and (d) thromboxane. Common names, chemical formulas, and molecular weights (g/mol) are shown.

Endocannabinoids

The psychotropic properties of the plant-derived chemical compounds in marijuana, most prominently Δ9-THC (delta-9-tetrahydrocannabinol), have been known since the mid-1960s. The mechanism through which these exogenous compounds act was a subject of controversy until the early 1990s when the first cannabinoid receptors, referred to as CB1 and CB2, were cloned (discussed below). This quickly led to the identification of endogenous cannabinoid receptor ligands, "endocannabinoids," which fall into two major classes: the N-acyl ethanolamines (NAEs) and monoacylglycerols (MAGs). The NAE, "anandamide" or N-arachidonoylethanolamide (Fig. 2a), is a high affinity agonist for both CB1 and CB2 and can produce all of the known neurobehavioral effects of marijuana. Anandamide is synthesized through a two-step process from the membrane phospholipid, phosphoethanolamine, and is created in a one-to-one ratio with phosphatidic acid, which is another important lipid signal (12). The second major endocannabinoid is 2-arachidonoylglycerol (2-AG), a MAG with lower affinity for both cannabinoid receptors and many distinct physiologic effects. It has an esterfied arachidonic acid acyl chain in the sn-2 position of the glycerol backbone (Fig. 2b); thus, 2-AG can also be a precursor for some prostanoids (13). Many enzymes are involved in the biosynthesis and catabolism of endocannabinoids, regulating their diverse physiologic functions (14). One important catabolic enzyme is the fatty acid amide hydroxylase (FAAH) and inhibitors of FAAH are currently being investigated for their analgesic properties (15).

The cannabinoid receptors are GPCRs that are known to couple to Gi/o, although other coupling may exist in different cell types (16, 17). CB1 and CB2 have distinct expression profiles, with the former primarily expressed in the nervous system and the latter expressed in immune cells and in intestine (18). Within the central nervous system (CNS), CB1 receptors are widely expressed, particularly in the basal ganglia, the hypothalamus, and limbic areas, which suggests roles in mood, endocrine regulation, and memory (19). Indeed, genetic deletion of CB1 in mice is associated with enhanced short-term memory and weight loss (20). CB1 is also expressed in the male and female reproductive systems, although the role of cannabinoid signaling in reproduction is still unknown.

Diseases, such as arthritis, atherosclerosis, liver disease, cystic fibrosis, and periodontitis (11).
Extracellular Lipid Signals

Endocannabinoids

Figure 2 Endocannabinoid chemical structures: (a) anandamide and (b) 2-arachidonoylglycerol. Chemical formulas and molecular weights (g/mol) are given.

Cannabinoid receptors have been of great therapeutic interest in a variety of conditions, in particular obesity and neuropathic pain (21). Several medicines targeting these receptors are now on the market, including nabilone (Cesamet; Valeant Pharmaceuticals International, Costa Mesa, CA), dranabinol (Marinol; Unimed Pharmaceuticals, Deerfield, IL), and the recently released rimonabant (Acomplia; Sanofi-Aventis, Bridgewater, NJ), a selective CB1 antagonist for obesity management.

Ether Lipids

Platelet-activating factor (PAF, 1-O-alkyl-2-acetyl-sn-glycero-3-phosphocholine) was the first extracellular lipid signal for which a cloned receptor was identified (22). A hallmark of the molecule is an alkyl ether at the sn-1 position of the glycerolphosphate backbone (Fig. 3), in contrast with ester-linked molecules such as the lysophospholipids. In addition, the inactivation for PAF by platelet activating factor acetylhydrolase (PAFAH, a specialized phospholipase A2) is highly specific for the short-chain acyl group off the sn-2 carbon (23). PAF is the best characterized of the ether lipid signaling molecules, with potent activity on platelets and other immune cells, as its name implies (3). However, since its initial characterization in the immune system, a broad range of additional physiologic or pathophysiologic processes have been linked to this lipid molecule, including reproduction, neural migration, and blood circulation (24). Its receptor is a GPCR linked to numerous intracellular signaling pathways. Although the PAF receptor has widespread expression in many tissues, the subtle phenotype of the null-mutant mouse suggests the involvement of compensatory signaling mechanisms that involve other as yet unidentified receptors or redundancy of signaling mechanisms. In-depth studies using both transgenic and receptor-null models have demonstrated influences on allergic response, inflammation, infection, as well as osteoporosis (25), reproduction, wound healing, and subtle CNS functions (26).

Free Fatty Acids (FFAs)

The most recent class of molecules to be characterized as extracellular lipid signals are free fatty acids (Fig. 4). Fatty acids are essential residues of biologic phospholipids, and they can exist in numerous permutations of length and saturation. FFAs are “free” in that they are not linked to other molecules such as the glycerolphosphate backbone found in cell membranes and have important roles in both energy production and cell signaling. They can be produced through several metabolic processes such as cleavage of membrane phospholipids by several different phospholipases (e.g., phospholipase A). Both short-chain and long-chain FFAs of varying carbon lengths and saturation have demonstrated biologic activity consistent with being lipid signals. Supporting this interpretation, recent receptor de-orphaning research has identified a set of four

 Ether Lipid

Figure 3 The chemical structure of a well-known ether lipid, platelet activating factor, showing stereochemistry. Chemical formula and molecular weight (g/mol) are included.

Ether Lipid

Figure 4 The chemical structure of a well-known ether lipid, platelet activating factor, showing stereochemistry. Chemical formula and molecular weight (g/mol) are included.
Extracellular Lipid Signals

Figure 4 Examples of short-chain and long-chain free fatty acids are shown: (a) 4-carbon butyric acid and (b) 16-carbon palmitic acid. Chemical names, formulas, and molecular weights (g/mol) are given.

GPCRs—GPR40, 41, 43, and 120—that mediate a response to a range of FFA forms (27–29). These GPCRs couple to Gq/11 resulting in an increase in intracellular calcium (coupling has not been reported for GPR120, although it can alter calcium levels consistent with a Gq/11 coupling). An important distinction to be made between FFAs and other extracellular lipid signals is that, unlike the other receptor–ligand interactions considered in this entry, these FFA receptors respond only to microM concentrations, and do so with a relative lack of selectivity for specific FFA forms. This raises the possibility that other endogenous ligands with higher affinities might exist, although this remains to be determined.

The FFA receptors serve as sensors for levels of FFAs that allow homeostatic metabolic mechanisms to be activated. Expression is primarily in the pancreas and gut (GPR40, -41, -120), which is consistent with a role in metabolic functions. For example, GPR120 increases circulating blood insulin levels indirectly through the release of glucagon-like peptide 1 (GLP1) (27). Other receptor expression patterns in adipose tissue (GPR41), brain (GPR40), and lymphocytes (GPR43) have also been reported (28, 29). However, their functional significance is not yet known. Data from receptor-null mutants have not yet been reported.

Lysophospholipids

Representing one the simplest forms of phospholipids, these molecules are characterized by a phosphate headgroup attached to a glycerol or sphingoid 3 carbon backbone with acyl chains that can vary with respect to length and saturation. The two best-studied forms of lysophospholipids are lysophosphatidic acid (LPA) and sphingosine 1-phosphate (S1P) (Fig. 5). Bioactive forms of LPA are numerous and include 1-oleyl LPA (18 carbons, 1 double bond acyl chain at the sn-1 position), as well as forms with acyl chains at the sn-2 position. LPA can be produced by the action of PLA2 on phosphatidic acid, as well as through the action of autotaxin, a lysophospholipase D enzyme that produces LPA from LPC (30). By contrast, S1P is produced by the action of sphingosine kinase 1 or 2 on sphingosine resulting in the phosphorylated bioactive lipid (31). Lysophospholipids can be degraded by multiple lipid phospholipases, lipid phosphate phosphatases, and/or lyases. Other lysophospholipids, such as lysophosphatidylcholine (LPC), lysophosphatidylethanolamine (LPE), lysophosphatidylserine (LPS), and sphingosylphosphorylcholine (SPC), may have biological activity as extracellular signaling lipids; however, bona fide signaling via defined receptors remains unclear. Many cell types can produce most forms of lysophospholipids, but it is often not possible to distinguish the signaling pools versus structural or other nonsignaling pools of these lipids within a cell.

To date, five GPCRs have been identified for LPA and five for S1P: LPA1–5 and S1P1–5. Although most of these receptors share a high degree of homology within and even between ligand classes, lower homology receptors for both classes have been identified, particularly two recently reported LPA receptors, LPA4 and LPA5 (32–34). A wide range of G proteins are activated by these receptors, including Gi/o, Gq/11, G12/13, and Gs. As a group, these receptors show widespread expression during development and postnatal life, and they have been documented across phylogeny, from humans through fish. Clear cellular effects include cell survival, actin cytoskeletal alteration and cell shape changes, proliferation, electrophysiological changes, and transcriptional activation, among others (35). These effects in turn contribute to a rich and complex physiology and pathophysiology that has been prominently revealed by receptor-null animals, as well as other approaches (36). Loss of LPA receptors can affect normal development, craniofacial formation, bleeding, CNS development and function, initiation of neuropathic pain, and normal fertility via uterine implantation. S1P receptor deletion can affect embryonic survival (S1P1), vascular formation, immunologic lymphocyte egress, cardiovascular function, and fertility (37).
Lipoxins and aspirin-triggered 15-epi-lipoxins are the biologic and therapeutic importance of these extracellular lipid signals.

Acknowledgements

Thank you to Christine Paczkowski and Danielle Letourneau for critical reading of this manuscript. The author gratefully acknowledges NIH support from NIMH, NINDS, NIDA, and NICHD.

References

Extracellular Lipid Signals


Further Reading

www.lipidlibrary.co.uk is a reference website with up-to-date information on the structure, function, and analysis of lipids.
www.xpharm.com is a database containing pharmacology reference information.
Hormone Signaling
Ross V. Weatherman, Department of Medicinal Chemistry and Molecular Pharmacology and the Purdue Cancer Center, Purdue University, West Lafayette, Indiana
doi: 10.1002/9780470048672.wecb229

Hormone signaling has always been a field that has required working at the chemistry–biology interface. The development of new chemical tools plays a major role in understanding the molecular mechanisms of hormone signaling and the physiologic outcomes of hormone receptor action. This article will outline the relevance of molecular endocrinology to the field of chemistry and the use of chemical approaches to solving problems in hormone research. The basic biologic outline of a hormone signaling system and the different classes of hormone molecules and their receptors will be discussed. In addition, several research areas in which new chemical tools have been playing key roles will be described in more detail, including the discovery of ligands for orphan hormone receptors, the development of non-natural hormone receptor mimics, and the use of selective hormone receptor modulators to understand the role of specific hormone receptor signaling pathways. Hormone signaling systems that will be discussed include the thyronamines and trace amine receptor, the liver X and farnesol X receptors, small molecules capable of binding to peptide hormone receptors, the melanocortin receptor family, the estrogen receptor, and chemically orthogonal hormone receptor–ligand pairs. Major challenges facing the field as well as some different experimental methods used to study hormone signaling will also be discussed.

At a fundamental level, multicellular life would not be possible if it were not for the ability of different cells to communicate and coordinate using a language based on biomolecular interactions. From the simplest two-component signaling systems found in primitive multicellular colonies to the complicated networks of overlapping signaling cascades found in higher vertebrates, hormone signaling is dependent on the production, diffusion, and recognition of small molecules and peptides. As such, hormone signaling has always been a productive area of study for people working at the chemistry–biology interface. Chemistry has played an important role in elucidating the molecular mechanisms and consequences of hormone signaling and will continue to be a vital tool in tackling the complicated new challenges that face the field. This advanced review will introduce some key features of hormone signaling that are relevant to the chemist: why a chemist should care about hormone signaling, some basic features of hormone signaling systems, and the archetypal classes of hormones and hormone receptors.

The Basic Biology of Hormone Signaling

A hormone, simply defined, is a chemical messenger that relays signals from one cell to another. With such a broad definition, almost any biomolecule could be a hormone in certain contexts, but this article will focus on classically recognized endocrine and paracrine hormones that typically are produced in specific groups of cells in one organ and target nearby cells (paracrine) or cells or tissues in distant organs (endocrine). Even with a focus just on these classes of hormones, the biology of hormone signaling is too vast to be covered by writing a large textbook let alone a small review article. As a result, this section will concentrate on issues relevant to the point of view of the chemist: why a chemist should care about hormone signaling, some basic features of hormone signaling systems, and the archetypal classes of hormones and hormone receptors.

Why chemists should care about hormone signaling

In a schematic of a typical hormone signaling pathway, one often sees a jumble of arrows pointing to various geometric shapes...
Hormone Signaling

Hormones also play a key role in the development and treatment of many hormone receptors (1–5). As will be discussed, the development of radiolabeling techniques and molecular biology led to the isolation and characterization of many hormones, and the development of radiolabeling techniques and molecular biology led to the isolation and characterization of many hormone receptors (1–5). As will be discussed, the development of new compounds with selective modulation of specific hormone signaling pathways has greatly increased our knowledge of the molecular underpinnings of hormone action (10–12).

Hormones also play a key role in the development and treatment of a large number of diseases ranging from breast and prostate cancer to diabetes and obesity (13, 14). A significant percentage of people in the United States target hormone signaling, including drugs used to treat inflammation, diabetes, hypertension, hypothyroidism, and other hormone-related indications (13–15).

Basic features of hormone signaling

At its most basic level, hormone signaling must involve a hormone and some sort of target. In most cases the target is a receptor protein, although some hormones have been proposed to function also by non-receptor-mediated mechanisms such as altering the local redox environment of a particular cell (16, 17). Basic hormone signaling follows a simple signaling loop (see Fig. 1). Every component of a hormone signaling loop can be affected through the use of chemical tools. In an endocrine or paracrine signaling system, two different cells exist: one making the hormone and one possessing the receptor and carrying out the biologic response. In the hormone-producing cell, typically some sort of stimulus exists to initiate production of the hormone, usually some sort of change in the concentration of a marker sensed by the cell or perhaps another hormone. The hormone is then produced either by biosynthetic enzymes or by release from storage vesicles inside the cell. The hormone then must be secreted out of the cell and transported to the target cell. This secretion and transport either can be a passive process or can involve other proteins. After reaching the target cell, the hormone can bind to the receptor, if it is extracellular, or it can be taken into the cell by either passive or active transport and bind to intracellular receptors. Once bound to the receptor, the signal must be transduced into a biologic response.

As mentioned, it is possible to modulate each of these processes using chemical tools and to affect the whole signaling loop. Sometimes this modulation can be intentional, such as the use of aromatase inhibitors to halt the production of estrogens and block estrogen signaling (18), or it can be unintentional, as is often the case with hormone receptor antagonists also blocking negative feedback inhibition and causing overproduction of the natural hormone (19). As such, it is important for any chemist who wishes to apply chemical tools to the study of hormone signaling in whole organism models, as in the case of new therapeutics, to consider hormone signaling as a system with many different components.

Classes of hormones and hormone receptors

Generally, hormones can be organized into three families: the peptide hormones, the steroid hormones, and a very loosely organized class of hormones derived from the modification of amino acids and lipids (see Fig. 2). The structure of the hormone dictates in many ways the nature of the hormone receptor. The peptide hormones can vary widely in length and amino acid composition, but they are generally too hydrophilic to cross the plasma membrane. As a result, most hormone receptors for peptide hormones are membrane receptors such as G-protein coupled receptors (GPCRs), receptor tyrosine kinases, and ion channels. As will be discussed, a key problem in the area of peptide hormone research is how to improve the physical properties of therapeutics by targeting peptide hormone receptors using peptidomimetics. The steroid hormone family is sufficiently hydrophobic to cross the membrane receptors, so its
Hormone Signaling

Figure 2
Examples of the different basic structural classes of hormones.

Receptors tend to be intracellular nuclear receptors, although extracellular receptors for steroid hormones have been proposed (20). The third family of hormones covers all remaining small molecules and, as such, targets several different extracellular receptors or intracellular receptors depending on the particular hormone. One class of hormones synthesized from amino acids includes biogenic amines such as dopamine, serotonin, histamine, and epinephrine. Although most of these biogenic amines are more closely identified as neurotransmitters, the biogenic amines are present in physiologically relevant concentrations in circulation and have receptors located in several tissues in the cardiovascular, digestive, and immune systems (21–24). Molecules well known to chemical biologists for other roles, such as adenosine and ATP, also are known to have hormone signaling functions (25). Also several hormone receptors exist for which no hormone has ever been discovered. These so-called “orphan” receptors are numerous, and discovering ligands to match these orphan receptors is one important area of hormone research that requires chemical tools.

Chemical Applications in the Study of Hormone Signaling

As stated, major leaps forward in the study of hormone signaling have often coincided with the development of new chemical tools. As such, an exhaustive history of these tools is far beyond the scope of this article. Instead, the article will focus on a few studies in which newly developed chemical tools have had an important impact on the understanding of different hormone signaling systems. The types of approaches used in these studies can be applied to several different types of hormone signaling systems and hopefully will serve as examples of the potential of studying hormone signaling at the chemistry-biology interface.

Discovering new ligands for orphan hormone receptors

One of the consequences of the Human Genome Project has been the identification of genes that encode proteins with no known function. Some of these proteins share sequence homology with known hormone receptors, which strongly suggests that these new receptors should have ligand partners. These “orphan” receptors exist in every major family subtype but they seem to be especially common in the GPCR and nuclear receptor families (26, 27). Although it is likely that some of these hormone receptors will not have a ligand, undoubtedly other orphan receptors exist that do have a hormone that has not yet been discovered. Finding methods to match ligands to orphan receptors would have a huge impact in discovering new drug targets and remains a major challenge in the field of molecular endocrinology. Several approaches are currently being used to attempt to accomplish this task. These methods include in vitro screening of large libraries of compounds, transgenic mouse models, structure-based screening methods, and bioinformatics; they have been applied most often to orphan members of the GPCR and nuclear receptor families (27, 28). Although some successes have been reported, it is important to remember that any potential ligand found during in vitro studies must also be...
Hormone Signaling

Present at sufficient concentrations in vivo for a sufficiently long enough period of time to qualify as an endogenous ligand.

Another strategy to discovering ligand matches for orphan receptors is to start with metabolites of known hormones, assuming that only a limited number of hormone structures exists that an organism can synthesize. This approach was used with great success in finding hormone ligands for the orphan nuclear receptors liver X receptor (LXR) and farnesoid X receptor (FXR). Based on expression patterns of the receptor in different tissues, Manglesdorf et al. hypothesized that LXR and FXR played some sort of role in cholesterol and bile acid metabolism. Through a combination of tissue extracts and screening and organic synthesis, they discovered that LXR could be regulated by sterols such as 24(S)-hydroxycholesterol and that FXR could be modulated by bile acids such as chenodeoxycholic acid at physiologically relevant concentrations (29, 30). The matching of ligands with these two receptors has had a huge impact in the field of lipid metabolism and has led to the development of potentially promising therapeutic candidates.

Another example of using hormone metabolites to discover new hormone signaling pathways is the case of a metabolite of thyroid hormone. Scanlan et al. noted the similarities between 3,3',5-triiodothyronine (T3), the highest affinity endogenous ligand for the thyroid hormone receptor, and various biogenic phenethylamines such as dopamine if T3 were to be enzymatically deacetylated to a thyronamine (see Fig. 3). Several different thyronamines were synthesized with different degrees of iodination and screened against GPCRs thought to bind to biogenic amines. One of these compounds, 3-iodothyronamine, was found with high affinity to an isoform of the trace amine receptor (TAAR1), an orphan GPCR with no previously identified endogenous ligand (31). Additional biologic characterization of this interaction showed that the ligand was found endogenously in rat and guinea pig brain and that when the compound was administered to mice it caused rapid slowing of heartbeat and an almost 8-degree drop in body temperature. The effects were reversible over time with no deleterious long-term effects on the mice. Interestingly, the effects of 3-iodothyronamine on the organism were opposite to those of T3. The thyroid hormone receptor is a member of the nuclear receptor superfamily and exerts most of its effects at the transcriptional level where it causes an increase in body temperature and heart rate. It seems like T3 and its iodothyronamine metabolite act in concert to maintain a balance in homeostasis as it relates to body temperature and heart rate—two of the most fundamental processes of an organism. Higher potency analogs have also been synthesized to explore this whole new area of hormone signaling research that was made possible by new chemical tools (32).

Non-natural hormone mimetics

In contrast to the case of orphan receptors without an endogenous ligand, many hormone receptors have well-characterized endogenous ligands. Even with well-characterized hormones, however, a strong need remains for synthetic analogs for these hormones that have different structures than their endogenous counterparts. Many reasons exist for needing synthetic analogs, including the possibility of designing a hormone antagonist that can be used as a tool to block hormone signaling. A rather reason that is especially relevant to peptide hormones is the need to change the physical properties of the hormone. Although several peptide hormones are used therapeutically, such as insulin and oxytocin, peptides generally are not orally bioavailable and usually are administered parenterally (33). The ideal hormone mimic would be orally available and would show increased half-life in the circulation.

Most successes in designing non-natural hormone mimics have come from nonpeptide hormone families. Nonsteroidal ligands exist for almost all steroid hormone receptors, and many are used therapeutically (34, 35). The same is true for hormones based on amino acids and lipids. The one hormone class where mimicry has been difficult to achieve is the peptide hormone class. Several strategies have been employed using peptide scaffolds to alter the physical properties of peptide hormones, including truncation, cyclization, and substitution with non-natural amino acids (36). These sorts of strategies have been greatly aided by the development of rapid peptide synthesis and screening techniques such as phage display (37). Many attempts have been made to create nonpeptide mimics of peptide hormones. It is a well-known problem that protein–protein interfaces are difficult interactions to mimic or block with a small molecule (38), but some notable successes exist in the field of hormone signaling (39, 40). A small molecule capable of mimicking the biologic activity of granulocyte colony stimulating factor (G-CSF) can induce the oligomerization of the hormone receptor in a manner similar to the endogenous peptide hormone (41). Also, small-molecule mimics have been discovered for the insulin receptor (42), fibrinolysis growth factor (43), and interleukin 2 (44). Although these mimics all seem to possess the functional equivalence of the endogenous hormones, it is still possible that these small molecules achieve their biologic effects by mechanisms other than direct binding to the hormone-binding site on the receptor.

Synthetic hormone signaling systems

In some cases, a particular application may call for some sort of hormone-regulated control that is not present, as in the case of selective transcriptional modulation of a specific transgene or signal transduction event. Several research groups have developed approaches to this problem by engineering a synthetic hormone signaling system that is orthogonal to all endogenous hormone pathways. If an effector is also placed under the regulation of this orthogonal receptor, then the response can be modulated by the addition of the matching ligand. The approaches have varied from using hormone receptors from other species to engineering receptors and ligands via a "bump-hole" approach. These approaches have been used to generate orthogonal versions of estrogen and the thyroid hormone receptor and can be used to explore the functions of a specific receptor (45–48). These synthetic hormone signaling systems have also been used to regulate transcription of specific transgenes as well as to control other hormone receptor families such as GPCRs, as was the case with the adenosine A2A receptor (49).
Selective hormone receptor modulation

Sometimes, simple mimicry of the hormone is not enough. One major reason that hormone signaling has been such a fruitful area of research is that many of these hormones have multiple effects. Endocrine hormones are systemically circulated, so it is not surprising that they can modulate receptors in many different tissues. Although this finding means that a hormone can have many different functions worthy of study, it also means that looking at the role or mechanism of one specific response is more difficult because of possible interference from other signaling events elicited by that hormone. This interference can have significant therapeutic ramifications if the other signaling events cause deleterious side effects. Several examples exist in endocrinology where selective modulation of hormone signaling is desired and where one of the most effective ways to achieve selective modulation is to use a selective ligand. The term “selective” can have different meanings when it comes to hormone signaling. This section will look at two different examples of selectivity: receptor isoform selectivity and “response” selectivity.

One key feature of many hormone signaling systems is that often multiple receptors can bind the same hormone. The same hormone can sometimes bind totally different receptors, as in the case with estradiol binding to the estrogen receptor, a member of the nuclear receptor superfamily, and to GPR30, a GPCR (50). Typically, however, the same hormone binds to variants of a single hormone receptor, either isoforms created by alternative splicing or subtypes actually encoded by different genes. Even though these variants may bind to the same hormone and have similar structural features, they can have significantly different functions. They may be expressed in different tissues or expressed at different times in the development of an organism and thus can regulate significantly different signal transduction pathways from other variants of the same receptor. To determine the function of individual variants in a biologic system, it is necessary to uniquely activate the variant of interest. Although genetic approaches that use knockout animal models have been necessary to uniquely activate the variant of interest, it is often multiple receptors can bind the same hormone, either isoforms created by alternative splicing or subtypes actually encoded by different genes.

Numerous examples exist where variant-selective ligands have been used to dissect complex hormone signaling pathways, but one of the best examples of the use of selective ligands to uncover the function of very disparate receptor variants is the case of the melanocortin receptor. The melanocortin receptor family is a group of five receptor subtypes that belong to the GPCR superfamily and bind to several similar peptide hormones, α, and γ-melanocyte stimulating hormones (MSHs) (51). The receptors are expressed at different amounts in a wide variety of tissues and seem to have roles in obesity, inflammation, and cardiovascular function, as well as the more expected role of controlling skin pigmentation. To better understand what specific receptor subtypes are doing, as well as to explore the potential of using melanocortin receptors as drug targets, several attempts have been made to design and synthesize selective receptor modulators. The different receptor subtypes have different binding preferences for the various hormones, which suggests that it might be possible to differentiate the receptor subtypes on the basis of ligand-binding affinity. Several selective ligands have been reported, although few of them are specific for just one subtype. Enough selectivity has been achieved, however, to start to understand the roles that some receptor subtypes are playing in the melanocortin hormone signaling system.

Because the melanocortin receptor subtypes are peptide hormone receptors, great effort has gone into using various peptide synthesis and screening methods to uncover selective ligands (52). This work has led to the development of several compounds with different patterns of selectivity toward the different receptor subtypes. One of the first reported selective compounds, a peptide termed MTII, showed selectivity for both the MC3R and the MC4R subtypes (53); it was an anti-inflammatory agent in a rodent model and could block overeating in an animal obesity model, which suggests that the peptide may play a role in both energy homeostasis and inflammation (54, 55). In addition, another peptide, labeled SHU9119, acted as an antagonist at the MC1R and MC5R subtypes, and was able to block the anti-inflammatory and anti-obesity effects of MTII. Selective nonpeptide ligands have also been developed for the MC1R subtype, which prevented inflammation in acute inflammatory mouse models (56), as well as for the MC4R subtype, which seemed to affect sexual function (see Fig. 4) (57). Although the various compounds are not quite selective enough to precisely determine what each melanocortin receptor subtype is doing, the development of selective chemical tools has greatly increased understanding of this complex hormone signaling network.

In some cases, it seems that different responses to a given hormone can come from just one receptor subtype. In such cases, the hormone is binding to the same receptor variant, but the response is different depending on the cell or tissue context. This has been reported for several hormone receptors, but the classic example is the estrogen receptor (ER) (58). Estrogens have a variety of responses in different tissues ranging...
developed and can help dissect this challenging problem. Selective estrogen receptor modulators (SERMs), have been chemical tools that show patterns of selective modulation, called responses depending on the cell context. Fortunately, several to the same compound but have totally different signaling challenges is to understand how the receptor can bind identically from binding to a single receptor subtype—ER alpha (59). The ER subtype ER beta, several tissue-dependent responses result from binding to a single receptor subtypes—ER alpha (59). The challenge is to understand how the receptor can bind identically to the same compound but have totally different signaling responses depending on the cell context. Fortunately, several chemical tools that show patterns of selective modulation, called selective estrogen receptor modulators (SERMs), have been developed and can help dissect this challenging problem. Some of the most widely used SERMs have already been clinically validated: tamoxifen, the most widely used drug to treat and prevent hormone-responsive breast cancer, and raloxifene, used for the prevention of osteoporosis and being considered as a breast cancer preventive (60). Both compounds act as anti-estrogens in the breast and block estrogen-induced proliferation but act as estrogens in the bone where they prevent osteoporosis. In the uterus, they have different activities: tamoxifen is estrogenic and induces proliferation in the uterus, whereas raloxifene blocks proliferation (61). Other compounds have different patterns of responses in different tissues, and all seem to be able to bind to the estrogen receptor with high affinity. So key questions remain: How can one receptor have so many different responses that are dependent on cell context? How can it be controlled? The combination of chemical tools and structural biology are beginning to provide the answers to these important questions. When comparing mechanisms to explain estrogenic responses versus anti-estrogenic responses, the structures of the ligand-binding domain of the estrogen receptor bound to estradiol, tamoxifen, or raloxifene were compared (see Fig. 5) (62). The SERMs tamoxifen and raloxifene cause a major perturbation in one alpha helix of the domain. The cleft created by that helix when an estrogen binds is recognized by a coactivating protein that allows further buildup of a complex that activates transcription at a particular promoter. The cleft is obstructed when tamoxifen or raloxifene binds to the ligand domain; this binding leads to an interaction with a copressor protein that, in turn, promotes the buildup of a complex that represses transcription. This mode of antagonism has become a common feature in drug design for nuclear receptors (10).

Although this structural comparison only reveals possible mechanisms by which estradiol can differ from tamoxifen and raloxifene, it also suggests that a single receptor can have several different biologic responses caused by different downstream effectors that interact with different parts of the receptor. Differential expression of those effectors can then dictate different responses to the same drug in different cells and tissues. Even though the tamoxifen and raloxifene structures are similar, the crystal structures contain only one domain of a three-domain receptor, the other domains could be playing a role in transducing slight changes in conformation of the ligand-binding domain into more significant changes in the overall receptor conformation. It is also known that the estrogen receptor modulates many different pathways, including the activity of different transcription factors (63). Some success has been reported at developing inhibitors to block interactions between ER and downstream effectors (64). Finding more chemical approaches like these to identify ligand-selective effectors as well as developing new chemical tools that can find new downstream estrogen signaling pathways are key areas where chemical biology will play a key role in untangling a complicated and therapeutically important hormone signaling network.

Chemical Tools and Techniques

For a field as broad as hormone signaling, the key chemical tools and techniques used in the field could include almost anything. The important thing to remember about studying the chemical biology of hormone signaling is that relying on only in vitro data usually will give you only a small part of the story. Ultimately, cell-based and preferably animal-based studies should be conducted to test any tools that might be developed. In general, several techniques are shared by most hormone signaling studies, but each hormone and hormone receptor will have unique assays. Very brief descriptions of these techniques will follow. Most of the assays described are very generalized and not described by single reference sources. The most useful source for many of these assays will be medicinal chemistry papers that describe a hormone receptor modulating compound from synthesis to animal testing, but another good source for protocols are the retailers of commercially available assay kits and can be easily found online.

Receptor-ligand binding assays

In almost all cases hormone signaling involves the binding of receptors and ligands, and the development of any chemical tool will need to include some assay to measure binding affinity. Typically, in vitro binding assays with either radiola beled or fluorescently labeled ligand and purified receptor are used, although sometimes membrane extracts have to be used for membrane-localized receptors (65–68). Several of these assays have been developed, are available commercially, and are amenable to high-throughput screening. It is important to calculate binding affinity with dose response curves using nonlinear regression analysis whenever possible. Several affordable software packages are available to do this sort of statistical analysis.
Hormone Signaling

Figure 5 A representation of the estrogen receptor alpha ligand-binding domain bound to estradiol, 4-hydroxytamoxifen, or raloxifene. Key helix 12 is highlighted in white. The hatched area indicates the coactivator-binding cleft formed upon estradiol binding, which is blocked by helix 12 upon 4-hydroxytamoxifen or raloxifene binding.

One caveat to these types of binding assays is that receptor binding affinity in vitro does not always correlate well to binding affinity in cells.

Reporter assays

Most hormone signaling studies will also need some assay to measure the cellular effects of the ligand being developed. These studies usually are performed in cultured cells and can range from transcriptional reporter assays using luciferase reporter plasmids (69) to enzymatic assays for the activation of various kinases and cyclases (70, 71) to cellular sensor assays looking for changes in the concentration of markers such as intracellular calcium (72). In these assays the cells are typically dosed with a compound for a certain period of time in multwell culture plates, then the cells are lysed, and the reporter is measured using some sort of spectrophotometric technique. Also whole-cell and whole-animal imaging techniques are beginning to be used to perform more complex reporter assays (73).

Mutagenic scanning methodologies

One key question that should be addressed when searching for a new hormone receptor or hormone receptor subtype is as follows: Which amino acids are involved in hormone binding? Homology modeling with similar receptors has been extremely useful in identifying potential ligand binding sites in orphan nuclear receptors, but one of the most useful techniques for peptide hormone receptors is scanning mutagenesis. In this technique, amino acids in a protein can be systematically replaced with other amino acids, typically alanine, and receptor binding can be assessed. The mutagenesis can be performed using several different approaches, but some sort of high-throughput screening of receptor binding or activity must be used in order to achieve thorough coverage (74). Mutagenesis has been used extensively to study the interaction between the human growth hormone and its receptor (75, 76). It has also been used with peptide hormones such as insulin and vasopressin (77, 78), as well as nonpeptide hormone receptors such as the vitamin D receptor (79).

Proteomics, genomics, and systems biology

Proteomic and genomic technology are other tools derived by chemists that are having a dramatic effect on the study of hormone signaling. Many questions regarding downstream effectors of a specific hormone receptor in a specific cell type can be answered using proteomic techniques to identify uniquely expressed proteins in specific tissues (80). If the hormone receptor modulates transcription, genomics can be very valuable in determining the subsets of genes affected by the receptor in a given tissue (81). In addition, the emerging field of systems biology should make significant contributions to the molecular study of hormone signaling at the organism level (82).

Major Challenges and Future Direction

Hormone signaling was one of the original fields of study that required working at the chemistry–biology interface, and it will continue to need people trained in multidisciplinary work to unravel many big questions facing the field. The topics described...
above—matching ligands to orphan receptors, finding unnatural mimics of hormones, and improving and understanding selective receptor modulation—will continue to remain major challenges in the field and to require the development of new chemical tools. The other major challenges are the development of real-time techniques for detecting the biological responses to hormones and hormone mimics, as well as the development of compounds and techniques to dissect the roles of multiple hormone signaling systems in a particular physiological response. It is becoming abundantly clear that most hormone receptors crosstalk with other receptors from the same hormone family as well as with other hormone families. For instance, the crosstalk between growth factor receptor signaling and estrogen receptor signaling is believed to play a major role in the development of several kinds of antiestrogen resistance in breast cancer (83). Major health problems facing western populations, such as obesity, are also endocrine disorders and will involve several different hormone receptor systems (84). These tough problems require people trained in the chemical sciences to develop the tools necessary to solve them. If the past is any guide, it is likely that a new chemical tool will be behind every great push forward in our understanding of the molecular mechanisms of hormone signaling.

References

10. Grönemeyer H, Gustafsson JA, Laudet V. Principles for modula-
14. Kahn S, Hull R, Utschneider K.M. Mechanism linking obe-
15. Williams DA, Lenke TL. Foye’s Principles of Medicinal Chem-
istry. 5th edition. 2002. Lippincott Williams & Wilkins, Philadel-
phia, PA.
20. Welling M, Losel R. Non-genomic steroid hormone effects: mem-
3172–181.
29. Janowski BA, Willy PJ, Doni TR, Falck JR, Magelssdorf DJ. An oxy-
35. Winneker RC, Fensome A, Wrobel JE, Zhang Z, Zhang P. Nons-
teroidal progestosterone receptor modulators: structure activity rela-


Hormone Signaling
Hormone Signaling


Further Reading

Here are a few textbooks that cover a number of the specific hormone signaling systems and protocols for both measuring receptor-ligand interactions and cell-based assays.


See Also

Cellular Communication Through Signal Transduction, Chemistry of Nuclear Receptors
Receptor-Ligand Interactions
Peptidomimetics
Small Molecules to Elucidate GPCR Signaling Pathways
Inositol Phospholipids, Biosynthesis and Biological Functions of

Karol S. Bruzik, Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, Chicago, Illinois
doi: 10.1002/9780470048672.wecb447

Phosphatidylinositol phosphates are cellular signaling molecules that came to prominence in the mid-1980s, after inositol trisphosphate was discovered as a key intracellular second messenger generated from phosphatidylinositol 4,5-bisphosphate in response to stimulation of a wide range of extra-cellular receptors. Since then, many novel inositol phosphates and phospholipids have been discovered, and their cellular functions have been studied. The roles of these molecules encompass regulation of many processes indispensable to organism homeostasis. On the cellular level, the processes include regulation of cell growth and differentiation, cell motility and invasiveness, vesicular trafficking, protein targeting to specific cell compartments. This review focuses on biosynthesis and biological roles of phosphatidylinositol phospholipids.

One of the most significant discoveries of the last two decades was the finding that multiple cellular signal transduction pathways are mediated by inositol phospholipids (PIPs) and that inositol phosphates are intracellular second messengers.1 Because of the sheer complexity, studies of the signaling function of these molecules remain an important field of biomedical research. Once deemed obscure minor components of biological membranes, inositol phospholipids are now known to be involved in the transduction of the vast array of extracellular signals, which include neurotransmitters, growth hormones, and photons (1, 2). Because of the broad scope of these signaling events, several thousand research publications related to inositol phospholipids are published every year (3). Much of this activity is enabled by the availability of phosphoinositides because of the success in developing synthetic methodologies that lead to these compounds during the last 20 years (4, 5).

Structure of Inositol Phospholipids

Overall, nine major inositol phospholipids (1-9) (Fig. 1) have been identified in cellular signaling pathways, and more are likely to be discovered. Of two possible enantiomers of the unsymmetrically substituted inositol residue, only the molecules in which the phosphatidate residue is attached to the D-1 position of myo-inositol are active physiologically (6, 7). Most naturally occurring PIPs that participate in signal transduction pathways in mammalian cells feature inositol ring phosphorylated at the 3-, 4-, and/or 5-positions. The addition or removal of the phosphate residue at the inositol ring is a simple biochemical process, yet given the large number of possible hydroxylation sites, it generates molecules that are characterized by a different net charge and feature distinct geometric distribution of the individual negative charges. In addition, the net negative charge of these molecules can be modulated easily by slight differences in pH of a local environment. Rapid generation of such a molecular array and facile interconversions between its members may be the basis for the omnipresence of inositol phospholipids in the signaling phenomena. The molecules that are nonphosphorylated at any of the 3-, 4- and 5-positions can also be glycosylated at the 6-position (GPI) (8). These phosphatidylinositols serve a multitude of other functions such as protein and protozoan VSG anchoring (9), and serving as precursors for the putative insulin mediators (10).

Regarding the lipid composition of phosphatidylinositols, the mammalian PIPs carry stearic and arachidonic acid esters at the glycerol sn-1 and sn-2-positions, respectively (11). The composition of inositol phospholipids isolated from plants varies from those present in mammalian tissues because they contain linoleoyl residue at the sn-2-position and palmitoyl residue at the sn-1-position.

As of 05/2007, the Chemical Abstract Service database included ca. 140,000 entries containing the "inositol" keyword.
In principle, only three members of the PIPn family, including phosphatidylinositol 4-phosphate ([PI(4)P], and 4,5-bisphosphate ([PI(4,5)P2]), are available in sufficient quantities by isolation from natural sources. Other phosphoinositides, although they play important roles in cell physiology, are formed transiently or are present in biological sources at very low concentrations, which makes preparative isolation of these compounds unfeasible. To date, all known naturally occurring phosphoinositides, with the exception of putative prostaglandyl inositol phosphate ([PI(3,4,5)P3]), have been synthesized, with some methodology published from the Department of Medicinal Chemistry and Pharmacognosy laboratory at the University of Illinois (14). Because of the synthetic expedience, most available synthetic inositol phospholipids are supplied as unnatural analogs that contain the saturated fatty acid chains, with either a long (dipalmityl) or shorter chains (dioctanoyl or diohexanoyl). These saturated analogs are not only easier to synthesize, but also are more stable chemically because they do not undergo peroxidation reactions common to unsaturated fatty acids. In addition, synthetic methods have been developed toward phosphatidylinositol analogs that carry fluorescent, biotin, and photoaffinity residues to identify the various phosphoinositide binding proteins and to study spatiotemporal aspects of inositol signaling pathways (15, 16).

Principal Biosynthetic and Metabolic Pathways

In plants, inositol is biosynthesized from glucose 1-phosphate via inosose 1-phosphate, which is reduced into 1D-inositol 3-phosphate (INO1 pathway). After dephosphorylation by inositol phosphate monophosphatase, myo-inositol is fed into biosynthesis of inositol phospholipids via a Kennedy-like sequence. In mammalian organisms, inositol is acquired from the diet and is also biosynthesized via INO1. The latter pathway, rather than the inositol phosphate phosphatase, seems to be a more promising target of antipsychotic drugs (17, 18). This compound is biosynthesized from myo-inositol and CDP-diacylglycerol by phosphatidylinositol synthase localized in the endoplasmic reticulum (19). The PI metabolism involves sequential phosphorylations at the 3 (20), 4- (21), and 5-positions (22) to provide compounds (2-8). In addition, PI is subject to degradation by phospholipases C and phospholipases A2. Although most mammalian phospholipases C prefer PI-4,5-P2 and PI-4-P (23), isolated reports of PI-specific phospholipases in mammalian tissues exist (24). In contrast to mammalian PI-PLC, the bacterial enzymes display strict specificity for unphosphorylated PI (25). Generally, PI, that bear a phosphate group at the inositol 3-position are regarded as resistant to cleavage by all PLC isozymes (26). PI seems also refractive to hydrolysis by two major isozymes of mammalian PLOD (27). PI is also subject to the decarboxylation reaction by sequential action of phospholipase A and lysophospholipase that generate glycerophosphoinositols (28), which was recently shown to inhibit invasion of cancer cells (29).

Phosphatidylinositol 3-phosphate

Phosphatidylinositol 3-phosphate ([PI(3)P]) is the second most abundant inositol phospholipid in biological membranes. It is a product of phosphorylation by four different species of PI 4-kinases (21), and it is a precursor to PI-4,5-P2 and PI-3,4,5-P3. These kinases are the two inositol lipids whose biological roles are understood best. The biological function of PI-4-P is not well known; however, the studies in yeast indicate its role extends beyond the substrate for PI 5-kinase (21), such as regulation of vesicular trafficking and protein secretion from Golgi. PI-4-P is removed by the subsequent phosphorylation to PI-4,5-P2 or dephosphorylation to PI by ER-localized phosphatase (21).

Phosphatidylinositol 5-phosphate

Phosphatidylinositol 5-phosphate ([PI(5)P]) is formed by the hydrolysis of 3-phosphate from PI-3,5-P2 by myotubulin,

Figure 1 Structures of PI.

For stereospecific numbering of the glycerol moiety, see Reference 6. For stereospecific numbering of myo-inositol residue, see References (5) and (7).

Inositol Phospholipids: Biosynthesis and Biological Functions of...
Figure 2  Principal metabolic pathways of inositol phospholipids.
M TM 1 (33). This molecule can be removed either by phospho-
rylation into PI-4,5-P2 by phosphorylation with the ρ-form of
PI-4K (34) or by dephosphorylation of PI-5-P by the novel form
of Pten-like proteins with selectivity toward phosphatidyli-
nositol phosphates at the 5-position (35). Recent studies
suggest a role for PI-5-P in a variety of cellular events, such
as tumor suppression, response to bacterial invasion (35), and
control of osmotic pressure (36). This phosphophosphid has been
shown recently to function as a second messenger that binds
to an Arabidopsis homolog of trithorax, which suggests that it
may have a regulatory function that connects lipid signaling
with nuclear functions (37).

Phosphatidylinositol 3,5-bisphosphate

Phosphatidylinositol 3,5-bisphosphate (PI-3,5-P2, 6) is the low
abundance, newest member of PIPn family (33). It is involved
in mediation of several cellular processes such as vascular
homeostasis, membrane trafficking, and vesicular protein sorting
(36, 38). The recently discovered PI-3,5-P2 effectors include a
family of β-propeller, epsin, and CHMP protein families (39).
The importance of PI-3,5-P2 in human physiology is demon-
strated by its role in insulin signaling, myotubular myopathy,
and corneal dystrophy (38).

Phosphatidylinositol 3,4-bisphosphate

Phosphatidylinositol 3,4-bisphosphate (PI-3,4-P2, 5) is biosyn-
thesized by phosphorylation of PI-4-P at the 3-position (26), by
dephosphorylation of PI-3,4,5-P3 at the 5-position by the SHIP
phosphatase (39, 40), and by phosphorylation of PI-3-P by the
Type II 4-kinase (41). The significance of this lipid is under-
standing the association between SHIP2 gene polymorphism
and type 2 diabetes mellitus (39). Therefore, SHIP2 constitutes
an important target for treatment of both type 2 diabetes and
obesity (40).

Phosphatidylinositol 4,5-bisphosphate

Phosphatidylinositol 4,5-bisphosphate (PI-4,5-P2, 7) is the third
most abundant inositol phospholipid in biological membranes.
The predominant biosynthetic pathway is via phosphoryla-
tion of PI-4-P by a 5-kinase (3h). This inositol lipid is best
known for participating in receptor-mediated cleavage by mam-
alian phospholipases C-β, γ, δ, and ε to produce inositol
1,4,5-trisphosphate second messenger (41, 42), and for initiating
extremely complex metabolic pathways of inositol phosphates,
which include formation of several new second messengers,
such as inositol 1,3,4,5-tetrakisphosphate, 1,3,4-trisphosphate,
and 3,4,5,6-tetrakisphosphate (3e, 43). The biological role
of PI-4,5-P2 is much larger because it is a substrate for Type
I 3-kinases that ultimately provide PI-3,4,5-P3 as an-
other important signaling molecule (44). In addition, PI-4,5-P2
is recognized specifically by pleckstrin homology domains
of many important proteins, which include several enzymes
(e.g., phospholipase D (45)), which provides an interesting
crosslink between the PLC and PLD lipolytic activities. A nother
example of the interrelationship between phospholipases is the
regulation of Ca2+ by PI-4,5-P2 (46). Furthermore, associa-
tion of certain transmembrane receptors with PI-4,5-P2 affects
their functional activity (e.g., interaction of vanilloid receptor
with PIP2 is necessary for receptor desensitization (47)). There-
fore, PI-4,5-P2 plays an important role in both human health
and disease (48).

Phosphatidylinositol 3,4,5-trisphosphate

Phosphatidylinositol 3,4,5-trisphosphate (PI-3,4,5-P3, 8) is the
most phosphorylated inositol phospholipid, and it is one whose
function is related to cell growth, survival, and differentiation
(49, 50). The proper physiological levels of PI-3,4,5-P3 are
maintained in a major part by phosphorylation of PI-4,5-P2
by the Type I phosphatidylinositol 3-kinases and by hydrolysis
of two 3-phosphoinositide phosphatases: PTEN and SHIP2 (20).
The latter activities lead to the regeneration of PI-4,5-P2 and to
formation of another second messenger, PI-3,4,5-P3, respectivelv
(39, 40). Strong evidence suggests that PI-3,4,5-P3 is an impor-
tant player of signaling pathways in the nucleus. Recent results
also indicate that nuclear translocation of cell surface receptors
could activate nuclear PI-3K, which suggests a new pathway
of signal transduction (51). The most important cellular func-
tion of PI-3P at the moment seems to be its strong interaction
with the PH domain of protein kinase A(κ). Binding of PIP3
causes translocation of Akt to the plasma membrane where it
becomes phosphorylated and activated (50). The activated Akt
then phosphophates downstream cellular proteins that promote
cell proliferation and survival, which results ultimately in tu-
morigenesis. Therefore, the proper concentration of PI-3,4,5-P3
is pivotal to cell homeostasis, and its elevated levels promote
tumor formation (50).

Phosphoinositide-Binding Protein Domains

A major advance in understanding signaling roles of phos-
phatidylinositols has been the discovery of several highly con-
served protein domains whose function is to bind head groups of
specific phosphatidylinositol phosphates. Such “cut and paste”
modules are attached to a diverse array of multidomain pro-
teins, and they enable recruitment of such proteins to spe-
cific regions in cells via binding to plasma or to intracellular
membranes enriched with such phosphoinositides (52, 53).
As a result, phosphoinositides can act as signal mediators in a
spatially- and temporally-defined manner, and they can control
various intracellular events such as cytoskeletal rearrangement
and membrane trafficking. The pleckstrin homology (PH) do-
main was the first identified phosphoinositide-binding domain.
It contains the largest number of members and is associated with
the formation of signaling complexes on the plasma membrane.
Recent studies identified other novel phosphoinositide-binding
domains such as FYVE, Phox homology (PX), and epsin
N-terminal homology (ENTH), which attest even more to the
functional versatility of phosphoinositides (54). The number of
known domains or modules that bind phosphoinositides has
increased dramatically over the past few years. Struc-
tural analysis of interactions of inositol phospholipids with
phosphoinositide-binding domains has provided significant insight into the mechanism of membrane recruitment by the different cellular phosphoinositides. Thus, the domains that target only the rare (3-phosphorylated) phosphoinositides must bind with high affinity and specificity. In the case of certain PH domains (which bind PI-3,4,5-P3 and/or PI-3,4-P2), this binding is achieved exclusively by headgroup interactions. In contrast, the PI-3P-targeting PX and FYVE domains require the same stringent affinity and specificity and tend to be more diverse in structure. The mode of phosphoinositide binding by different domains also seems to reflect their distinct functions. For example, PX domains that serve as simple targeting domains recognize only the phosphoinositide headgroup. By contrast, certain other domains, notably the epsin targeting domains recognize only the phosphoinositide headgroup interactions. In contrast, the phosphoinositide binding domains is listed in Table 1.

Table 1 Recognition of inositol phospholipids by specific phosphoinositide-binding domains

<table>
<thead>
<tr>
<th>Phosphoinositide-binding domain</th>
<th>PI-3P</th>
<th>PI-4P</th>
<th>PI-5P</th>
<th>PI-3,4-P2</th>
<th>PI-3,5-P2</th>
<th>PI-4,5-P2</th>
<th>PI-3,4,5-P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FYVE</td>
<td>PH</td>
<td>PH</td>
<td>PH</td>
<td>PH</td>
<td>PH</td>
<td>PH</td>
<td>PH</td>
</tr>
<tr>
<td>PHX</td>
<td>E/ANTH</td>
<td>PH</td>
<td>E/ANTH</td>
<td>PH</td>
<td>E/ANTH</td>
<td>PH</td>
<td>E/ANTH</td>
</tr>
</tbody>
</table>

In summary, transmembrane and intracellular signaling with phosphoinositides is a dynamically expanding field. The progress in this area is critical to understanding the intricate spatiotemporal effects of generation of these small molecular mediators in response to extracellular signals on the functional roles of intracellular proteins.

References

3. As of 05/2007 Chemical Abstract Service database contained ca. 140,000 entries containing the “inositol” keyword.

Inositol Phospholipids, Biosynthesis and Biological Functions of...
Inositol Phospholipids: Biosynthesis and Biological Functions

Inositol Polyphosphates

Adam C. Resnick, Division of Neurosurgery at the Children’s Hospital of Philadelphia, Department of Neurosurgery at the University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania
Adolfo Saiardi, Medical Research Council (MRC) Cell Biology Unit and Laboratory for Molecular Cell Biology, Department of Biochemistry and Molecular Biology, University College London, London, United Kingdom
doi: 10.1002/9780470048672.wecb252

Inositol polyphosphates comprise a large family of water-soluble molecules derived from the combinatorial phosphorylation of the six hydroxyls of myo-inositol. Second messenger roles for inositol polyphosphates in Ca2+ mobilization were first identified for what is now the best characterized family member, inositol 1,4,5-trisphosphate (InsP3). Additional anabolic and catabolic metabolism of InsP3 results in the formation of a large, diverse family of higher inositol polyphosphates whose signaling roles and biologic functionality remain largely undefined. However, the recent cloning and identification of the kinases and phosphatases involved in the combinatorial modification of inositol polyphosphates has served to further define and characterize this complex metabolic network and to identify its preeminence in nearly all aspects of cell biology. Conserved from yeast to humans, inositol polyphosphates regulate a wide array of processes, including ion-channel conductance, membrane dynamics, transcription, nucleic acid metabolism, and protein phosphorylation.

The scientific study of inositol polyphosphates began in 1850 with the isolation of a crystalline “sugar” from heart muscle extracts (1). The substance was named “inosit” from the Greek root inos for muscle, fiber, or sinew. The later identification of this optically inactive inositol (the “ol” suffix added in French and English) as one of nine possible cyclohexanehexol isomers necessitated the use of prefix designations for the specific stereoisomeric configurations of the secondary hydroxyl groups about the six-carbon ring (cis-, epi-, allo-, muco-, neo-, scyllo- (+)-chiro,-(-)-chiro, and myo-) (Fig. 1). Although several, if not all, inositol isomers occur in nature, myo-inositol (myo, once again derived from the Greek word for muscle) is the biologically relevant stereoisomer in most species and functions as the structural building block for the inositol polyphosphates. Viewed in its favored chair conformation, myo-inositol possesses five equatorial groups and one axial hydroxyl group. The modern D-numbering system for inositol is by convention counterclockwise as viewed from above and assigns the axial hydroxyl group to C2. This conformation is best illustrated by Agranoff’s turtle (2) in which the six hydroxyls are envisaged as the appendages (1-, 3-, 4-, 6-hydroxyl), head (2-hydroxyl), and tail (5-hydroxyl) of a friendly turtle. It is worth noting that among the stereoisomers of inositol, myo-inositol is unique in containing a single axial OH group (Fig. 1). This results in an achiral molecule with a plane of symmetry through C2–C5 (head to tail) and two pairs of enantiomorphic hydroxyls (C1–C3 and C4–C6, the turtle’s arms and legs, respectively). Presumably, it is the chemical uniqueness of this configuration that resulted in nature’s selection of the myo isomer over others for enzymatic modification and biologic significance.

Myo-inositol’s most basic function, like other polyols (such as sorbitol), is as an osmolyte whose increased cellular concentration reflects responses to hyperosmolarity. However, inositol’s full biologic potential is only realized via chemical modification. Although inositol participates in several varied enzymatic reactions, it is the combinatorial substitution of phosphate moieties to the six hydroxyls that impart preeminent metabolic and functional significance to this deceptively simple molecule. Mathematically, 63 such combinations are possible; this number, however, is an underestimate as di- and triphosphates (also known as pyrophosphate) moieties also exist. To date, over 30 different inositol polyphosphates have been observed across euarchdolic evolution resulting in a fairly crowded metabolic map. Complicating matters, the cellular biosynthesis of inositol polyphosphates—composed of only myo-inositol and phosphates and therefore water-soluble—is not achieved by the mere sequential phosphorylation and/or dephosphorylation of the myo-inositol ring (although there are exceptions in...
Inositol Polyphosphates

(a) Myo-inositol and its polyphosphate derivatives. Although nine stereoisomeric configurations of inositol are possible, the turtle-like myo-inositol with its single axial hydroxyl is the most biologically relevant. Also depicted are the scyllo- and neo-inositols with zero and two axial hydroxyls, respectively. The modern D-numbering system for inositols is counterclockwise viewed from above and assigns the axial hydroxyl group to C2 of myo-inositol. (b) Inositol polyphosphates. Depicted are representatives of phosphorylated derivatives of myo-inositol, including Ins(1,4,5)P3, the calcium releasing factor; InsP6, the naturally most abundant fully phosphorylated inositol polyphosphate; and InsP7, an inositol diphosphate (pyrophosphate).

Receptor-Stimulated Inositol Metabolism and the “Inositol Cycle”

The birth of receptor-stimulated inositol polyphosphate metabolism and its role in cell signaling was the observation in the early 1950s of acetylcholine-stimulated 32P incorporation into inositol lipids (termed the “PI response”) (3). However, it would take more than two decades for the mechanisms and significance of this response to come to light. Key to the eventual elucidation of the “PI response” was the realization that receptor stimulation leads to the activation of phosphoinositol-lipid-specific phospholipase C (PLC). Upon activation, PLC hydrolyzes the glycerol-phosphate bond in PtdIns(4,5)P2 causing the release of water-soluble inositol 1,4,5-trisphosphate (InsP3 or Ins(1,4,5)P3) and the lipid diacylglycerol (DAG). After receptor-mediated catabolism, a regenerative cycle undertakes to restore PtdIns(4,5)P2 to the plasma membrane. Released DAG acts as the now familiar activator of PKC, but subsequently it reenters the inositol metabolism pathway as CMP-phosphatidic acid (CMP-PtdOH, alternatively named CDP-DAG). Synthesis of CMP-PtdOH proceeds via the phosphorylation of DAG by DAG kinase and conjugation to a cytidine nucleotide. Inositol lipid synthesis initiates as the phosphatidic acid (PA) moiety of CMP-PtdOH is enzymatically donated to the C1-hydroxyl of myo-inositol forming PtdIns. Sequential phosphorylation of PtdIns to PtdIns(4)P and PtdIns(4,5)P2 completes the “inositol cycle” and regenerates the substrate for receptor-activated PLC hydrolysis ([Fig. 2]).
Figure 2  The "inositol cycle." A simplified representation of the regenerative metabolism responsible for the synthesis of InsP₃. Generation of InsP₃ and DAG upon PLC-mediated hydrolysis of PtdIns(4,5)P₂ is followed by sequential dephosphorylation of InsP₃ and DAG’s modification to CDP-DAG. The enzymatic attachment of CDP-DAG to inositol regenerates the lipid precursors that ultimately replenish PtdIns(4,5)P₂.
Inositol Polyphosphates

PtdIns synthesis is often dephosphorylated Ins(1,4,5)P3. Alteration of an "inositol cycle" is concomitantly dependent on the continuous supply of free inositol. The source of myo-inositol for PtdIns synthesis is often dephosphorylated Ins(1,4,5)P3. Alternatively, cells take up extracellular inositol or synthesize inositol de novo.

Myo-inositol uptake and synthesis

Sodium- or proton-coupled myo-inositol transporters provide the most direct route for the initiation of inositol polyphosphate metabolism (affording the inositide researcher a simple avenue for the analysis of inositol polyphosphate dynamics using radiolabeled inositol). To date three different mammalian cotransporters have been identified (SMIT 1 and 2 are Na+-coupled, whereas HMIT uses a proton gradient). Although transporters are widely transcribed in many animal tissues, access to extracellular inositol varies among organs. In the absence or low levels of extracellular inositol, de novo synthesis of free myo-inositol is transcriptionally induced; this seems to be a universal capacity of cells conserved from bacteria to humans. Synthesis is initiated via the cyclization of glucose-6-phosphate by 1D-myo-inositol-3-phosphate synthase (MIPS) and the formation of inositol-3-phosphate (Ins(3)P). Dephosphorylation of Ins(3)P by inositol monophosphatase (IMP) yields the free inositol that can then be incorporated into inositol lipids (PtdIns).

In addition to its role in the de novo synthesis, IMP also functions to dephosphorylate downstream hydrolytic products of InsP1, Ins(1)P, and Ins(4)P. As such, IMP is situated at a metabolic intersection of de novo synthesis and the regeneration of myo-inositol, functioning as a gateway for the completion and maintenance of the "inositol cycle."

InsP1, and InsP2, lithium, and the inositol depletion hypothesis

By and large the inositol monophosphates and bisphosphates are thought to lack messenger functions and are most often conceived as catabolic products of the regenerative portion of the inositol cycle or as components of the "off" switch of inositol polyphosphate signaling. Sequential dephosphorylation of Ins(1,4,5)P3 in most eukaryotes proceeds with the concomitant dependence on the continuous supply of free inositol. The source of myo-inositol for PtdIns synthesis is often dephosphorylated Ins(1,4,5)P3. Alternatively, cells take up extracellular inositol or synthesize inositol de novo.

The use of lithium to treat manic-depressive illness dates back to 1949 (4). J.F.J. Cade, while working at a psychiatric hospital, was using guinea pigs to evaluate the effects/toxicity of injections of urine collected from psychiatric patients, testing the hypothesis that a "toxin" may be responsible for the patients' illnesses. In the course of these and other related experiments, he began using the lithium salt of uric acid—chosen merely for its high solubility—and noted a marked depressive effect on the animals' behavior after injection, which he later attributed to the lithium rather than to the uric acid. These experiments were followed by the successful clinical use of lithium in patients suffering from mania and bipolar disorders (4).

Studies in the early and mid-1970s evaluating the effects of lithium administration on inositol levels in rat brains identified a dramatic decrease in the free inositol levels accompanied by a concomitant increase in inositol monophosphates and suggested IMP as the plausible pharmacologic target mediating the therapeutic actions of lithium (5). Subsequent detailed biochemical studies of IMP's inhibition by lithium demonstrated its uncompetitive inhibition by therapeutically relevant concentrations of the ion (6). Surprisingly little attention was paid to the early studies of lithium's effects on inositol metabolism, and it was not until the formulation of the "inositol depletion hypothesis" by Berridge et al. that inositol's mechanism of action and role in aberrant neuronal signaling came to the forefront of the psychopharmacology of lithium's actions in the brain (7).

Stated explicitly, the proposal put forth by Berridge et al. suggests that the cells of the central nervous system (CNS) are uniquely sensitive to the inhibitory effects of lithium on IMP as a result of their limited access to extracellular inositol because of its poor penetration of the blood-brain barrier. Thus, the brain is extensively dependent on IMP's role in de novo inositol synthesis and IMP's functionality in the recycling of inositol monophosphates. Lithium's inhibition of IMP results in the slowing down of the "inositol cycle" and depletes the inositol pool necessary for the production of Ins(1,4,5)P3 in response to receptor stimulation. The corollary of this hypothesis is that, in part, it is the overstimulation of inositol metabolism in the brains of manic-depressive patients that is responsible for the disease's manifestation. In the nearly 20 years since the formulation of the "inositol depletion hypothesis," additional substrates for lithium's actions have been identified, including both the upstream Ins(1,4)P2;Ins(1,3,4)P3, 1-phosphatase as well as the serine/threonine kinase glycogen synthase kinase-3 (GSK-3), whose regulation seems to be independent of inositol metabolism, which suggests lithium's effects on the CNS are more complicated than first envisioned. Nevertheless, it was in fact Berridge et al.'s use of lithium in examining the "PI response" that paved the way for the identification of Ins(1,4,5)P3 as the now well-known calcium-mobilizing second messenger.
Inositol Polyphosphates: Form and Function

Ins(1,4,5)P₃ and Ca²⁺

The identification of Ins(1,4,5)P₃ and its biologic function and significance in the PI3K pathway continually eluded investigators until a conceptual connection between calcium dynamics and receptor-stimulated phosphoinositol turnover was proposed by M ichel (18). Working with blowfly salivary glands, Berridge and Finch demonstrated the cellu lar capacity to rescue calcium release upon cell-surface receptor activation to Ca²⁺ release from intracellular stores. Using permeabilized cells, Streb et al. conclusively demonstrated that application of exogenous InsP₃ mobilized Ca²⁺ from the endoplasmic reticulum (10). A search for the InsP₃ receptor ensued and culminated in the cloning, purification, and reconstitution of the receptor as well as the demonstration that the protein itself is an InsP₃-gated Ca²⁺ channel (11, 12). Calcium dynamics and the regulation of the InsP₃ receptor are still very active research areas. However, by the end of 1980s, a functional and elegant paradigm for soluble inositol metabolism was established.

Ins(1,3,4,5)P₄:C₃-hydroxyl phosphorylation

It was the unexpected identification of an alternative isomer of InsP₃, Ins(1,3,4,5)P₄, derived from the dephosphorylation of Ins(1,3,4,5)P₄ that revealed that there was clearly more to inositol polyphosphate metabolism than circumscribed by the “inositol cycle” (13). Indeed within a few years’ span in the mid-to-late 1980s, the discovery of the C₃-hydroxyl phosphorylation of the inositol ring ushered in an era of “proliferation” for the inositides, within both the lipid and water-soluble arenas of their metabolism. During this time, an inositol lipid 3-kinase activity was found associated with viral Src and the middle-T antigen of the polyoma virus (14). Phosphoinositide-3 kinase (PI3K) produces PtdIns(3,4,5)P₃ from PtdIns(4,5)P₂ upon cell-surface receptor activation and thus they remain relegated to the “lipid switch” for inositol polyphosphate metabolism and function. However, by the end of 1980s, a functional and elegant paradigm for soluble inositol metabolism was established.

Ins(1,3,4,5)P₄ and Ca²⁺

The identification of Ins(1,4,5)P₃ and its biologic function and significance in the PI3K pathway continually eluded investigators until a conceptual connection between calcium dynamics and receptor-stimulated phosphoinositol turnover was proposed by Michel (18). Working with blowfly salivary glands, Berridge and Finch demonstrated the cellular capacity to rescue calcium release upon cell-surface receptor activation to Ca²⁺ release from intracellular stores. Using permeabilized cells, Streb et al. conclusively demonstrated that application of exogenous InsP₃ mobilized Ca²⁺ from the endoplasmic reticulum (10). A search for the InsP₃ receptor ensued and culminated in the cloning, purification, and reconstitution of the receptor as well as the demonstration that the protein itself is an InsP₃-gated Ca²⁺ channel (11, 12). Calcium dynamics and the regulation of the InsP₃ receptor are still very active research areas. However, by the end of 1980s, a functional and elegant paradigm for soluble inositol metabolism was established.

Ins(1,3,4,5)P₄ and Ca²⁺

The identification of Ins(1,4,5)P₃ and its biologic function and significance in the PI3K pathway continually eluded investigators until a conceptual connection between calcium dynamics and receptor-stimulated phosphoinositol turnover was proposed by Michel (18). Working with blowfly salivary glands, Berridge and Finch demonstrated the cellular capacity to rescue calcium release upon cell-surface receptor activation to Ca²⁺ release from intracellular stores. Using permeabilized cells, Streb et al. conclusively demonstrated that application of exogenous InsP₃ mobilized Ca²⁺ from the endoplasmic reticulum (10). A search for the InsP₃ receptor ensued and culminated in the cloning, purification, and reconstitution of the receptor as well as the demonstration that the protein itself is an InsP₃-gated Ca²⁺ channel (11, 12). Calcium dynamics and the regulation of the InsP₃ receptor are still very active research areas. However, by the end of 1980s, a functional and elegant paradigm for soluble inositol metabolism was established.
Inositol Polyphosphates

Ins(3,4,5,6)P₄ - Ca²⁺-activated chloride (Cl⁻) channel regulation

Two separate biosynthetic routes have been proposed for the production of Ins(3,4,5,6)P₄. The "correctness" of both pathways, whether they coexist or merely display species differences, is still a matter of debate. Regardless, both depend on the initial C₃-hydroxyl phosphorylation of Ins(1,4,5)P₃. The first pathway, currently only supported by experiments in avian erythrocytes, depends on several sequential dephosphorylation and phosphorylation reactions that isomerically interconvert between InsP₇ and InsP₅. This pathway can be summarized as follows: Ins(1,3,4,5,6)P₅ → Ins(1,3,4,5)P₄ → Ins(1,3,4,5,6)P₆ (18). A second proposed pathway, which is prevalent in mammalian cells, proceeds via the C₁-hydroxyl dephosphorylation of Ins(1,3,4,5,6)P₅. Surprisingly, despite its use of Ins(1,3,4,5,6)P₅ as a precursor for the formation of Ins(3,4,5,6)P₄, this latter pathway, like the first, also seems to depend on the dephosphorylation of Ins(1,3,4,5)P₃. In a theme of competitive protection that is likely repeated often in inositol metabolism, the generation of Ins(1,3,4,5)P₃ after PLC activation results in the protection of the cellular pools of Ins(3,4,5,6)P₄ by competing as a substrate for a dual specificity kinase that phosphorylates Ins(1,3,4,5,6)P₅ back to Ins(1,3,4,5,6)P₆. Ins(1,3,4,5,6)P₇ 5,6-Ins(3,4,5,6)P₄ 1-kinase (also named ITPK1) is one of several inositol polyphosphate kinases displaying substrate promiscuousity, an enzymatically conservative yet metabolically proliferative measure on behalf of evolution that also affords the potential of competitive "crosstalk." Startlingly, the 1- Ins(1,3,4,5,6)P₅ phosphatase responsible for Ins(1,3,4,5,6)P₅ production turns out to be the very same ITPK1. Thus, ITPK1 functions as a reversible kinase capable of the interconversion of Ins(1,3,4,5,6)P₅ and Ins(3,4,5,6)P₄ (19). The balance of phosphatase/kinase activities seems to be regulated by Ins(1,3,4,5)P₃ such that Ins(1,3,4,5)P₃ stimulates the phosphatase reaction. Reflecting the regulatory "crosstalk" involved in its production, Ins(3,4,5,6)P₄ accumulation in response to persistent PLC activation occurs relatively slowly but persists long after other inositol metabolites have returned to their prestimulated levels, a feature likely important in its regulation of Ca²⁺-activated Cl⁻ channels. Cl⁻ secretion via Ca²⁺-activated channels serves to physiologically regulate epithelial salt and fluid secretion in the gastrointestinal tract, exocrine glands, and lungs of animals. A putative role for a PLC-dependent factor in the negative regulation of Cl⁻ secretion was suggested by experiments in which a Ca²⁺-mediated activation of Cl⁻ secretion upon receptor activation of PLC were followed by a period refractory to the stimulating effects of Ca²⁺. Correlating channel activation and deactivation with cellular levels of downstream PLC-dependent inositol polyphosphates, Sheets and associates identified Ins(3,4,5,6)P₄ as the likely mediator of the inhibition of Cl⁻ conductance and went on to show that a cell-permeant analog of Ins(3,4,5,6)P₄ decreased Ca²⁺-dependent Cl⁻ secretion (20). The apical distribution of ITPK1 in polarized epithelial cells localizes Ins(3,4,5,6)P₄ near its presumed site of action. However, the precise mechanism or target of inhibition has remained elusive. Although much of the delineation of the physiologic importance of Ins(3,4,5,6)P₄ has been derived from the study of epithelial cells, Ca²⁺-activated Cl⁻ channels occur in most cell types raising the potential of Ins(3,4,5,6)P₄-mediated regulation of diverse cellular processes ranging from neurotransmission to smooth-muscle contraction (21).

As indicated, the regulation of Ca²⁺ and/or Ca²⁺-dependent processes by Ins(1,4,5)P₃, Ins(1,3,4,5)P₄, and Ins(3,4,5,6)P₄ likely represents a recent addition to the repertoire of inositol polyphosphate functionality. Soon after the discovery of C₃-hydroxyl phosphorylated inositol polyphosphates came the realization that more phosphorylated inositol species were indeed conserved ubiquitous inositol metabolites predominating the metazoan emergence of InsP₃-receptor-mediated Ca²⁺ signaling. These highly phosphorylated species included inositol hexakisphosphate (Ins(1,2,3,4,5,6)P₆), known at the time to occur in avian erythrocytes, and the fully phosphorylated inositol hexakisphosphate (Ins(1,2,3,4,5,6,7)P₇, also called phytic acid or InsP₇), thought to be predominantly relegated to the plant kingdom. Although the biosynthetic pathways of highly phosphorylated inositol polyphosphates have yet to be fully elucidated, over the last decade significant strides have been made in their metabolic and functional characterization, including the identification of inositol polyphosphate kinases ancestral to the Ca²⁺-dependent Ins(1,4,5,6)P₃ 3-kinases. In large part these advances are the byproducts of the genetic tractability of one of the simplest model organisms, the budding yeast Saccharomyces cerevisiae.

Ins(1,4,5,6,7)P₆ and Ins(1,3,4,5,6,7)P₆: transcriptional regulation

In the budding yeast, inositol polyphosphate synthase's proceeds via what is likely to be one of the earliest incarnations of a PLC-dependent pathway for higher inositol polyphosphate metabolism in eukaryotes. Early biochemical studies in yeast (and plants) failed to identify a calcium-sensitive Ins(1,4,5)P₃ 3-kinase activity analogous to that found in mammalian cells. Instead, these studies identified C₁-hydroxyl phosphorylation of Ins(1,4,5)P₃ and formation of Ins(1,4,5,6)P₄ as the most likely first anabolic step in the production of higher inositol polyphosphates (22). Additional biochemical studies identified the sequential phosphorylation of Ins(1,4,5,6)P₄ to Ins(1,3,4,5,6,7)P₇ followed by Ins(1,2,3,4,5,6,7)P₇ (23). These findings were interpreted as proof of the existence of disparate pathways in yeast and mammals for the metabolism and functionality of Ins(1,4,5)P₃. In contrast, the eventual cloning of the yeast Ins(1,4,5)P₃ kinase activity found that mammalian and yeast inositol metabolism were more closely related than initially suspected.

Biochemical and genetic characterization of the yeast Ins(1,4,5,6)P₃ kinase revealed it to be a dual-specificity Ins(1,4,5)P₃, 3/6-kinase, rapidly converting Ins(1,4,5)P₃ to Ins(1,3,4,5,6)P₄ via the intermediate production of Ins(1,4,5,6)P₃. Sequence comparisons of the yeast Ins(1,4,5,6)P₃ 3/6-kinase with the mammalian Ins(1,4,5,6)P₃ 3-kinases demonstrated them to be part of a closely related family and suggested the Ins(1,4,5)P₃ 3-kinases were merely a recent evolutionary modification and specialization of this catalytically versatile kinase family (14). The cloning
and characterization of the yeast Ins(1,4,5)P3 3/6-kinase was accomplished contemporaneously by two independent groups (24, 25). Noting its catalytic versatility, the Snyder group dubbed the enzyme inositol polyphosphate multikinase (IPMK), whereas the York group chose inositol polyphosphate kinase 2 (Ipk2) as it was one of two simultaneously identified kinases necessary for the synthesis of InsP6 (26). However, yeast IPMK Ipk2 was in fact identical with the previously characterized Arg82, which was a protein with a history of nearly two decades of research that had clearly identified it as a nuclear, transcriptional regulator. Studies of yeast deficient for Arg82 identified it as a required factor for the transcriptional regulation of genes comprising arginine anabolic and catabolic metabolism. Transcriptional regulation of these arginine-sensitive pathways permits yeast to use alternative nitrogen sources such as arginine or ornithine when the preferred sources of ammonia or glutamate are limiting. In the absence of Arg82, yeast fail to grow on media with arginine or ornithine as the sole nitrogen source (27). The identification of yeast IPMK/Ipk2 as Arg82 (henceforth called IPMK) prompted the evaluation of the role its kinase activities may play in regulating these arginine-sensitive transcriptional responses. Monitoring growth phenotypes of various inositol polyphosphate mutants in media containing arginine or ornithine as the sole nitrogen source, York and associates identified the necessity of Pcl1 and IPMK but not downstream inositol polyphosphate kinases for growth on selective media (25). These results supported the conclusion that it is the production of Ins(1,4,5)P6 and/or Ins(1,3,4,5,6)P5 that is required for the regulation of arginine-dependent gene expression. Although determinants of IPMK may play a role in its functional contribution to arginine-sensitive transcription (28), additional experiments using IPMK orthologs found in the fly and plant that lack previously identified yeast regulatory sequences suggest IPMK (1,4,5)P6 and/or Ins(1,3,4,5,6)P5 roles are themselves sufficient for modulating arginine-dependent transcription; however, the molecular mechanism remains unknown (29, 30). Roles for Ins(1,4,5)P6 and/or Ins(1,3,4,5,6)P5 in transcriptional regulation have been supported by subsequent genetic screens in yeast that identified the requirement of Ins(1,4,5)P6 and/or Ins(1,3,4,5,6)P5 in phosphate-responsive regulation of chromatin remodeling and transcription (31). The ipmk null yeast are defective in the induction of the PHO5 gene in response to phosphate limitation. Remodeling of the PHO5 promoter chromatin is impaired in ipmk mutant yeast as a result of defects in the recruitment of ATP-dependent chromatin-remodeling complexes. Specific roles for either Ins(1,4,5)P6 or Ins(1,3,4,5,6)P5 were once again suggested as chromatin remodeling defects were also evident in pcl1 mutant yeast, but not in yeast deficient in downstream inositol polyphosphate kinases (33). A mechanism for Ins(1,4,5,6)P5 mediating repression of chromatin remodeling was suggested by in vitro nucleosome mobilization assays in which Ins(1,4,5,6)P5 and Ins(1,3,4,5,6)P5 were found to have direct stimulatory effects on chromatin remodeling complexes (32). However, these latter studies have been faulted for the use of nonphysiologic concentrations of the inositol phosphates, which suggests additional characterization of the process is necessary. The identification in yeast of a concise anabolic pathway for higher inositol polyphosphates in the nucleus suggests primordial roles for their metabolism consistent with earlier suppositions on the nuclear origin of lipid inositol metabolism (33). Indeed, IPMK not only functions as an inositol polyphosphate kinase, but it is also a nuclear lipid inositol PI3K kinase that predates the evolutionary emergence of the unrelated viral-Src- and middle-T-antigen associated kinases (discussed above). Like its water-soluble inositol polyphosphate activities, IPMK-mediated synthesis of nuclear PtdIns(3,4,5)P2 has also been linked to transcriptional regulation in yeast (34). Whether transcriptional regulatory roles for Ins(1,4,5,6)P5/Ins(1,3,4,5,6)P5 (or PtdIns(3,4,5)P2) are evident in higher eukaryotes remains unknown, although roles in nuclear processes are likely, as nuclear IPMK paralogs have been identified across evolution, including mammals. Characterization of mammalian IPMK showed it had retained its catalytic versatility and was capable of sequential phosphorylation of Ins(1,4,5)P3 to Ins(1,3,4,5,6)P5 (35); however, like the Ins(1,4,5)P3 3-kinase mammalian IPMK prefers to initiate this metabolism via C3-hydroxyl phosphorylation. Additional studies have extended IPMK’s catalytic repertoire and have suggested that in mammalian systems, IPMK serves to synthesize Ins(1,3,4,5,6)P5 via the C5-hydroxyl phosphorylation of Ins(1,3,4,6)P4, the product of Ins(1,3,4,5)P5 phosphorylation by ITPK1 (discussed above) (36). Thus, in mammals, two pathways may exist for the production of Ins(1,3,4,5,6)P5. One pathway is analogous to that found in yeast, mediating exclusively via IPMK, but the second follows a more circuitous route initiated by Ins(1,4,5)P3 3-kinase and ultimately is still dependent on IPMK. This latter pathway can be summarized as follows: Ins(1,4,5)P3 \( \rightarrow \) Ins(1,3,4,5)P4 \( \rightarrow \) Ins(1,3,4)P3 \( \rightarrow \) Ins(1,3,4,5,6)P5 \( \rightarrow \) Ins(1,3,4,5,6)P5. Whether both routes do indeed exist in mammals or whether one pathway predominates is still a matter of some debate. Perhaps the occurrence of both pathways simply reflects cytoplasmic versus nuclear signaling. Whichever the route, production of Ins(1,3,4,5,6)P5 and its more phospho-derivative relatives seems to be of preeminent importance for mammalian systems as deletion of IPMK in mice results in early embryonic lethality and severe developmental defects (37). Ins(1,2,3,4,5,6)P6: from chelator to cofactor Production of Ins(1,3,4,5,6)P5 is the penultimate step in the synthesis of the most abundant inositol phosphate on earth, Ins(1,2,3,4,5,6)P6. Its synthesis is achieved via the phosphorylation of Ins(1,3,4,5,6)P5’s lone remaining axial hydroxyl by an Ins(1,3,4,5,6)P5 2-kinase (IPK2) (28). InsP6’s quantitative dominance on earth is owed largely to its use by plants for phosphate storage in seeds. Although it was the first inositol polyphosphate discovered, it was not until the development and application of high-pressure liquid chromatography (HPLC) in the analysis of inositol polyphosphates that its prevalence in animal cells was appreciated. With intracellular concentration in mammalian cells ranging between 10 and 100 μM, it is the most...
abundant inositol polyphosphate species, often exceeding most others by an order of magnitude or more. Even higher concentra-
tions have been observed in the slime mold Dictyostelium discoideum, which possesses unique metabolic pathways for
inositol polyphosphate synthesis independent of inositol lipids and their hydrolysis by PLC. In fact, in D. discoideum (and in
some plants), higher inositol polyphosphate synthesis can pro-
cceed via the direct sequential phosphorylation of InsP3(3P) after
cyclization of glucose-6-phosphate. In these unusual organisms
InsP4 concentrations can reach as high as 700 µM. Because of
its high charge density, InsP4 is a strong chelator that read-
ily forms insoluble salts with polyvalent cations resulting in its
precipitation at higher concentrations. Under cellular ionic con-
dition found in animal cells, soluble InsP4 concentrations are
limited to ~50 µM (likely in the form a neutral, stable pentam-
agranium salt). Thus, it is likely that much of cellular InsP4 is
actually found in a “bound state.”

Several proteins have been identified to bind InsP6. As
InsP6 interacts strongly with positively charged residues, cau-
tion has been urged in interpreting in vitro binding exper-
iments. Nevertheless, physiologic roles for InsP6 have been
suggested in several processes throughout the cell. In the cy-
toplasm, InsP6 binds tightly to clathrin assembly proteins neg-
avely regulating the assembly of clathrin-coated vesicles and
receptor-mediated endocytosis at the plasma membrane (38). Nu-
clear roles for InsP6 have also been suggested. In mammals,
DNA double-strand breaks can be repaired by nonhomologous
end-joining (NHEJ) requiring the activity of DNA-dependent
DNA-PK. In vitro studies demonstrated the stimulation of DNA-PK dependent NHEJ activity by InsP6 and
identified the Ku70/80 subunits of DNA-PK as direct InsP6 tar-
gets (39). Recently, a crystallographic study unexpectedly found
InsP6 bound within the enzymatic core of an RNA editing
enzyme (ADAR2) belonging to a class of adenosine deami-
nases and further demonstrated its requirement as a cofactor for
this class of enzymes (40). Consistent with a conserved role
in processes involving nucleic acids, a yeast genetic screen for
required factors involved in mRNA export unabiguously iden-
tified Pcl1, IPMK, and Ipk1 and their sequential enzymatic roles
in InsP6 synthesis (26). Finally, like the IPMK mouse knock-
out, IPK1 knockout mice also display early embryonic lethality
demonstrating a critical role for higher inositol polyphosphate
synthesis in mammalian development (41).

In contrast to the rapid changes upon receptor activation
in intracellular concentrations of lower inositol polyphosphates
(e.g., Ins(1,4,5)(P3), InsP3, and InsP4) concentrations seem immute.
This finding, along with observations of the sluggish incorporation
of radiolabeled inositol into InsP6, prompted interpretations of
slow turnover rates for InsP6 in vivo. However, the identification of
InsP6 as a precursor to even more phosphorylated inositol
polyphosphates overturned these misconceptions and revealed
the true dynamic nature of cellular InsP6.

**PP-InsP6 and PP2-InsP6: signaling and protein phosphorylation**

That more than six phosphates can fit onto an inositi-
ol ring was first discovered in D. discoideum (42) and
fluoride-treated pancreatoma cells (43). Subsequent stud-
ies confirmed the widespread evolutionary diversity of the
diphosphate-containing inositols. These “high energy”
molecules contain the fully phosphorylated inositol ring of
InsP6 with additional pyrophosphate moieties on either one
or two α-phosphates. The best characterized diphosphorylated
inositol polyphosphates are the diphosphoinositol pentakisphos-
sphate (InsP5, PP-InsP5) and bis-diphosphoinositol tetrakisphos-
sphate (InsP6, PP2-InsP6) species. In most cell types, these
species are present in submicromolar concentrations, repre-
senting 1-5% of total InsP6 levels. Remarkably, in mam-
alian cells, up to 50% of InsP6 cycles through the more
phosphorylated diphosphates every hour (43). Once again the
slime mold represents the extreme of higher inositol polyphos-
phates metabolism with diphosphates species reaching concen-
trations of 200 µM. Nuclear magnetic resonance spectroscopy
analysis of inositol diphosphates in D. discoideum showed a
single InsP6 isomer present in the cytoplasmic fraction of CS and
C6 (3,6-(PP2)-InsP6) and the two respective InsP5 isomers,
5-PP-InsP5 and 6-PP-InsP5. So far, in mammalian cells, only
a single InsP6 isomer has been detected and demonstrated to be
the CS- and PP2-InsP6 species. The standard free energy of
hydrolysis of the pyrophosphate bond in InsP6 (the nonphysi-
ologic l-PP-InsP6) has been estimated at 6.6 kcal/mol, higher
than that of adenosine-5′-diphosphate (ADP) (6.4 kcal/mol) and
only slightly lower than that of ATP (7.3 kcal/mol). These val-
ues are likely to be quite higher for InsP6 isomers with vicinal
pyrophosphates as a result of steric constraints and strong elec-
 trostatic repulsion. In addition to InsP6-derived diphosphates, in-
ositol species derived from the pyrophosphorylation of InsP6 and
InsP5 have also been observed while triphosphate-containing
InsP6 species have not been ruled out, implying an overall
more diverse ensemble of high energy inositols than first sus-
pected (44).

The enzymatic capacity for the synthesis of diphosphory-
lated inositols has so far been attributed to a single class of
InsP6-kinases conserved across the evolutionary spectrum.
Indeed the InsP6-kinases, along with IPMK and Ins1,4,5P3
3-kinases, define a single, highly related, conserved family with
representative members identified in every eukaryotic genome
studied thus far. Surprisingly, studies examining the occurrence
of this kinase family in some of the “earliest” eukaryotes suggest
InsP6-kinases may be founding representatives as exemplified in
the Giardia genome, which contains a single InsP6-kinase (45).
Along with IPMK and Ipk1, the IPK3-kinase (denoted Kcs1 in
yeast) completes the list of identified inositol polyphosphate ki-
nases in the yeast genome. InsP6-kinase can mediate the produc-
tion of both InsP6 and InsP5 as well as diphosphate derivatives of
InsP5 and InsP6 (46, 47). However deletion studies in yeast
(as well as the absence of a clear paralog in plants, although
they contain IPMK) suggest additional enzymes—including an
InsP6-kinase capable of inositol pyrophosphorylation may ex-
ist (48, 49). It is worth noting that at least in vitro mammalian
and yeast IPMK have also been shown to generate inositol
diphosphates (35).

As yeast contains only a single InsP6-kinase (mammals
contain three isoforms), much of what we know regarding
the functional roles of inositol diphosphates is the result of
loss-of-function studies in the kas1 yeast mutant. Expanding on roles suggested for InsP6 in vesicle formation and endocytosis, in yeast, InsP3, InsP7, and InsP8 seem to regulate vesicular trafficking. Loss of Kas1 results in altered vacuolar morphology with the appearance of smaller fragmented vacuoles and the accumulation of membranous, vesicular structures as a result of aberrant endosomal processing (50). Inositol diphosphates also likely complement the transcriptional regulatory roles of InsP6–InsP8 in phosphate metabolism. InsP6-kinase was in fact identified as a cDNA stimulating inorganic phosphate uptake into Xenopus eggs (51). Furthermore, yeast deficient in inositol diphosphate synthase’s lack polyphosphate synthesis/storage (52). Recently, a remarkable role for inositol diphosphates in telomere length maintenance was revealed by studying ipk1-deficient yeast (49, 53). Telomeres consist of chromosomal caps of long, repetitive DNA sequences that prevent nucleolytic degradation and protect more internal coding sequences from chromosomal shortening associated with cell division. In the absence of InsP6, InsP7-kinase can use InsP6, which accumulates in yeast in the absence of ipk1. As a result, in ipk1-deficient yeast, the products of InsP6 pyrophosphorylation, PP-InsP7, and (PP)2-InsP8 also accumulate to levels not observed in wild-type yeast and in fact exceed concentrations normally observed for InsP7 [PP-InsP7] and InsP8 [(PP)2-InsP8]. When compared with wild-type yeast, ipk1-deficient yeast displays shortened telomere length, whereas yeast deficient in InsP6-kinase activity display increased telomere length, which suggests inositol diphosphates (increased in ipk1-deficient yeast and absent in kas1-deficient yeast) negatively regulate telomere extension. Although the mechanism for these effects remains unknown, inositol diphosphates may target the regulatory actions of phosphatidylinositol 3-kinase-related kinases (protein kinases of evolutionary relatives to PI3Ks) known to be involved in telomere length maintenance (44). While the functional significance for inositol diphosphates observed in yeast is likely to be conserved in mammalian cells, the roles for InsP6, InsP7, and InsP8 in mammalian physiology are more difficult to study because of the presence of three isoforms of InsP6-kinase with distinct subcellular distributions. In mammals, several lines of evidence suggest inositol diphosphates regulate cell death/apoptosis (54) and are dynamically regulated in response to environmental stress (48). As is the case for other inositol polyphosphates, binding allosteric mechanisms for inositol diphosphates’ functional identity have also been proposed, although it must be noted that binding partners must demonstrate significant specificity to allow for meaningful regulation by InsP6–InsP8 within the cellular context of the much more abundant InsP6. In D. discoideum, where the levels of InsP6/InsP8 rival those of InsP6, evidence for the specific regulation by inositol diphosphates of PH-domain lipid binding (analogous to that suggested for InsP1,2,4,5,PK discussed above) has revealed a role in both InsP6 and InsP8-dependent chemotaxis (55). However, the high energy potential of the β-phosphates in inositol diphosphates suggests unique functionality for this class of inositol polyphosphates beyond allosteric competition. Indeed, early characterization of InsP6-kinase demonstrated its capacity for ATP formation in the reverse by transferring a β-phosphate from InsP6 to ADP (56). More recently, Snyder and associates have demonstrated the occurrence of likely more functionally significant acceptors of the high energy β-phosphates than ADP: proteins (57). InsP6 seems to physiologically phosphorylate a variety of protein targets, although its activity seems restricted to eukaryotes. Identification of specific targets in yeast suggested a phosphorylation consensus sequence consisting of an acidic, polyserine stretch interspersed with aspartate or glutamate residues. Such stretches proved to be excellent CK2 (formerly casein kinase-2) substrates. However, careful characterization of the process has demonstrated InsP6-mediated phosphorylation to be quite distinct from ATP/kinase-mediated phosphorylation. In contrast to ATP, InsP6 does not require separate protein kinases but directly phosphorylates its targets in a nonenzymatic and temperature-dependent reaction. Nevertheless, protein kinases such as CK2 are indeed critical for the process but function only to “prime” targets. The significance of kinase-mediated pre-phosphorylation/priming and its requirement for InsP6-mediated nonenzymatic phosphorylation has only recently been elucidated. Biochemical characterization of phosphorylated substrates suggests the chemical nature of InsP6-phosphorylated serines is quite distinct from canonical phospho-serines. In a dramatic twist to protein phosphorylation, InsP6-mediated phosphorylation is in fact serine pyrophosphorylation (R. Bhandari and S.H. Snyder, personal communication). Although the physiologic roles of this novel posttranslational modification need to be evaluated, serine pyrophosphorylation likely expands on the already well-defined roles of canonical protein phosphorylation in modifying protein conformation, regulating catalytic activity, determining protein localization, or altering protein–protein interactions. Experiments assessing these possibilities are currently underway.

Perspective

As the most recently identified members of the inositol polyphosphate family, inositol diphosphates have already complex and universal signaling network. Although they have been known for more than 20 years since the discovery of Ins(1,4,5)P3 as a second messenger, much of the functional and cell signaling roles of most inositol polyphosphates remains poorly understood. Nevertheless it is clear that inositol polyphosphates serve to dynamically organize, regulate, and orchestrate nearly all aspects of cell biology (Fig. 3). Full elucidation of their complex metabolism and roles will likely necessitate the development of better and less laborious analytic methods for the evaluation/identification of their dynamic regulation as well as methods to track their subcellular localization. Although such tools (as well as pharmacologic inhibitors) exist for their lipid counterparts, the inositol polyphosphates have lagged behind, in part because of their greater complexity. Nevertheless, the molecular identification and cloning of the inositol polyphosphate kinases (and phosphatases) achieved in the last decade has allowed for the generation of new experimental/genetic models. Although preliminary characterizations of these have confirmed inositol polyphosphates’ preeminent importance in cell biology
and physiology, they have also revealed how much more we have yet to learn.

Although every effort was made to cite relevant original research, unavoidable omissions because of space constraints exist, for this, the authors apologize. More complete descriptions and referencing of original sources can be found under “Further Reading.”

References

Inositol Polyphosphates


Further Reading


Riboy V. m6o-Inositol-1,2,3,4,5,6-hexakisphosphate. Phytochemistry 2003;64:1033–1034.


See Also

Calcium Signaling

Phosphatidylinositol
Integrin Signaling, Bidirectional-Signaling of
Xiaoyang Wu and Jun-Lin Guan, Cornell University, Ithaca, New York
doi: 10.1002/9780470048672.wecb253

Integrins are a large family of heterodimeric receptors that mediate the adhesive behavior of cells. Most integrins bind to extracellular matrix (ECM) molecules, and they transmit signals that are critical in growth, development, tissue homeostasis, and host defense. A central feature of these receptors is their ability to transduce bidirectional signals into and out of the cell. In this article, we will give an overview of our current understanding of the structure and cellular signaling functions of integrins.

An amazing feature of life is that cells, the tiny units of an organism, can develop such elaborate molecular systems to sense all kinds of environmental information and translate it into various cellular responses to survive. The extracellular matrix (ECM) is composed of a complex mixture of polysaccharides and large fibrous proteins such as fibronectin, collagen, and laminin. ECM serves as a key environmental cue for cells in multicellular organisms, because many fundamental cellular processes, including proliferation, survival, migration, and differentiation, are regulated by the cells’ adherence to the ECM and the composition of the ECM (1, 2). In the 1980s, integrins were recognized as the major metazoan receptors for the ECM (3). Ever since, intensive efforts have been made to unravel its complex functions as an important signal transducer. Importantly, the underlying structure–activity relationships have also started to be elucidated by resolving the high resolution three-dimensional (3-D) structure of integrins. It is now widely appreciated that integrins serve as a pivotal module to mediate a bidirectional signal transduction, namely outside-in and inside-out signaling pathways (4). Upon engagement of ECM ligands, integrins make transmembrane connections to cytoskeleton and regulate many intracellular responses through the outside-in signaling. Conversely, intracellular signaling pathways can modulate integrin-mediated cell adhesion to ECM via the inside-out signaling. In this article, we summarize recent progress in this area by discussing the bidirectional signaling of integrins.

Integrin Family (An Overview)

A functional integrin is a heterodimeric protein complex consisting of an α and β subunit. So far, the mammalian integrin family consists of 8 β subunits and 18 α subunits, which are known to assemble into 24 distinct integrins (Fig. 1). The different combination of subunits accounts for heterodimer specificity, with certain integrins showing preference for particular ECM molecules (4–6). Each subunit has a large extracellular domain, a single transmembrane (TM) segment, and a short cytoplasmic tail (with the exception of β4). The N-terminal portion of the α and β subunits associate to form the headpiece, which contains the ligand-binding site, whereas the C-terminal segments traverse the plasma membrane and mediate interaction of the integrin with the cytoskeleton and other signaling molecules. Hence, the exterior and interior of a cell are physically linked by integrins, which allows the bidirectional transmission of mechanical and biochemical signals across the plasma membrane, and leads to a cooperative regulation of cellular functions, including adhesion, migration, growth, survival, and differentiation.
Most integrins bind to ECM components. The ECM is a complex network of polysaccharides and proteins with high molecular weight, such as laminins, collagens, vitronectin, and fibronectin, and with the basement membrane. It has also been demonstrated that the ECM sequesters growth factors and plays a critical role in the differentiation and growth of various cell types. Integrins containing the $\alpha_2$, $\alpha_5$, $\alpha_6$, or $\alpha_{10}$ subunits recognize the RGD (Arg-Gly-Asp) motif of ECM components, namely fibronectin and vitronectin (4–6). Although laminins and collagens also contain the same RGD sequences, they are normally cryptic and inaccessible to the integrin receptors. Instead, these ECM proteins are recognized by integrins containing the $\alpha_2$, $\alpha_5$, and $\alpha_{10}$ subunits (laminin-binding receptors) or the $\alpha_1$, $\alpha_2$, $\alpha_3$, and $\alpha_{13}$ subunits (collagen-binding receptors).

In addition to serving as ECM receptors, some integrins (e.g., integrins of hematopoietic cells) can bind to counter-receptors on other cells, such as intercellular adhesion molecules (ICAMs) and the vascular cell adhesion molecule-1 (VCAM-1) (7). Integrins of hematopoietic cells also recognize plasma proteins that are deposited at sites of injury (e.g., fibrinogen or von Willebrand factor) and complement factors. Other ligands of integrins include inhibitors of platelet aggregation that are secreted by endothelial cells, such as intercellular adhesion molecules (ICAMs) and the vascular cell adhesion molecule-1 (VCAM-1) (7).

Integrin Regulation of Outside-In Signaling

Integrins recognize positional cues encoded by the ECM and convert them into biochemical signals that control a wide spectrum of cellular behaviors. ECM binding to integrins leads to integrin clustering and recruitment of actin filaments and signaling molecules to the cytoplasmic domain of integrins (4). These specialized, cell adhesion organelles and signaling centers are named focal complexes (nascent adhesion structures) or focal adhesions (FA; fully formed, mature adhesion structures). It has been demonstrated that more than 20 different important signaling proteins are recruited to the ECM-integrin-binding site. Because of the length limit of this article, we will only focus on the role of the MAPK (mitogen-activated protein kinase) pathway, FAK, ILK, and Rho family small G proteins in mediating the outside-in signaling of integrin (summarized in Fig. 2).

MAPK Pathway

Activation of the MAPK pathway provides a common route leading to transcriptional regulation of genes that are crucial for cell-cycle progression and differentiation (10). In a classic model, activation of Ras via growth factor receptors (GRFs) leads to sequential stimulation of the protein kinases Raf, MEK, and, finally, the MAP kinases Erk1 and Erk2. It is now well established that, in addition to growth factors, many normal cells require adhesion to the ECM to proliferate (11). Without cell adhesion to ECM, soluble mitogens alone cause a transient and relatively modest activation of Erk. On the other hand, integrin-mediated adhesion also causes a weak activation of Erk in cells deprived of growth factors. Transient activation of Erk is not sufficient to promote transcription of Cyclin D, but it induces transcription of the CDK inhibitor p21 (12).

Several signaling proteins involved in this pathway have been found in focal adhesion complexes. Based on current understanding, integrin signaling activates ERK through two major mechanisms: the SFK (Src Family Kinase)/FAK pathway, which is activated by most, perhaps all, integrins (will be discussed in detail later); and the SFK/Shc pathway (15), which is activated by a subset of integrins through the transmembrane segment of their $\alpha$ subunits. Both pathways are required for efficient joint integrin/GFR signaling to ERK, as supported by the effect of mutations that inhibit either one of the two pathways (16–18). The $\alpha$-subunit-dependent SFK/Shc signaling is necessary and sufficient for activation of ERK in several cell types (18, 19). Shc is an adapter protein containing SH2 and a phosphotyrosine binding (PTB) domain, which links tyrosine-phosphorylated signaling molecules to Ras (15). Upon binding to activated receptors, Shc is phosphorylated on tyrosine, which serves as a binding site to the G-b2mSOS complex. This process leads to the juxtaposition of the GTP exchange factor domain of mSOS to Ras, and it subsequently activates Ras. In addition, recent results indicate that Shc is also a potent PI3K/Akt activator (20).

In hematopoietic cells, phosphorylated Shc recruits p85b2 through Grb2. Gab2 in turn recruits the regulatory subunit of PI3K, which leads to PI3K activation.

In addition to the Shc-mediated response, some integrins may directly cooperate with GFRs. For example, the $\alpha_5\beta_3$ integrin receptor has been found to uniquely associate with insulin receptor substrate 1 (IRS-1), which is a cytoplasmic signaling transducer of the insulin and insulin-like growth factor receptors (IGFRs) (21). IRS-1 is tyrosine phosphorylated by the activated IGFR and subsequently binds to a variety of signaling molecules. In cells that have adhered to ECM, such as vitronectin through the $\alpha_5\beta_3$ integrin, a subset of IRS-1 binds to the integrin. This interaction substantially enhances the growth-stimulating function of insulin and IGF. In another example, a 190-kDa protein that is tyrosine phosphorylated as a result of PDGF receptor activation also binds to $\alpha_5\beta_3$ receptor, which suggests that there could be cooperation between this integrin and the PDGF signaling pathway (22).

Previous studies have also shown that engagement of $\alpha_5\beta_3$ integrin can induce tyrosine phosphorylation of EGFR (epidermal growth factor
Integrin Signaling, Bidirectional-Signaling of

Integrin receptors relay signals from ECM to regulate a wide spectrum of cellular processes, including migration, cell cycle, survival, and cell differentiation. (Please see text for details of the outside-in signaling of integrins.)

Integrin-mediated adhesion not only activates ERK but also JNK (25–27). JNK is another member of the MAPK family, which is activated by stress stimuli, such as UV-radiation, hyperosmolar conditions, and inflammatory cytokines (28). After activation, JNK translocates into the nucleus and phosphorylates the transcription factor c-Jun, thereby activating c-Jun-dependent transcription (29). c-Jun activation plays an indispensable role in initiation of DNA synthesis (30). Thus, cells derived from mice lacking JNK1 and JNK2 or c-Jun display defective proliferation (31, 32). Integrins can activate signaling to JNK in the absence of a significant contribution from GFRs. α6β4 can activate JNK through Shc and a downstream pathway involving Ras, PI3-K, and Rac (25), whereas β1 and αv integrins signal to JNK through the FAK/Cas/Rac pathway (27, 33). In addition, the β2 integrins are associated with JAB1, a coactivator of c-Jun. Upon their association, a subset of JAB1 moves to the nucleus and transactivates c-Jun-dependent transcription (34). The existence of multiple mechanisms of integrin-dependent regulation of c-Jun further highlights the importance of this pathway in integrin signaling.

FAK

Early studies on integrin-mediated cell adhesion and signaling demonstrated that cell adhesion to the ECM was accompanied by integrin aggregation, and that this clustering could trigger increased tyrosine phosphorylation of various intracellular proteins (35). As integrins lack intrinsic tyrosine kinase activity, proteins in the integrin/ECM-binding site (focal adhesions) were dissected carefully to identify potential tyrosine kinases that could mediate this signaling event. Interestingly, a predominant protein in focal adhesions, which was shown to undergo rapid tyrosine phosphorylation after integrin ligation and clustering, is a 120-kDa nonreceptor tyrosine kinase known as FAK (25, 36, 37). FAK lacks SH2 and SH3 domains, but contains an N-terminal FERM domain, a central kinase domain, and a C-terminal focal adhesion targeting (FA T) domain (38, 39). Direct association of FAK and integrin cytoplasmic tails has been demonstrated (40). However, the role of this direct interaction remains unclear. The integrin-binding site in FAK was mapped to the region N-terminal to the central kinase domain; however, this region is not required for localization of FAK to focal adhesions. Instead, the C-terminal FAT domain of FAK (43) and also interacts with vinculin and thereby paxillin (42). Thus, the recruitment of FAK to activated integrin clusters could be indirect.

A tempting model has been proposed that, upon recruitment to focal adhesion complexes, FAK undergoes conformational change and interacts through its amino terminal domain with the integrin cytoplasmic tail (43–45). The amino terminal domain of
FAK could fold back onto the kinase domain and play a negative autoregulatory role. Thus, this conformational change may be a prerequisite for FAK’s catalytic activity. The phosphorylation of FAK then initiates a cascade of phosphorylation events and new protein–protein interactions. Phosphoryrosine 397 serves as a binding site for Src family kinases via the SH-2 domain (46–48). Conversely, it has also been shown that tyrosine 925, which creates a binding site for the Grb2–mSOS complex, resulting in activation of MAPK (50, 53).

The two major substrates of SFKs at focal adhesions are p130CAS and paxillin. CAS associates with FAK through the SH3 domain of CAS and the proline-rich motif at FAK C terminus (52). CAS contains a C-terminal proline-rich segment, through which it interacts with the SH3 domain of Src and likely other SFKs, and a large substrate domain, which contains several phosphoryrosine sites mediating the interaction with adaptor proteins, such as Crk and Nck (53). Crk associates with DOCK180 in many cells (54). DOCK180 is the mammalian counterpart of C. elegans gene, ced-5, which is involved in the regulation of phagocytosis and cell migration in nematode (55). Overexpression of DOCK180, along with CAS and Crk, can stimulate membrane ruffling, which is a hallmark of Rac activation. Crk also binds to the GTP exchange factor (GEF) C3G, which activates the Ras-related small G protein Rap1 (56, 57). Rap1 in turn reinforces integrin adhesion to the ECM by an inside-out activation mechanism (58–60). Additionally, because B-Raf is a target of Rap1 and, like other Raf isoforms, promotes signaling to Erk, C3G/Cas signaling at focal adhesions results in activation of Erk in cells that express B-Raf (16).

Paxillin is phosphorylated by SFKs at Tyr 31 and 118 and can recruit Crk as CAS (61). Furthermore, paxillin can associate with Csk, which suppresses SFKs (62), and PTP-PEST (63), which can dephosphorylate CAS (64). Thus, paxillin could potentially participate in a negative feedback loop. Paxillin also associates with p120CA as GAP (GTPase-activating protein) and prevents it from binding to and suppressing p190RhoGAP (65). Therefore, paxillin may inhibit Rho activity during the initial phase of cell adhesion. Furthermore, paxillin is associated with a complex containing the adaptor Pkl, the GEF Pia/Coid, and the Cdc42/Rac target-effector PAK, and it may play a role in linking Cdc42 and Rac to PAK and its downstream targets, such as LIMK and MLCK (66, 67). Lastly, paxillin can recruit the Abl tyrosine kinase from the nucleus to focal adhesions upon integrin ligation (68). At focal adhesions, Abl phosphorylates Ena/VASP proteins, thus regulating actin cytoskeleton (69). Phosphoryrosine 397 of FAK associate with the SH2 domain of p85 subunit of PI3-K (70). This direct interaction could activate PI3-K and thus promote cell survival. The same residue on FAK also mediated association with other proteins containing SH2 domain, including p130 Ras GAP, PLCγ, and GRB7 (71–73). In addition, FAK contains two proline-rich motifs at the C-terminal portion, which mediates interaction with various SH3-domain-containing proteins, such as CAS, Graf, ASAP1, and Endophilin A2 (74–77). N-terminal region of FAK has also been shown to associate with signaling proteins, such as N-WASP (78). Thus, FAK can relay the integrin signaling to affect a wide variety of pathways, which in turn regulate various cellular processes, including growth, survival, differentiation, and migration. In addition to the study on in vivo cell culture systems, the analysis of FAK knockout mice has provided important insight into the biological function of FAK in vivo. Constitutive as well as tissue-specific knockout of FAK have implicated it in many physiological and pathological processes, ranging from embryonic development, angiogenesis, and cortical basement membrane assembly or remodeling to tumorigenesis in skin, and heart hypertrophy (79–84).

ILK

ILK was identified by yeast two-hybrid screen for proteins that could bind to the cytoplasmic tail of integrin (85). The N-terminal domain of ILK contains three ankyrin repeats, which mediate protein interactions, and a putative fourth ankyrin repeat that lacks some conserved residues. The C-terminal portion shares significant sequence homology to Ser/Thr protein kinases. A pleckstrin homology (PH) domain is situated between these two domains. It is now widely accepted that ILK functions as the central component to organize a heterotrimeric protein complex named IPP (ILK, PINCH, and Parvin) complex (86, 87). ILK binds PINCH through the N-terminal ankyrin-repeat domain and parvins through the kinase domain. ILK may also link the IPP complex to the cytoplasmic tails of integrins. The IPP complex functions both as an adaptor between integrins and the actin cytoskeleton and as a hub that regulates various signaling pathways. It remains controversial whether ILK is a real protein kinase. The kinase domain of ILK shows significant homology to Ser/Thr protein kinases, except residues within the catalytic loop and the conserved DFG motif (88). Thus, ILK lacks a conventional catalytic base and Mg2+–chelating residues. However, recombinant, purified ILK has been shown to be able to phosphorylate several substrates in vitro (89–91), including GSK3β and AKT, which regulate many different signaling pathways. It is well established that AKT activation requires phosphorylation of Thr at position 308 by phosphatidylinositol 3-kinase (PI3-K) and this notion is further supported by immunoprecipitation assays showing that ILK directly binds to AKT (90). However, it is not clear whether ILK possesses sufficient activity to function as a physiologically relevant kinase in vivo. The function of the IPP complex as a signaling platform could be achieved mainly through its interaction with other proteins. Many IPP functions are consistent with a role for the IPP complex at focal adhesions, for example, the regulation of podocyte adhesion and spreading (92), platelet aggregation (93, 94), neuronal spreading and outgrowth (95, 96), and leukocyte recruitment (97) all support the notion that the IPP complex regulates actin–cytoskeleton dynamics and integrin activation in focal adhesions. As mentioned, ILK binds directly to the cytoplasmic tails of integrins (98), and it is connected to the actin
cytoskeleton through its interaction with parvins (98). Interactions with the cytoskeleton can also be mediated by the adaptor protein paxillin, which binds to ILK through a paxillin-binding site (PBS) within the kinase domain of ILK. Paxillin binds to F-actin via interactions with α-parvin and the actin-binding adaptor molecule vinculin (61, 98, 99). In addition, ILK binds to the LIM5 domain of PINCH1, but not PINCH2. In addition, PINCH1, but not PINCH2, associates with Thymosin-β1 through the LIM domains -4 and -5 of PINCH1 (106). This interaction can increase ILK activity and positively influences migration and survival of cardiac cells. PINCH1 also binds to NCK2, an adaptor protein in vitro through a LIM4-3H3-domain (Src-homology-3 domain) interaction (107, 108). However, the physiological relevance of this interaction is not clear. α-Parvin can bind to F-actin directly. It also associates with paxillin, which could further bridge it with actin cytoskeleton. HIC5, a paxillin-related protein, also associates with α-parvin (98). Interestingly, HIC5 shuts to the nucleus, where it modulates the expression of several genes (109, 110). Furthermore, α-parvin significantly binds to TESK1, a Ser/Thr kinase that phosphorylates coflin (111). β-Parvin can bind to the actin-cross-linking protein α-actinin and the guanine nucleotide-exchange factor α-Pix (112). It therefore provides a connection between the IPP complex and the Rho family GT-Pases, Rac1, and Cdc42. α-Pix, in turn, binds to PAK1 (114), a Rac/Cdc42 effector that regulates cytoskeletal dynamics. Furthermore, α-Pix can associate with the protease subunit calpain-4 (115). Calpain-4 has been shown to cleave talin, which serves as the rate-limiting step in the disassembly of focal adhesions (116).

Cytoskeleton and Rho Family Small GT-Pase

The initial study from Alan Hall's group described that addition of serum or LPA stimulates the formation of stress fibers and focal adhesion in serum-starved Swiss 3T3 cells, which retain only a few stress fibers and are devoid of focal adhesions (117-120). Later, it was found that the induction of cortical actin assemblies and formation of focal adhesions can be ascribed to the action of different Rho family proteins. Rho family proteins are Ras-related small GT-Pases consisting of Rho A, B, C, D, and E; Rac1, 2, 3, and 6; and Cdc42, Rho G, and TC10 (121). It is now well established that Rho is involved in the organization of focal adhesions and stress fibers; Rac is responsible for the membrane ruffling and extension of lamellipodia; and Cdc42 controls the formation of filopodia (121).

In addition to the growth factor receptors, Rho family members can also be activated by integrins. During adhesion and spreading on an ECM, cells develop filopodia and lamellipodia, which are regulated by Cdc42 and Rac, respectively. Integrin-mediated adhesion can activate Cdc42 and Rac (122). Like other small GT-Pases, Rho family protein is active when GDP-bound and inactive when GTP-bound. Activation of Rho proteins is catalyzed by GEFs and inactivation is induced by GAPs that stimulate the intrinsic GT-Pase activity of the Rho proteins. Vav2, a hematopoietic cell-specific GEFT, is activated upon integrin engagement (123). However, the closely related GEFT Vav3, which is widely distributed, is activated downstream from growth factor receptors but not from integrins (124, 125). Nevertheless, a dominant-negative mutant of Vav2 blocked lamellipodium formation and spreading on fibronectin, suggesting that Vav2 plays a role in Rac activation after integrin engagement (126). Furthermore, integrins may activate Rac through other pathways. P130cas and paxillin associate with FAK, and both of them have been linked to Rac activation. As mentioned above, tyrosine phosphorylation of p130cas promotes the formation of the focal complexes consisting of Crk, DOCK180, and ELMO (127). DOCK180 can function as a Rac GEF, even though it lacks the Dbl-homology/pleckstrin-homology tandem domains characteristic of conventional Rho family GEFTs (127, 128). Paxillin associates with another complex consisting of PKL/GIT and Pak-interacting exchange factor (PIX), the latter being a conventional Rac GEF (161). Interestingly, recent work demonstrated that, in addition to their effect on GTP loading, integrins independently control the translocation of active Rac to the plasma membrane (129). This step is required for Rac binding to its downstream effectors. Integrins can increase membrane affinity for Rac, leading to RhogD1 disassociation and effector coupling locally, in the vicinity of integrins. Furthermore, integrin-regulated Rac binding sites are within cholesterol-enriched membrane microdomains or lipid rafts. Integrins control Rac signaling by preventing the internalization of Rac-binding sites in lipid rafts. Regulation of Rho A activity by integrins is more complicated. It has been shown that integrin engagement leads to a transient depression in RhoA activity, followed by activation (130). The depression could play an essential role in lamellipodial extension during cell migration (131). The mechanism for this depression is not completely understood, but it may involve Src, FAK, paxillin, and p190RhoGAP (64, 65, 132, 133). However, different integrins may initiate different response of Rho A activity. Engagement of α6β4 resulted in stimulation of RhoA activity, in contrast to the depression induced by clustering of β1 integrins (134). Overexpression of β3 integrin in CHO cells resulted in a pronounced increase in Rho-GTP levels when the cells were plated on fibronectin or fibrinogen, whereas β1 integrin overexpression had no effect (135). In contrast, re-expression of β1 subunits in β1-deficient cells stimulated RhoA activity, whereas β3 had no effect (136). This controversy could be due to the differences of cell types used in these two studies. Rho family proteins serve as important mediators for integrin outside-in signaling, integrin clustering and focal complex

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

- 1716 -
Regulation of Integrin Through Inside-Out Signaling, a Structural View

As the signals that are transduced by integrins from the outside of the cell to the inside were being elaborated, parallel investigations on the regulation of integrin affinity revealed that information also flowed in the opposite direction—from integrin cytoplasmic tails to the extracellular ligand-binding domain, usually named the “inside-out” signaling of integrin. This process was actually described even before the molecular definition of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.

Earlier studies using biophysical approaches and antibodies that specifically recognize the activated state of integrins demonstrated that their activation involves alterations in integrin conformation (168-172) and that the cytoplasmic tails of leukocyte adhesion receptors were subjected to a similar regulation of integrins. Bennett and Vilaire found that the binding of platelets to fibrinogen is subject to regulation by agonists (165). It was then shown that this regulation was not dependent on the recruitment of receptors to the surface, rather on an increase in the binding activity of the receptor, and that leukocyte adhesion receptors were subjected to a similar regulation of ligand-binding affinity (166, 167). After cloning of integrins, it became clear that most leukocyte and platelet integrins, including β₂-containing integrins, exist in a resting state until activated, and that a variety of agonists could regulate this process (4, 6). The physiological relevance of this regulation is obvious. It prevents the spontaneous adhesion of platelets and leukocytes within the circulation or to the blood vessel wall, but allows the rapid response upon injury.
Therefore, although 

conserved phosphotyrosine-binding domain (PTB)-like interac-
tion (176). Overexpression of talin fragments containing this
PTB-like domain activates integrins, whereas knocking down
talin expression with small interfering RNAs or sequestration of
talin blocks α1 and β3 integrin activation in different cells (175,
177). Furthermore, mutations in conserved residues in talin or
the integrin β tail that disrupt their interaction prevent activa-
tion (175). Thus, the binding of the talin PTB-like domain to the
integrin β tail serves as a common step in integrin activation.

The conserved membrane-proximal portions of integrin α and
β subunits and cytoplasmic tails control integrin activation. Mutation or trun-
cation of this region results in constitutive integrin activation
(178). Furthermore, replacement of the region by heterodimeric
coiled-coil peptides or a covalent cross-linking of the tails in-
activates integrins, whereas breaking the coiled-coil or the co-
valent bond activates them (179). It has been proposed that
the membrane-proximal regions of α and β cytoplasmic tails
associate through a salt bridge between a conserved Arg in
the α tail and an Asp in the β tail, thereby keeping the in-
tegrin heterodimer in an inactive state (178). The talin head
domain can disrupt the salt bridge formed between α1β1-Arg596
and β3-Asp723 (180) (Figure 3). Therefore, talin binding may
activate an integrin by disrupting a “clasp” between the α and
β cytoplasmic tails, leading to tail separation and integrin acti-
vation. However, this clasp has not been consistently observed
(179) in structural studies of isolated integrin cytoplasmic do-
 mains, and the interaction between tails was of low affinity.
Therefore, although α and β tail interactions may occur via a
salt bridge, the strength of this interaction appears to be modest.

In addition to the “clasp” model, a “piston” model has been
proposed. In this model, it is the displacement of the membrane-
proximal domain from the membrane that leads to integrin activation by shortening of the TM domain. Integrin cytoplasmic domains contain remarkably conserved sequences consisting of a Trp-Lys (Arg) doublet that is predicted to terminate the TM sequence. This doublet is followed by the membrane-proximal region, a hydrophobic stretch of four or five residues terminated by strongly polar residues, which are hotspots for mutations that activate integrins (181, 182). It has been shown that fragments of talin that activate integrins perturb NMR resonances in the membrane-proximal region, whereas a subdomain of talin that binds to the tail but fails to activate

ulib3 perturbs the NMR resonances to a much lesser extent. In
addition, the ulib cytoplasmic tail, which blocks talin interaction
with the membrane-proximal region of the β3 tail, prevents
ulib3 activation (183). Glycosylation mapping has indicated
that the membrane-proximal domains of the α and β subunits
can reside within the membrane and that certain activating
mutations in this region can displace them from the membrane,
thereby shortening the TM domains (184, 185). Various targeted
and random activating mutants of integrin ulib3 would be
predicted to shorten the TM domain in a similar manner (173,
181, 182, 186). However, studies with isolated ulib and β3
peptides and liposomes suggest that an upward movement
of membrane-proximal helices upon talin binding could contribute
to integrin activation (187). Resolution of this controversy will
require further structural analysis of integrin.

Propagation of the inside-out activation signal across the
plasma membrane is mediated by the TM helices. The TM he-
lies of the inactive, but not active, integrin show a periodic
disulfide cross-linking pattern (188), suggesting that the TM he-
lies specifically interact in the inactive state and that this inter-
action is disrupted upon activation. Many membrane-embedded
activating mutations identified by random and site-directed mu-
tagenesis studies further support this notion (181, 189, 191).
However, the molecular details of these interactions are not
fully understood. Furthermore, CFP and YFP fluorophores fused
to the C-terminus of α and β subunits showed decreased fluores-
cence resonance energy transfer (FRET) upon integrin activation
by a variety of stimuli (191), implying cytoplasmic domain sep-
oration during integrin activation. So the cytoplasmic domain
separation, and therefore TM separation, might be the critical
step in activation. However, another possibility cannot be ruled
out—that mutations of the TM domain might lead to altered ori-
teation of the TM domains instead of actual separation; such
an alteration of orientation is also consistent with the reported
FRET results (191).

A number of key questions concerning the integrin activa-
tion remain unresolved. Although we know that the membrane-
proximal region and TM region of integrins controls the acti-
vation, other mutagenesis results indicate that the C-terminus
membrane distal region of the α or β cytoplasmic tails is also
important in regulating integrin activation via a mechanism that
is yet unknown. Thus, the whole picture for the inside-out acti-
vation of integrins can be substantially more complicated. In addition, there
may exist other proteins besides talin that bind to the cytoplas-
mic tails and that regulate the conformational change required
for integrin activation. The current “clasp” or “piston” model
explains how the alterations of the cytoplasmic tail and TM
region of integrins relieve the structural constraint and allow
the unbending of the extracellular domain to attain the high
affinity ligand-binding state (192). However, a thorough molecu-
lar understanding of integrin inside-out signaling awaits further
structural analysis of the intact receptor in inactive and active
states.
Concluding Remarks

Recent years have witnessed exciting progress in understanding integrin signaling, owing to structural analyses and characterization of proteins that interact with integrins. The role of integrins in vivo has also been revealed by the functional studies of gene-targeted mice (193). Interestingly, despite the fact that the binding of integrin-binding proteins involves a loss of almost any integrin subunit to defects with varying severity in knockout mice, providing the strongest evidence for biological relevance of integrin signaling. However, many molecular details of integrin signaling remain elusive. A number of successful approaches have been carried out to study the integrin interactions at the molecular level. However, findings over the past two decades in this field have led to the development of clinically useful or promising integrin antagonists combating diseases such as cancer, cardiovascular defect, and inflammatory diseases (194). Therefore, a thorough understanding of integrin signaling is likely to lead to even greater opportunities for novel therapies.

References

Integrin Signaling, Bidirectional Signaling of


102. Tsuda T, Mariniott M, Maozil L, Cutler ML. The Ras sup- pressor RAS-I localizes to 10p13 and its expression in the U251...
Integrin Signaling, Bidirectional-Signaling of


Integrin Signaling, Bidirectional Signaling of


Integrin Signaling, Bidirectional-Signaling of


The isoprostanes are a unique series of prostaglandin-like compounds formed in vivo via a nonenzymatic mechanism involving the free radical-initiated peroxidation of arachidonic acid. This article summarizes our current knowledge of these compounds. Herein, a historical account of their discovery and the mechanism of their formation are described. Methods by which these compounds can be analyzed and quantified are also discussed, and the use of these molecules as biomarkers of in vivo oxidant stress is summarized. In addition to being accurate indices of lipid peroxidation, some isoprostanes possess potent biological activity. This activity will be discussed in detail. Finally, in more recent years, isoprostane-like compounds have been shown to be formed from polyunsaturated fatty acids, including eicosapentaenoic acid and docosahexaenoic acid. These findings will be summarized as well.

Free radicals, largely derived from molecular oxygen, have been implicated in a variety of human conditions and diseases, including atherosclerosis and associated risk factors, cancer, neurodegenerative diseases, and aging. Damage to tissue biomolecules by free radicals is postulated to contribute significantly to the pathophysiology of oxidative stress. Measuring oxidative stress in humans requires accurate quantification of either free radicals or damaged biomolecules. The targets of free radical-mediated oxidant injury include lipids, proteins, and DNA. Several methods exist to quantify free radicals and their oxidation products, although many of these techniques suffer from lack of sensitivity and specificity, especially when used to assess oxidant stress status in vivo. In a recent multi-investigator study, termed the Biomarkers of Oxidative Stress Study (BOSS), sponsored by the National Institutes of Health, it was found that the most accurate method to assess in vivo oxidant stress is the quantification of plasma and urinary isoprostanes (IsoPs) (1). IsoPs, a series of prostaglandin (PG)-like compounds produced by the free radical-catalyzed peroxidation of arachidonic acid independent of the cyclooxygenase, were first discovered by our laboratory in 1990 (2). Since that time, we and others have shown that levels of IsoPs are increased in several human diseases. Furthermore, several of these compounds possess potent biological activity and thus may be mediators of oxidant injury. In recent years, additional related compounds, derived from various polyunsaturated fatty acids, such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), have been discovered to be formed as products of the IsoP pathway. It is the purpose herein to summarize our current knowledge regarding the IsoPs, including their discovery, the chemical biology of their formation, the utility of these compounds as markers of in vivo oxidant stress, and their bioactivity.

### Historical Aspects

In the 1960s and 1970s, it was shown that PG-like compounds could be formed by the autoxidation of purified polyunsaturated fatty acids. Seminal studies by Pryor, Porter, and others (see further reading) led to a proposed mechanism by which these compounds are generated via bicycloendoperoxide intermediates. However, this work was never carried beyond in vitro studies. Furthermore, it was not determined whether PG-like compounds could be formed in biological fluids containing unsaturated fatty acids.

In the 1980s, we showed that PGD_2_ derived from the cyclooxygenase is primarily metabolized in vivo in humans to form 9,11PGD_2_ by the enzyme 11-ketoreductase. In aqueous solutions, however, PGD_2_ is an unstable compound that undergoes isomerization of the lower side chain and these isomers can likewise be reduced by 11-ketoreductase to yield isomers of 9u,11u-PGF_2_. In studies undertaken to further characterize these compounds using gas chromatography (GC)/mass spectrometry (MS), we found that when plasma samples from normal volunteers were processed and analyzed immediately, a series of peaks was detected possessing characteristics of F-ring PGs. Interestingly, however, when plasma samples that had been stored at −20°C for several months were reanalyzed, identical chromatographic peaks were detected but levels of putative PGF_2-like compounds were up to 100-fold higher. Subsequent experiments led to the finding that these PGF_2-like compounds were generated in both freshly processed and stored plasma, not by a COX-derived mechanism, but nonenzymatically by autoxidation of arachidonic acid. Because these compounds contain F-type prostanate rings, they are referred to as F_2-IsoPs.
Formation of the F2-Isoprostanes

A mechanism to explain the formation of the F2-IsoPs from arachidonic acid is outlined in Fig. 1 and is based on that proposed by Pryor for the generation of bicycloendoperoxide intermediates. After abstraction of a bisallylic hydrogen atom and the addition of a molecule of oxygen to arachidonic acid, the peroxyl radical undergoes 5-exo cyclization and a second molecule of oxygen adds to the backbone of the compound to form PGG2-like compounds. These unstable bicycloendoperoxide intermediates are then reduced to the F2-IsoPs. Based on this mechanism of formation, four F2-IsoP regioisomers are generated. Compounds are denoted as 5-, 12-, 8-, or 15-series regioisomers depending on the carbon atom to which the side-chain hydroxyl is attached. An alternative nomenclature system for the IsoPs has been proposed by Rokach et al. in which the abbreviation iP is used for isoprostane and the regioisomers are denoted as III–VI based on the number of carbons between the omega carbon and the first double bond.

Although the initial abstraction of any bisallylic hydrogen atoms of arachidonic acid is equally likely, the different IsoP regioisomers are not formed in equal amounts. When arachidonic acid is oxidized, the 5- and 15-series regioisomers are formed in significantly higher amounts than the 8- and 12-series regioisomers. An explanation for this difference has been elucidated by Yin et al., who demonstrated that the arachidonyl hydroperoxides which give rise to the 8- and 12-series regioisomers readily undergo further oxidation to yield a newly discovered class of compounds that contain both bicycloendoperoxide and cyclic peroxide moieties termed dioxolane-IsoPs (Fig. 2). Overall, 5- and 15-series regioisomers cannot undergo this further oxidation but instead accumulate at higher concentrations in tissues and fluids. A method allowing the separation and quantification of each of the four classes of IsoP regioisomers, as well as their individual stereoisomers, has recently been developed in our laboratory.

An important structural distinction between IsoPs and cyclooxygenase-derived PGs is that the former contain side chains that are predominantly oriented cis to the prostane ring, whereas the latter possess exclusively trans side chains. A second important difference between IsoPs and PGs is that IsoPs are formed in situ esterified to phospholipids and are subsequently released by a phospholipase(s), whereas PGs are generated only from free arachidonic acid.

Quantification of F2-Isoprostanes

Several methods have been developed to quantify the F2-IsoPs. Our laboratory uses a gas chromatographic/negative ion chemical ionization–mass spectrometric (GC/NICI-MS) approach employing stable isotope dilution. For quantification purposes, we measure the F2-IsoP, 15-F2-IsoP, and other F2-IsoPs that coelute with this compound. Several internal standards are available from commercial sources to quantify the IsoPs. In our
Isoprostanes

assays, we typically use either $[^{2}H_4]$-15-F$_2$-IsoP ($[^{2}H_4]$-8-iso-PGF$_{2\alpha}$) or $[^{2}H_4]$-PGF$_{2\alpha}$ as internal standards. The advantages of mass spectrometry over other approaches include its high sensitivity and specificity, which yields quantitative results in the low picogram range. Its drawbacks are that it is labor intensive and requires considerable expenditures on equipment.

Several alternative mass spectrometric assays have been developed by different investigators, including Rokach et al. (4, 7, 8). Like our assay, these methods quantify F$_2$-IsoPs using stable isotope dilution GC/MS and require solid-phase extraction using a C18 column, thin layer chromatography (TLC) purification, and chemical derivatization. These assays, however, measure F$_2$-IsoP isomers other than 15-F$_2$-IsoP, but they are comparable with ours in terms of utility. In addition to these GC/MS assays, several liquid chromatographic (LC) MS methods for F$_2$-IsoPs have been developed. One advantage of LC/MS methods is that the sample preparation for analysis is simpler than that for GC/MS because no derivatization of the molecule is required. The method reported earlier this year by Taylor et al. is the first of these LC/MS methods to be validated for quantitation of 15-F$_2$-IsoPs in biological fluids (9). In this assay, a gradient reverse-phase LC tandem mass spectrometric approach is used to separate several 15-series IsoP stereoisomers for quantitation with $[^{2}H_4]$-PGF$_{2\alpha}$ as the internal standard. The outcomes of this assay correlated significantly with GC/MS results, and the coefficients of variation in the measurements were lower for their LC/MS assay than for their GC/MS assay. The authors thus suggest that this LC/MS method potentially offers greater precision than existing methods while allowing for the quantitation of more compounds with simpler sample preparation.

Alternative methods have also been developed to quantify IsoPs using immunological approaches (6, 10). A number of antibodies have been generated against 15-F$_2$-IsoP, and at least three immunoassay kits are commercially available. Although mass spectrometric methods of IsoP quantification are considered the best methods for analysis, immunoassays have expanded research in this area due to their low cost and relative ease of use. Only limited information is currently available regarding the precision and accuracy of immunoassays. In addition, little data exist comparing IsoP levels determined by immunoassay with MS, though Wang et al. offers one example of an MS-validated immunoassay (11) out laboratory’s experiments have not validated the commercial available kits. Analogous to immunological methods to quantify cyclooxygenase-derived PGs, immunoassays for IsoPs suffer from a lack of specificity. Furthermore, the sensitivity and/or specificity of these kits vary substantially between manufacturers.

**F$_2$-Isoprostanes as an Index of Oxidant Stress In Vivo**

The true utility of the F$_2$-IsoPs is in the quantification of lipid peroxidation and thus oxidant stress status in vivo. F$_2$-IsoPs are stable, robust molecules and are detectable in all human tissues and biological fluids analyzed, including plasma, urine, bronchoalveolar lavage fluid, cerebrospinal fluid, and bile (6). The quantification of F$_2$-IsoPs in urine and plasma, however, is most convenient and least invasive. And, based on available data, quantification of these compounds in either plasma or urine is representative of their endogenous production and thus gives a highly precise and accurate index of in vivo oxidant stress. Although measurement of F$_2$-IsoPs in plasma is indicative of their endogenous formation at a specific point in time, analysis of these compounds in urine is an index of systemic or “whole-body” oxidant stress integrated over time.
is important to note that the measurement of free F2-IsoPs in urine can be confounded by the potential contribution of local IsoP production in the kidney. In light of this issue, we have identified the primary urinary metabolite of 15-F2t-IsoP to be 2,3-dinor-5,6-dihydro-15-F2t-IsoP, and we have developed a highly sensitive and accurate mass spectrometric assay to quantify this molecule. However, the extent to which 15-F2t-IsoP is converted to the urinary metabolite remains unclear. Nevertheless, the quantification of 2,3-dinor-5,6-dihydro-15-F2t-IsoP may represent a truly noninvasive, time-integrated measurement of systemic oxidation status that can be applied to living subjects.

Normal levels of F2-IsoPs in healthy humans have been defined (6, 12, 13). Defining these levels is particularly important in that it allows for an assessment of the effects of diseases on endogenous oxidant tone and allows for the determination of the extent to which various therapeutic interventions affect levels of oxidant stress. Elevations of IsoPs in human body fluids and tissues have been found in a diverse array of human disorders, some of which include atherosclerosis, hypercholesterolemia, diabetes, obesity, cigarette smoking, neurodegenerative diseases, and rheumatoid arthritis. Furthermore treatments for some of these conditions, including antioxidant supplementation, treatments, cessation of smoking, and even weight loss, have been shown to decrease production of F2-IsoPs. Thus, the clinical utility of F2-IsoPs has been great and continues to grow. For manuscripts and comprehensive reviews on IsoP detection in human disease, please refer to the Further Reading Section.

**Biological Activities of the F2-IsoP Prostanoids**

In addition to being robust markers of tissue oxidant stress, F2-IsoPs can exert potent biological activity and potentially mediate some adverse effects of oxidant injury. As mentioned, IsoPs are initially formed in vivo esterified to glycerophospholipids. Molecular modeling of IsoP-containing phospholipids reveals them to be remarkably distorted molecules. Thus, the formation of these abnormal phospholipids would be expected to exert profound effects on membrane fluidity and integrity, well-known sequelae of oxidant injury. All studies exploring their bioactivity, however, have been performed using unesterified IsoPs. Recent studies by Stafforini et al. have shown that F2-IsoP hydrolysis from phospholipids is regulated, at least in part, by the platelet-activating factor (PAF) acylethanolamides I and II (14).

One particular F2-IsoP that is produced abundantly in vivo and has been extensively tested for biological activity is 15-F2t-IsoP (8-iso-PGF2α), which differs from cyclooxygenase-derived PGF2α, or in the inversion of the upper side-chain stereochemistry. This IsoP has been found to be a potent vasoconstrictor in a variety of vascular beds, including the kidney, lung, heart, and brain. In addition, 15-F2t-IsoP induces endothelin release and proliferation of vascular smooth muscle cells. There is also additional evidence that this molecule can increase resistance to aspirin inhibition of platelet aggregation in platelets as well as inhibit platelet aggregation in human whole blood. These vasoactive effects of 15-F2t-IsoP have been shown to result from interaction with the thromboxane receptor, a G-protein-coupled transmembrane eicosanoid receptor, based on the finding that these effects can be abrogated by thromboxane receptor antagonists.

The testing of other F2-IsoPs for biological activity has been limited. It has been shown, however, that 15-F2t-IsoP (12-iso-PGF2α) activates the PGF2α receptor and induces hypertropy in cardiac smooth muscle cells.

**Formation of Isoprostanes with Alternative Ring Structures**

Since the initial discovery of the F2-IsoPs, our laboratory has shown that the IsoP pathway provides a mechanism for the generation of various classes of IsoPs from arachidonic acid, which differ in regard to the functional groups on the prostane ring. In addition to undergoing reduction to yield F2-IsoPs, the arachidonyl endoperoxide intermediate can undergo isomerization to yield E- and D-ring IsoPs (Fig. 3), which are isomeric to PGE2 and PGD2, respectively. 15 E2/D2-IsoPs are formed competitively with F2-IsoPs, and recent studies have demonstrated that the depletion of cellular reducing agents, such as glutathione (GSH) or α-tocopherol, favors the formation of E2/D2-IsoPs over that of reduced F2-IsoPs. Importantly, depletion of GSH and α-tocopherol occurs in various human tissues under conditions of oxidant injury, including the brains of patients with Alzheimer’s disease (AD). Thus, the ratios of F-ring to E/D-ring IsoPs in postmortem brain tissues from patients with AD were examined, and not only were levels of both E2/D2- and F2-IsoPs significantly elevated, but also E2/D2-IsoPs were the favored products of the IsoP pathway in affected brain regions. This increased ratio of E2/D2-IsoPs to F2-IsoPs provides information not only about lipid peroxidation in a given organ but also about the reducing environment in that tissue.

E2/D2-IsoPs, however, are not terminal products of the IsoP pathway. These compounds readily dehydrate in vivo to yield A2J2-IsoPs (Fig. 3), which are also known as cyclopentenone IsoPs because they contain an α,β-unsaturated cyclopentenone ring structure (16, 17). A2J2-IsoPs are highly reactive electrophiles that readily form Michael adducts with cellular thiols, including those found on cysteine residues in proteins and glutathione. These cyclopentenone IsoPs are rapidly metabolized in vivo by glutathione transferase enzymes to water-soluble modified glutathione conjugates. The major urinary cyclopentenone IsoP metabolite in rats, a 15-A2J2-IsoP mercapturic acid sulfoxide conjugate, was recently identified in our laboratory 18.

The chemical reactivity of cyclopentenone IsoPs suggested that these compounds might be biologically active. The recent synthesis of two cyclopentenone IsoPs regioisomers, 15-A1-isoP and 15-J1-isoP, has allowed us to examine their bioactivity. Studies employing primary cortical neuronal cultures demonstrated that both 15-A1-isoP and 15-J1-isoP potently induce neuronal apoptosis and exacerbate neurodegeneration caused by other insults at concentrations as low as 100nM (19).
veal that 15-A2-IsoPs potently suppress lipopolysacharide (LPS)-induced inflammatory signaling via inhibition of the NF-κB pathway (20). 15-A2-IsoPs abrogate inducible nitric oxide synthase and cyclooxygenase-2 expression in response to LPS, as well as the elaboration of several pro-inflammatory cytokines. Similar anti-inflammatory effects were observed with 15-J2-IsoPs. However, 15-J2-IsoPs also activate the peroxisome proliferators activated receptor-gamma (PPARγ) with an EC50 of approximately 3 μM. This receptor modulates a wide variety of biological processes, including inflammatory signaling and fatty acid metabolism. 15-J2-IsoPs also induce macrophage apoptosis at low micromolar concentrations in a PPARγ-independent manner. Thus, there is a diversity of actions among cyclopentenone IsoP isomers. It seems that these compounds could act as negative-feedback regulators of the inflammatory response, because oxidative stress and lipid peroxidation often occur under conditions of chronic inflammation.

Formation of Isoprostanes from Other PUFAs

A rachidonic acid is not the only polyunsaturated fatty acid that can be oxidized to generate IsoPs. The basic requirement for cyclization to occur is the presence of at least three double bonds. F-ring IsoPs have been shown to be generated from the peroxidation of linolenic acid [C18:3, ω-3, F3-IsoPs] (21), EPA [C20:5, ω-3, F5-IsoPs] (22), and DHA [C22:6, ω-3, F6-IsoPs] or F4-neuroprostanes (NPs) (23). Similar to the distribution of F2-IsoP regioisomers, certain F3-IsoP and F4-NP regioisomers are more abundant than others (24). The 5-series and 18-series F3-IsoPs are the most abundant because the 8-, 12-, and 15-series compounds can further oxidize (unpublished data). Analogously, the 4-series and 20-series F4-NPs are formed in the largest amounts. In addition to F-ring compounds, E- and D-ring as well as A- and J-ring, or cyclopentenone, IsoPs are generated from the oxidation of EPA (unpublished data) and DHA.

In recent years, emerging evidence has implicated increased dietary intake of fish oil, which contains large amounts of EPA and DHA, as being beneficial in the prevention and treatment of several diseases, including atherosclerotic cardiovascular disease and sudden death, neurodegeneration, and various inflammatory disorders. Furthermore, recent data have suggested that the anti-inflammatory effects and other biologically relevant properties of ω-3 fatty acids are due, in part, to the generation of various bioactive oxidation products (25, 26). We thus hypothesized that EPA- and DHA-derived IsoPs could contribute to the beneficial biological effects of fish oil supplementation. Indeed, one report states that the EPA-derived IsoP, 15-F2t-IsoP, possesses activity that is different from 15-F2t-IsoP in that it does not affect human platelet shape change or aggregation. The lack of activity of 15-F2t-IsoP is consistent with observations regarding EPA-derived PGs in that these latter compounds exert either a weaker agonist or no effects in comparison with arachidonic acid-derived PGs.

Interestingly, our laboratory has recently shown that the levels of these compounds generated from the oxidation of EPA significantly exceeded those of F2t-IsoPs generated from arachidonic acid, perhaps because EPA contains more double bonds and is therefore more easily oxidizable. Additionally, in vivo in mice, levels of F2t-IsoPs in tissues such as heart were virtually undetectable at baseline but supplementation of animals with EPA markedly increased quantities up to 27.4 ± 5.6 ng/g heart. But, of particular note, we found that EPA supplementation markedly reduced levels of arachidonate-derived F2t-IsoPs by up to 64% (p < 0.05). Furthermore, in a recent small clinical trial, termed the K.A.N.W.U. study, of 162 healthy men and women, plasma F2t-IsoP levels significantly decreased (by as much as 29%) after 3 months of supplementation with 3.6-g/day fish oil (27). Together these observations are significant because F2t-IsoPs are generally considered to be pro-inflammatory molecules associated with the pathophysiological sequelae of oxidant stress (28–31). It is thus intriguing to propose that part of the mechanism by which EPA prevents certain diseases is its ability to decrease F2t-IsoP generation. In addition, it suggests that supplementation with fish oil may be of benefit to populations associated with increased levels of F2t-IsoPs.

Conclusions

The discovery of the IsoPs as products of nonenzymatic lipid peroxidation has been a major breakthrough in the field of free radical research. The quantification of these molecules has opened up new areas of investigation regarding the role of free...
radicals in human physiology and pathophysiology, and it seems to be the most useful tool currently available to explore the role of lipid peroxidation in the pathogenesis of human disease. Our understanding of the IsoP pathway continues to expand, providing new insights into the nature of lipid peroxidation in vivo, and revealing new molecules that exert potent biological actions and might serve as unique indices of disease. Basic research into the biochemistry and pharmacology of the IsoPs, coupled with clinical studies employing these molecules as biomarkers, should continue to provide important insights into the role of oxidant stress in human disease.

References


Further Reading

Free Radicals and Disease

Mechanisms of Lipid Peroxidation

F2-Isoprostanes: Formation and Metabolism

Isoprostanes as Biomarkers of Oxidative Stress

Neurodegeneration

Atherosclerosis

Isoprostanes
Isoprostanes


D- and E-ring Eicosanoids: Formation and Metabolism


Cyclopentenone (A-and D-ring) Eicosanoids: Formation and Metabolism

**Isoprostanes**


**Omega-3 Fatty Acids**


See Also

Prostaglandins

Chromatography of Lipids

GC-MS of Lipids

Redox Regulation and Signaling: Reactive Oxygen Species (ROS)

Chemistry of Oxidative DNA Damage
Large G-Proteins

David P. Siderovski, Adam J. Kimple and Francis S. Willard, University of North Carolina at Chapel Hill, North Carolina
doi: 10.1002/9780470048672.wecb219

Large G-proteins, also known as heterotrimeric G-proteins, cycle through inactive (GDP-bound) and active (GTP-bound) states, thereby coupling the activation of cell-surface receptors to the modulation of various second messengers and intracellular effector proteins. High-resolution structures of the constituent subunits of the heterotrimer (G\text{α}, G\text{β}, and G\text{γ}) have revealed the determinants involved in heterotrimer assembly and effector protein engagement as well as the mechanism of nucleotide hydrolysis by the G\text{α} subunit and its acceleration by “regulator of G-protein signaling” (RGS) proteins. Mechanistic details have also been gleaned from studies of G-protein modulation by pathogen toxins and chemicals such as aluminum tetrafluoride. Recent identification of peptide-binding “hotspots” on G-protein subunits should facilitate discovery and design of new chemical entities that can modulate nucleotide cycling and receptor/effector coupling at the level of the G-protein heterotrimer.

A wide variety of extracellular signaling molecules, such as hormones, neurotransmitters, growth factors, tastants, and odorants, communicate their information via binding and activating cell membrane-bound receptors and thus eliciting changes in various intracellular processes. The largest class of such receptors, the superfamily of seven-transmembrane, G protein-coupled receptors (GPCRs), has been an attractive target historically for discovery of small molecule therapeutics. GPCRs remain the largest single fraction of the druggable proteome, with GPCR-targeted drugs continuing to have annual sales in the tens of billions of dollars worldwide [1]. Thus considerable effort has been made over the past few decades to establish a complete mechanistic understanding of how GPCRs communicate extracellular signals into the cell, in the hopes of elucidating the mechanism of action of existing therapeutics as well as facilitating new modalities of GPCR-directed drug discovery/design (beyond the receptor/ligand binding interface per se). The precise structural determinants of GPCR-mediated G-protein activation on binding activating ligand remain enigmatic; however, the mechanisms by which heterotrimeric G-proteins couple receptor activation to modulation of intracellular processes are now known in great detail.

Background

As their name suggests, G protein-coupled receptors link ligand binding to intracellular changes functionally via their engagement of three distinct proteins (G\text{α}, G\text{β}, and G\text{γ}) known collectively as the G-protein heterotrimer [Fig. 3]. The G\text{α} subunit of the heterotrimer is the bona fide G-protein that binds guanine nucleotides in one of two states: guanosine 5′-diphosphate (GDP) when in the inactive, heterotrimeric state (i.e., complexed with G\text{β} and G\text{γ}) and guanosine 5′-triphosphate (GTP) when in the activated state. An activating ligand binds its respective GPCR and changes the conformation of the receptor’s transmembrane regions and intracytosolic loops to activate the receptor’s guanine nucleotide exchange factor (GEF) activity on the G\text{α} GDP/G\text{βγ} complex. Receptor-catalyzed exchange of GTP for the bound GDP within G\text{α} leads to release of bound G\text{βγ}, and subsequent, engagement of a- and bγ-effector proteins by GTP-bound G\text{α} and the now-freed G\text{βγ}-heterodimer.

Four general classes of G\text{α} subunits have been defined based on their functional couplings (in the GTP-bound state) to various effectors [Fig. 4]. G\text{α}-subfamily G\text{α} subunits are stimulatory to membrane-bound adenylyl cyclases that generate the second messenger 3′-5′-cyclic adenosine monophosphate (cAMP); cellular adenylyl cyclase activity can be stimulated directly (i.e., in the absence of G\text{βγ} heterotrimer activation) by forskolin, a plant-derived diterpene. Conversely, G\text{βγ}-subfamily G\text{α} subunits generally are inhibitory to adenylyl cyclases. G\text{βγ}-subfamily G\text{α} subunits are potent activators of phospholipase-C\text{β} enzymes that catalyze breakdown of the cell membrane lipid constituent phosphatidylinositol-4,5-bisphosphate into the second messengers diacylglycerol and inositol triphosphate, which leads to transient increases in intracellular calcium content. The G\text{βγ} subunits released by activated G\text{βγ} heterotrimers are also known to stimulate the activity of phospholipase-C\text{ζ} enzymes, as well as to modulate adenylyl cyclase activity, to activate phosphatidylinositol-3′ kinase and inward-rectifying potassium channels, and to inhibit calcium channel current (reviewed in...
Reference 2). G_{12/13}-subfamily G_{α} subunits activate the small G-protein RhoA through stimulation of RhoA-specific GEFs such as p115-RhoGEF, LARG, and PDZ-RhoGEF.

G_{α} subunits from all four subfamilies have the intrinsic ability to hydrolyze bound GTP to GDP and inorganic phosphate (Fig. 1). This intrinsic GTPase activity can be accelerated dramatically by the GTPase-accelerating proteins (GAPs) specific to G_{α} subunits: namely, the “regulator of G-protein signaling” (RGS) proteins (3). α–Effectors can also exhibit GAP activity [e.g., phospholipase-C_{β} activity in accelerating GTP hydrolysis by G_{α}q (4)]. GTP hydrolysis controls the timing of signal duration, as the loss of the third phosphate reverts the G_{α} back to the GDP-bound state with characteristic high affinity for G_{βγ} and low affinity for α–effectors. Hence, the root mechanism for coupling receptor activation to the modulation of intracellular processes is a cycle of GTP binding and GTP hydrolysis transacted by the G_{α} subunit of the G-protein heterotrimer.

Structures

X-ray diffraction crystallography has been the main technique employed to establish the structures of GDP-bound G-proteins (both as isolated G_{α} subunits and G_{αβγ} heterotrimeric complexes; Table 1), as well as activated states of G_{α} induced either by the binding of the nonhydrolyzable GTP analog guanosine 5'-γ-thiotriphosphate (GTPγS) or by the addition of aluminum tetrafluoride (AlF_{4}–) to GDP-bound G_{α} subunits. These efforts have unveiled the precise secondary and tertiary structures of both G_{α} and G_{βγ} (Figs. 2 and 3) (5), how the heterotrimeric complex is formed by these two binding partners, the conformational changes that are induced within G_{α} by the exchange of GDP for GTP (Fig. 2b), and the mechanisms of both intrinsic (Fig. 4) and RGS protein-accelerated GTP hydrolysis (6–8).

G_{α}

The G_{α} subunit, in its inactive state, binds GDP within a nucleotide-binding pocket circumscribed by residues derived...
brane targeting as well as assembly with $G_{\beta\gamma}$ fatty acids myristate and/or palmitate, which facilitates $G_{\alpha}$ N-terminal $\alpha$-helix and $G_{\alpha}$ N-terminal helix is modified by covalent attachment of the fatty acids myristate and/or palmitate, which facilitates mem-

Active conformation of $G_{\alpha}$ subunits (bound to GTP$\gamma$)

Transition-state conformations of $G_{\alpha}$ subunits (bound to GDP, $G_{\alpha}$)

Activated $G_{\alpha}$ bound to adenyl cyclase and activated $G_{\alpha}$ bound to p115-RhoGEF; basis for $G_{\alpha}$/effector interactions

$G_{\alpha}$ bound to GDP, $G_{\beta\gamma}$, $G_{\alpha}$, and GDP; basis for $G_{\beta\gamma}$-mediated GDI activity

$G_{\alpha}$ bound to GDP and RGS4; basis for RGS protein $G_{\alpha}$-effectors (e.g., $G_{\alpha}$-effectors, RGS proteins, and GoLoco proteins)

$G_{\alpha}$ bound to GDP and RGS4, GoLoco motif; basis for GoLoco-mediated GDI activity

$G_{\alpha}$ bound to GEF peptides derived from phage-display (KB-752) and the dopamine D2 receptor (D2N); roles of $\alpha$3u2 loop and $\alpha$3 strand in $G_{\alpha}$ GDP release

*Structural data is accessible via http://www.wwpdb.org.*

Large G-Proteins

G$\beta$ and G$\gamma$ subunits form tightly associated heterodimers (Fig. 3). G$\beta$ begins with an extended N-terminal $\alpha$-helix and is composed mainly of a $\beta$-propeller fold formed by seven individual segments of a ~40-amino-acid sequence known as the WD-40 repeat. G$\gamma$ is an extended stretch of two $\alpha$-helices joined by an intervening loop. Assuming no significant ter-

The particular conformations of these three switch regions are critical to the protein–protein interactions that $G_{\alpha}$ makes with its nucleotide-selective binding partners such as G$\beta$-$G_{\alpha}$ effectors, RGS proteins, and GoLoco motifs (6, 8). Structures of complexes between activated $G_{\alpha}$ subunits and several different $\alpha$-effectors (e.g., $G_{\alpha}$-adenyl cyclase, $G_{\alpha}$-) or $G_{\alpha}$ bound to phage-display peptides (KB-752) and the dopamine D2 receptor (D2N); roles of $\alpha$3u2 loop and $\alpha$3 strand in $G_{\alpha}$ GDP release

**Table 1** Important heterotrimeric G-protein subunit structures obtained by X-ray crystallography

<table>
<thead>
<tr>
<th>Structure(s) of G-protein subunit(s)</th>
<th>Database identifier(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inactive conformation of isolated $G_{\alpha}$ subunits (bound to GDP)</td>
<td>1GDD; 1TAG</td>
</tr>
<tr>
<td>Inactive heterotrimeric G-protein complexes of $G_{\alpha}$ GDP/G$\gamma$</td>
<td>1GPD; 1GOF</td>
</tr>
<tr>
<td>Active conformation of $G_{\alpha}$ subunits (bound to GTP$\gamma$)</td>
<td>1GA; 1TND</td>
</tr>
<tr>
<td>Transition-state conformations of $G_{\alpha}$ subunits (bound to GDP, $G_{\alpha}$)</td>
<td>1GFI; 1TAD</td>
</tr>
<tr>
<td>Activated $G_{\alpha}$ bound to adenyl cyclase and activated $G_{\alpha}$ bound to p115-RhoGEF; basis for $G_{\alpha}$/effector interactions</td>
<td>1AZS; 1SHZ</td>
</tr>
<tr>
<td>$G_{\alpha}$ bound to GDP, $G_{\beta\gamma}$, and GDP; basis for $G_{\beta\gamma}$-mediated GDI activity</td>
<td>1AGR</td>
</tr>
<tr>
<td>$G_{\alpha}$ bound to GDP and RGS4, GoLoco motif; basis for GoLoco-mediated GDI activity</td>
<td>1K1Y</td>
</tr>
<tr>
<td>$G_{\alpha}$ bound to GEF peptides derived from phage-display (KB-752) and the dopamine D2 receptor (D2N); roles of $\alpha$3u2 loop and $\alpha$3 strand in $G_{\alpha}$ GDP release</td>
<td>2H1L</td>
</tr>
</tbody>
</table>

*Structural data is accessible via http://www.wwpdb.org.*
Mechanism of Guanine Nucleotide Hydrolysis

The mechanism of GTP hydrolysis by Gα, as well as RGS protein-mediated acceleration of this hydrolysis, has been discerned from x-ray diffraction crystallographic structures of the Gα transition state-mimetic form (Gα bound to GDP and AlF₄⁻), both in isolation and bound to the archetypal RGS protein RGS4 (7, 18), as well as hydrolysis reaction intermediates including Gα bound to guanosine 5′-(β,γ-imido)triphosphate (GppNHP) or GDP plus inorganic phosphate (19, 20). The GTP hydrolysis reaction is mediated by three conserved Gα amino acids (Fig. 4; residues numbered as found in Gαi1). Glutamine-204 in switch II coordinates the critical nucleophilic water molecule responsible for hydrolysis of the γ-phosphate, whereas arginine-178 and threonine-181 (both from switch I) help to stabilize the leaving group (as mimicked by the planar anion AlF₄⁻), with the latter coordinating a bound Mg²⁺ ion (18). These three residues within Gα are both necessary and sufficient for GTP hydrolysis to GDP and inorganic phosphate; thus, the mechanism of action of RGS proteins, the GTPase-accelerating proteins (GAPs) specific for Gα subunits, necessarily differs from those of the GAPs for small G-proteins that introduce an additional catalytic residue in trans (e.g., Reference 21). Instead, RGS protein binding to Gα stabilizes the transition state conformation, which lowers the activation energy required for the hydrolysis reaction (7, 22).
The γ subunit is subjected to posttranslational geranylgeranylation, which mimics the phosphate leaving group in the hydrolysis reaction, is highlighted in sticks configuration. The relative positioning of the N-terminal α-helix of the Gα subunit (often in the Gα-GDP/Gαi-heterotrimeric complex) is also highlighted. Coordinates are from PDB record 1OMW.

The causative agent of the Black Plague, Yersinia pestis, harbors several essential virulence determinants that it injects into host cells. One protein, the serine/threonine kinase YpkA, inactivates Gαi activation by phosphorylating serine-47, a residue located in the highly conserved diphosphate-binding loop of Gα (Fig. 4i, which impairs GTP binding (26). Although it is not a toxin, YM-254890, a cyclic depsipeptide identified from Chromobacterium sp. Q53666 culture broth that blocks ADP-induced platelet aggregation, is also an inhibitor of signaling from Gα11-containing heterotrimers, albeit with an unresolved mechanism of action potentially related to effects on receptor-catalyzed nucleotide exchange (27). Conversely, Pasteurella multocida, a common cause of animal infections, produces a toxin (PMT) that activates specifically Gα12 and not Gα11. Although the differential responsiveness of these two

**Figure 4** Residues within Gα that are critical to the GTP hydrolysis mechanism include arginine-178 and threonine-181 from switch I and glutamine-204 from switch II (colored as in Fig. 2 and numbered as in Gα11; coordinates are from PDB record 1GFI). Magnesium ion is highlighted in yellow. The planar anion aluminum tetrafluoride, which mimics the phosphate leaving group in the hydrolysis reaction, is depicted in metallic red. Note the position of serine-47, the target of ADP-ribosylation of arginine-178 within Gα. As mentioned previously, the planar anion aluminum tetrafluoride, which is used experimentally to activate Gα, is also an inhibitor of receptor-catalyzed nucleotide exchange (27).

**Figure 3** Overall structural fold of the Gαi heterodimer. The Gα subunit is colored to highlight the seven WD40 repeats that comprise the β-propeller (or "torus") fold: WD1, green; WD2, purple; WD3, cyan; WD4, orange; WD5, grey; WD6, red; WD7, blue. Residues within Gβ that contact the Gαiα2-binding peptide SIGK, which constitutes the Gβ-1 "hotspot" as defined by Bocaccio et al. (5), are highlighted in light green. The cysteine residue within Gγ (red) that is subjected to posttranslational geranylgeranylation is highlighted in sticks configuration. The relative positioning of the N-terminal α-helix of the Gα subunit (often in the Gα-GDP/Gαi-heterotrimeric complex) is also highlighted. Coordinates are from PDB record 1OMW.
particular mechanism of G\(\alpha_q\) activation by PMT has not yet been determined.

G protein-binding peptides reveal mechanistic insights and “hotspots” for chemical intervention

Using various random-peptide screening strategies, several groups have recently identified linear polypeptide sequences that bind to heterotrimeric G-protein subunits in unique and informative ways (5, 13, 29, 30). One example is KB-1753, a 10-amino acid peptide derived from bacteriophage screening (29) with selective binding affinity for G\(\alpha\)-subclass G\(\alpha\) subunits in their activated form (either GTP\(\gamma\)S or GDF AIF\(\gamma\)S-bound); KB-1753 can block \(\alpha\)-effector and RGS protein binding to these two activated G\(\alpha\) states. Its structural determination by X-ray crystallography (12) was instrumental in highlighting a universal site of effect/G\(\alpha\)-GTP engagement (as described above). A second peptide derived from the same screen, KB-752, has selective binding affinity for G\(\alpha\)-subclass G\(\alpha\) subunits in their inactive (GDP-bound) form. On binding, it enhances the rate of spontaneous GDP release; structural determination of a K\(B\)-752/G\(\alpha\) GDP complex by X-ray crystallography (29) provided strong support for a prevailing model of receptor-catalyzed nucleotide exchange in which the \(\alpha\)2 loop of G\(\alpha\), normally an occlusive barrier to GDP release, is remodeled by receptor-mediated binding of the G\(\alpha\) dimer (10).

The 36-amino acid GoLoco motif, a naturally occurring, G\(\alpha\) GDP-binding peptide sequence found in several G-protein regulators such as GoS12 and GoS14, exhibits the opposite biochemical activity to that of KB-752: namely, guanine nucleotide dissociation inhibitor (GDI) activity (i.e., reducing the rate of spontaneous GDP release) (12). Whole-cell electrophysiological studies of GPCR coupling to ion channel modulation have established that GoLoco motif-derived peptides are useful tools to uncouple heterotrimeric G-protein signaling, but they have no intrinsic ability to activate directly G\(\alpha\)-dependent signaling per se (31). By the use of mRNA display, Ja and Roberts (30) screened a set of semi-randomly permuted peptide sequences based on the GoLoco motif and identified a family of R\&A peptides (including a minimal 9-amino-acid sequence called R\&A-1) that interact with G\(\alpha\) bound GoS subunits in a manner competitive to G\(\alpha\) binding partners via engagement of switch II (13, 29, 32). Identifying this particular region of G\(\alpha\) as a “hotspot” potentially amenable to targeting by small molecules for future chemical modulation of G\(\alpha\) function. That peptides can define binding hotspots on G\(\alpha\) proteins targets has already been exploited successfully by Bonacci et al. (5) for the G\(\alpha\)phr heterodimer. Using bacteriophage display, this group identified four distinct groups of G\(\alpha\)phr-binding peptides that, despite divergent sequences, were found to bind the same site on G\(\alpha\)phr—a site identified previously in mutagenesis studies as critical to \(\alpha\)-effector interactions. A crystal structure of G\(\alpha\)phr bound to one of these peptides (SIGK) was determined by this group, which revealed a partial α helical conformation of the SIGK peptide reminiscent of that adopted by switch II of G\(\alpha\) within the G\(\gamma\)1b heterotrimer.

Small molecules that modulate the function of G-protein subunits and their regulators

Bonacci et al. (5) identified several G\(\alpha\)phr-binding compounds using the G\(\alpha\)phr/SIGK interface obtained by X-ray crystallography as the basis for computational docking screens of virtual compound libraries (5). One lead compound from this virtual screening, (S,S)-2-(3,4,5-trihydroxy-6-oxoxanthen-9-yl)cyclohexane-1-carboxylic acid (or “M119”; Fig. 5a), possesses several in vitro and in vivo activities consistent with sterol blockage of the G\(\gamma\)1b hotspot, which includes inhibition of G\(\gamma\)1b-mediated PLC\(\gamma\)2 activation and chemotactic-activated calcium signaling in immune cells, as well as sensitization of mice to the antinociceptive effects of morphine, an activator of the \(\mu\)-opioid GPCR. In a screen for compounds that can inhibit cholesterin-stimulated (but not forskolin-stimulated) cAMP accumulation in intact cells, Pelvod et al. (35) identified an imidazo-pyrazine derivative, 2-amino-1-(2-cyclohexyl)-8-(cyclohexylmethyl)-6,8-dihydro-5H-imidazo(2,1-c)pyrazin-7-yl-3-sulfanyl-propan-1-one or (“BIM-46174”) as a pan-inhibitor of heterotrimeric G-protein signaling that emanates from G\(\alpha\), G\(\beta\), and G\(\gamma\)-coupled GPCRs as well as Frizzled receptors (33). As this screen was conducted using a library of cysteine-related compounds designed originally as farnesyltransferase inhibitors, one likely mechanism of action of BIM-46174 is alteration of the G\(\gamma\)-isoprenylation that is critical to proper heterotrimeric membrane targeting and receptor coupling. Statin drugs are used clinically to reduce serum cholesterol; this class of compounds also exerts pleotropic, cholesterol-independent effects, which include the reduced production of isoprenoids because statins inhibit the rate-limiting enzyme (HMG-CoA reductase) in the mevalonate pathway. At least one statin, atorvastatin, has been found to affect β-adrenergic signaling in isolated cardiac myocytes by reducing G\(\gamma\) isoprenylation and thus decreasing the amount of functional G\(\alpha\)G\(\gamma\) heterotrimers at the cell membrane (34).

Another promising venue for developing chemical modulators of G-protein function is the G\(\gamma\)/RGs protein interface. With a chemical genetics approach that employs the egg-laying behavior of the nematode Caenorhabditis elegans, Fitzgerald et al. (35) identified two G\(\alpha\)-subunits (G\(\alpha\)-1/G\(\alpha\)-30/G\(\alpha\)-30/G\(\gamma\)-1) and an R\(\gamma\)-proteins (EAT-16) as the molecular targets for two related inhibitors of rat bladder muscle tone and spontaneous contractions [BMS-192364, BMS-195270; Fig. 5a]. A model of trapping a G\(\alpha\)-GTP/R\(\gamma\) protein pair in an unproductive complex, similar to the action of brefeldin A on the ARF1/G\(\gamma\)1/2/SIGK interface obtained by X-ray crystallography (36), has been proposed for the mechanism of action of BMS-192364 and BMS-195270 (33); however, no in vitro biochemistry has yet been published to support this model directly. Neubig et al. (37) used a flow-cytometry-based protein interaction assay to identify methyl-1-(4-chlorophenyl)sulfanyl)-1-nitrobenzene (MNS-
doate (CCG-4986) as a selective inhibitor of RGS4 (and not the related RGS protein RGS8). However, it has been shown subsequently that the selective nature of CCG-4986 lies in its mechanism of action—namely, covalent modification of surface-exposed cysteine-free thiols (Fig. 5b), which include cysteine-132 present in the Gα-interacting surface of RGS4, but not RGS8 (38). Placing this cysteine within RGS8 by site-directed mutagenesis engenders sensitivity to CCG-4986 inhibition. Along with earlier evidence that subtle changes in the Gα/RGS protein interface can lead to loss of RGS protein-mediated GAP activity (37), the precedence established by the early RGS protein inhibitors BMS-192364, BMS-195270, and CCG-4986 supports the notion that these particular G-protein regulators will be valuable drug discovery targets in the future (3). In addition, continued identification of peptide-binding “hotspots” on G-protein subunits, akin to recent Gβ1γ2/SIGK peptide and dopamine receptor loop/Gαi1 crystal structures (5, 39), should facilitate the rational design or discovery of direct chemical modulators of heterotrimeric G-protein action.

References


Wall MA, Coleman DE, Gilman AG, Sprang SR. Structure of Gα55 bound to ARF4-activated Gαi alpha); stabilization of the transition state for GTP hydrolysis. Cell 1997;89:251-261.


Further Reading

Mathematical Modeling of Biological Signaling Networks

Jason M. Haugh, North Carolina State University, Raleigh, North Carolina
doi: 10.1002/9780470048672.wecb646

Intracellular signaling networks, which are composed of interconnected biochemical pathways, regulate and actuate responses such as cell-cycle progression and cell migration, survival, and differentiation. Although our knowledge of the intricate biochemical mechanisms at the level of individual proteins and molecular interactions is ever expanding, those details leave us with an even murkier view of how the complex network operates as a whole. True understanding requires knowing not only what happens at the molecular level but also how these mechanisms influence the precise magnitude, timing, and spatial localization of signal transduction processes. Hence, mathematical modeling and analysis has emerged in recent years as a legitimate approach for interpreting experimental results and generating novel hypotheses for additional study and model refinement. Once conducted in isolation and scorned by most biologists, quantitative modeling has moved into the mainstream as a powerful tool for the analysis of cell signaling. In this article, the biological, chemical, and physical underpinnings of this approach are presented, as are its current applications and future challenges.

Biological research has been focused increasingly on the basis for cell regulation and function at the molecular level, and as a result, we now have a detailed map of how intracellular molecules are organized to form signal transduction pathways and cascades (Fig. 1). The “protein jungle,” as articulated by Bray a decade ago (1), is characterized by a wealth of qualitative information about the connectivity of pathways (the so-called interactome) but relatively little in the way of quantitative measurements. For example, one might wish to know how some meaningful quantity, such as the activity of a particular enzyme, changes as a function of time and stimulus in a given cell type. A quantitative assay could be developed for measuring that quantity. Understanding those kinetics fully, however, would require detailed knowledge of the underlying regulation mechanisms, of which several typically exist. Reconstructing those mechanisms in a test tube, in most cases, is prohibitively difficult, as is their systematic manipulation in the cell. Ultimately, the regulatory mechanisms would have to be characterized in terms of equilibrium and rate constants and intracellular concentrations. Finally, it would be useful to develop a correlation between the magnitude and timing of that pathway and the quality of cellular responses, such as rates of proliferation and probability of survival. Even if obtaining such a data set were feasible, it is not obvious how one would analyze it and extract useful insights and predictions from it.

This article deals with the mathematical modeling of signal transduction pathways and networks, which has emerged as a powerful tool that can aid in interpreting such quantitative data sets. In principle, quantitative models offer three advantages over the conceptual “arrow diagrams” that are encountered routinely in the signaling literature. First, a consequence of their mathematical construction is that they are precise, and the inherent assumptions may be laid out clearly. Second, to the extent that the underlying molecular biology is known, quantitative models can be mechanistic. Mechanistic models are based on established physico-chemical principles, in which case the form of the model equations is determined to a significant extent by the hypothetical mechanisms assumed. Thus, it is possible to evaluate the relative merits of different candidate mechanisms by comparison with experiment. In contrast with mechanistic models, phenomenological models are meant only to capture experimentally observed relationships in an empirical way. They naturally are less powerful, but they serve a definite and useful role and are appropriate in situations in which the underlying mechanisms are less certain. A prominent example is the sort of statistical, correlative models central to the field of bioinformatics; although such approaches
Receptor–ligand interactions and receptor dynamics

Receptors are responsible for transmitting information about the external environment to the cell interior, and therefore their mechanism of activation and action must be well characterized before the details of the intracellular networks can be analyzed and modeled mathematically. In most cases, receptor activation is induced by the noncovalent binding of a specific ligand. Expression of the cognate receptor, typically existing at copy numbers of 10^2–10^4 per cell, is tantamount to the ability of a cell to respond to the ligand, and thus chemical ligands signal specific cell types. Most ligands are soluble and diffusible, and such ligands are termed growth factors, cytokines, hormones, or agonists depending on the type of receptor engaged and/or the response elicited. Other ligands are associated with the extracellular matrix or other cells and mediate cell–matrix adhesion and cell–cell interactions, respectively, in addition to signaling functions. In many cases, ligation induces receptor dimerization or oligomerization as a prerequisite for signal transduction. For example, it is well established that receptor tyrosine kinases (RTKs), receptors that engage certain growth factors in a variety of cell types, must dimerize for their intrinsic enzymatic activity to phosphorylate specific tyrosine residues in the cytoplasmic portion of each receptor. The decoration of the receptor with phospho-tyrosine provides a scaffold for the recruitment of a host of signaling enzymes (4).

Receptors and receptor–ligand complexes on the cell surface are not static. Their lateral mobility in the plasma membrane enables receptor dimerization, as described above, but another critical aspect of receptor dynamics is receptor trafficking, whereby components of the plasma membrane are internalized, delivered to intracellular compartments called endosomes, and sorted either for recycling back to the cell surface or for enzymatic degradation (5). Thus, differential sorting of activated and inactive receptors provides a mechanism for regulating the number of cell surface receptors in response to chronic ligand exposure, a process called receptor downregulation, and can contribute to the clearance of external ligands over time. Typically, activated receptors are internalized at a faster rate and their intracellular sorting fate depends on the persistence of the receptor–ligand interaction in endosomes. Any realistic model of receptor-mediated signal transduction should take into consideration the processes of ligand binding, receptor dimerization or oligomerization, and receptor/ligand trafficking. Receptors that are well characterized, such as the epidermal growth factor receptor, therefore have attracted the most attention from modelers (6).

Modular functions of signaling proteins

Signaling proteins form complexes with other molecules using conserved motifs or domains, which are modular in the sense that the domain by itself typically is sufficient for its function (7, 8). Protein-protein interaction domains include those responsible for binding to phosphorylated receptors and other tyrosine-phosphorylated proteins (Src homology 2 and phospho-tyrosine binding domains) or to proline-rich protein...
sequences (Src homology 3 domains); other domains are responsible for interactions with lipids (pleckstrin homology, FYVE, C1, and C2 domains). Together, these domains mediate the formation of multimolecular complexes that regulate the activities and mediate the targeting of signaling enzymes. Many signaling enzymes have a handful of such domains, whereas other signaling proteins called adaptors have no enzymatic function but possess domains that mediate the binding of enzymes to signaling complexes. From the standpoint of modeling, this complexity presents a definite challenge.

Physical organization and compartmentalization of signaling processes

Cells are not simply well-mixed reaction vessels, and cell structure plays a critical role in signal transduction. Besides its obvious barrier function, the physical properties of the plasma membrane and other cell membranes are particularly important. In the inner leaflet of the plasma membrane, certain lipids such as phosphatidylinositol’s and phosphatidylcholine are substrates for signaling enzymes and can be hydrolyzed to form soluble and membrane-associated products or phosphorylated to yield other lipids with second-messenger functions. In signal transduction pathways that do not involve lipid modification reactions, typically a lipid-tethered protein is involved, such as small GTPases of the Ras and Rho families. Given the near ubiquity of biochemical reactions and interactions that occur at cell membranes, it is clear that aspects of both surface and solution chemistry should be considered when formulating models of signaling processes.

In modeling interactions at cell membranes, it is important to consider the mobility of the membrane-associated molecules, which is affected by the fluidity of the membrane bilayer. Often it is assumed that the membrane is isotropic and homogeneous in that respect, at least macroscopically; however, it has long been appreciated that the plasma membrane is organized into ordered and disordered subcompartments (9). Low-density microdomains, including lipid rafts and caveolae, have been found to be enriched with certain receptors, lipids, and lipid-tethered proteins that are involved in signaling, but their role in facilitating and/or regulating signal transduction remains controversial (10, 11). Even in the bulk, disordered plasma membrane, single-molecule imaging has revealed complex molecular mobility dynamics consistent with “hop” diffusion across corral-like barriers (12). These and other aspects of cell membrane organization increasingly have brought signal transduction into the realm of biophysics.

Beyond the distinctions between signaling processes that occur at the plasma membrane and in the cell cytoplasm, certain other intracellular compartments may need to be considered. As a prominent example, the signaling competency of internalized receptors has been studied for 2 decades (13, 14). Some receptor–ligand complexes remain active in endosomes and signal through certain pathways but not others, whereas other complexes dissociate and cease to signal after the delivery to the endosomes. Other intracellular compartments, such as the Golgi and endoplasmic reticulum, have been found more recently to serve as platforms for certain signaling interactions (15).

Signal Transduction Models: Chemical/Physical Foundations

The premise of molecular biology is that cellular processes are governed by physico-chemical principles, and accordingly, those principles may be used to translate known or hypothetical molecular mechanisms to mathematical equations. In this section, the general principles of chemical kinetics, mass transport, and fluid mechanics used to model chemical reaction systems are reviewed briefly.

Reaction and mass-action kinetics

In chemical kinetics, each molecule type in the system is considered a distinct species and a conservation equation is formulated to account for the change in the amount of each species per unit time. In this context, the system could be a single cell and the species could be cell-associated molecules; different states of a molecule, the simplest distinction being a two-state model (e.g., active versus inactive or phosphorylated versus dephosphorylated), are considered distinct species as are complexes of multiple molecules. The rate of each biochemical reaction or interaction is expressed in terms of the concentrations of the reacting/interacting species, defining the so-called rate law, which appears as a positive or negative term in the corresponding conservation equations depending on whether the transition generates or consumes the conserved species.

In formulating rate laws, it is common to invoke the law of mass action (Fig. 2a). In the biochemical context, this principle applies to only two types of processes: 1) the union of two species (complex formation), the rate defined as the product of both of their concentrations and a bimolecular rate constant, and 2) a spontaneous, unimolecular transition, such as the dissociation of two species from a complex or the covalent modification of a substrate in a complex with an enzyme, which occurs with a constant mean probability per unit time that defines its rate constant. A rate process that follows mass-action kinetics often is termed an elementary reaction, although that definition technically implies specific criteria that generally are not met for reactions in liquids much less for biomolecules. The advantage of mass action kinetics is that the mathematical form of the rate law is dictated by the mechanism, in which case one needs to specify only the values of the corresponding rate constants. For complicated rate processes, a notable example being the action of highly cooperative enzymes, a more abstract rate law that does not reflect the precise mechanism but nonetheless is in quantitative agreement with experimental measurements might be assumed.

Diffusion and mass transport

Biochemical rate processes depend on local species concentrations, which are not necessarily constant. When significant concentration gradients exist, it is appropriate to include a spatial dependence in the conservation equations, which then are
A. Consider the following mechanism for ligand/receptor dynamics: Cell surface receptors (R) are synthesized at a constant rate and bind reversibly with a ligand (L) to form a complex (C). Both free and bound receptors are internalized and later degraded, but they do so at different rates. Based on this mechanism, the law of mass action is used to construct the conservation equations for R and C.

B. Addition of transport effects. In the case of autocrine signaling, the cell is both the source of the secreted ligand and the responder. Spherical geometry is adopted, and a simple reaction/diffusion model is used to calculate the ligand concentration profile at steady state, assuming a dilute suspension of cells and the same receptor dynamics as in A. The profile is given by 

$$ [L] = \frac{[L]_s}{a/r} $$

where 

$$ [L]_s $$ is the value of 

$$ [L] $$ at the cell surface (r = a). It is shown readily that the maximum value of 

$$ [L]_s $$, achieved when no receptor binding exists, is equal to 

$$ \frac{V_L}{D_L} $$, where 

$$ V_L = \frac{V_R}{k_t} $$, and 

$$ D_L $$ are used here.

The actual value of 

$$ [L]_s $$ relative to the maximum is found to be a function of only two dimensionless variables: the ratio of receptor and ligand synthesis rates (equal to 1 for the plotted results) and a parameter that characterizes the efficiency of the receptor-mediated ligand capture (defined as 

$$ \frac{a k_f R_0}{D_L (k_t + k_e) } $$, where 

$$ R_0 = \frac{V_R}{k_t} $$; values of 100, 10, 1, and 0.1 were used here).

Solved in conjunction with appropriate boundary conditions that together specify the problem (Fig. 2b). For a species in solution present at a concentration 

$$ C_i (x, t) $$, where 

$$ x $$ is a spatial position vector (coordinates in 

$$ x, y, z $$) and 

$$ t $$ is time, the conservation equation is represented precisely as follows:

$$ \frac{\partial C_i}{\partial t} = - \nabla \cdot N_i + \sum_j v_{ij} r_j $$

Here, 

$$ N_i $$ is a vector that specifies the outward flux of species 

$$ i $$; thus, taking its gradient (\nabla = \partial/\partial x) and adding a minus sign gives the net rate of species 

$$ i $$ into location 

$$ x $$ by mass transport. The second term accounts for the net generation of species 

$$ i $$ by reactions 

$$ j $$, defined by the aforementioned rate laws 

$$ r_i $$ and stoichiometry coefficients 

$$ v_{ij} $$. Concentrations of species associated with membranes usually are expressed on a per unit area basis, and a planar geometry generally is adopted for the membrane.

For molecules in solution, the flux 

$$ N_i $$ is the sum of two contributions: a convection term that accounts for the bulk flow of the fluid, given by 

$$ v C_i $$ (v is the fluid velocity vector), and a diffusive flux term 

$$ J_i $$ that accounts for the tendency of molecules to disperse in solution. In the cellular context, it is important to note that the convective term also might include consideration of active transport processes, such as those actuated by motor proteins. To proceed, a semi-empirical constitutive equation that relates the form of 

$$ J_i $$ to macroscopic variables must be invoked. The only suitable theory that has been offered in that regard is attributed to Fick, who in 1855 asserted that the net diffusive flux of species 

$$ i $$ is proportional to its gradient in mole fraction; Einstein’s theory of Brownian motion, published 50 years later, is the microscopic analog of Fick’s Law. If species 

$$ i $$ is present in a dilute solution, Fick’s Law reduces to

$$ J_i = -D_i \nabla C_i $$

The diffusion coefficient 

$$ D_i $$ characterizes the mobility of species 

$$ i $$ in solution, determined by the thermal energy that promotes translational motion and the viscous drag force that opposes that motion. Substituting the expression above into the conservation
Fluid mechanics and mechanical forces

Mechanical stress and strain have not been given due consideration in the broad context of cell biochemistry, yet mechanical forces can affect intracellular signaling processes in at least two distinct ways. First, at the level of macroscopic fluid flow, the conservation of fluid momentum determines the velocity field \( \mathbf{v} \), which along with active transport considerations determines the contribution of convective mass transport to the conservation of species \( i \), as discussed in the previous section. Typically, this contribution to the reaction kinetics is neglected. Second, at the microscopic level, mechanical forces might alter directly the functions of macromolecules in prescribed ways, for example by exposing a cryptic binding pocket. This mode of regulation is thought to be at the core of mechanotransduction, which refers to the ability of cells to sense and respond to applied forces (17,18).

Mathematical Modeling Tools and Techniques

Armed with reasonably good knowledge or hypotheses of the underlying biochemistry and the ability to formulate models based on physico-chemical principles, the models must be implemented to obtain and analyze the quantitative results and to generate predictions. Except in very simple, idealized cases, this task requires the use of various numerical methods and tools. In this section, an overview of the various model types and associated methods is offered.

Continuum models

The most common approach in mechanism-based modeling is to model the system as a continuum. The underlying assumption is that the state variables of the system (e.g., species concentrations) vary continuously in time and space and that, therefore, they evolve in a deterministic fashion, according to the aforementioned conservation equations and their initial and boundary constraints. The implementation of such models involves the solution of ordinary and partial differential equations (ODEs and PDEs, respectively). ODEs are commonly encountered in kinetic models of cell signaling systems in which the species concentrations are assumed to be spatially homogeneous within the cell/cell compartment or are averaged over the domain in an appropriate way. In kinetic models of even modest complexity, the ODEs are nonlinear and therefore must be solved using numerical methods. With the advent of efficient, implicit solution algorithms and steadily increasing computing power, handling even large systems of ODEs is straightforward and the integration of these algorithms in various software packages is widespread.

Solution of PDEs, commonly encountered in spatial modeling, is more complicated and computationally intensive by comparison, but several approximate numerical methods are available for this purpose. These include the finite difference, finite volume, and finite element methods, which vary according to their ease of implementation and applicability. The finite element method is the most complicated but generally applicable method for complex domain geometries; numerous software packages that implement this method are available, although the most powerful commercial packages are quite expensive. The freely available and user-friendly virtual cell software environment, which uses the finite volume method, has been used extensively for spatial modeling of intracellular processes (19). Several challenges are encountered in spatial modeling, regardless of the method used. First, the method and the accuracy of its computed results need to be validated. This requirement typically is achieved by testing the method on a simple problem that has an analytical solution for comparison and by confirming that changes in the mesh size and time step do not affect the results significantly. Second, the accuracy of some models that need comprehensive, which is geometric complexity of the model increases from one to three spatial coordinates, and a temptation exists to oversimplify the domain geometry to reduce the dimensionality of the problem. Finally, certain cell biological processes, such as cell motility, demand advanced modeling features such as the coupling of cell mechanics and chemical kinetics and the handling of moving boundaries, which can make the problem computationally prohibitive without the introduction of simplifying assumptions.

Stochastic modeling techniques

By comparison with continuum models, stochastic models aim to account for the inherent fluctuations and discrete nature of chemically reacting molecular systems. In cell biology, this approach has been applied most prominently in the arena of gene regulation (20–22) because these systems tend to be characterized by relatively small numbers of molecules (<100) and nonlinear, switch-like transitions. For stochastic systems that vary in time only, the most commonly used method for computing probabilistic realizations of such a model is the Monte Carlo-based algorithm developed by Gillespie (23), wherein running totals of the numbers of molecules in each state are tracked and adjusted in each time step according...
to probabilities of the various rate processes. This method is easy to implement and exact, but it is computationally inefficient for so-called stiff problems that possess a broad range of characteristic time scales. Aproximate adaptations of the technique, using a method called tau leaping, allows more efficient computation of stiff systems albeit with a potential loss in accuracy (24, 25). A nother approximate approach, viable in the limit of large numbers of molecules, is the stochastic differential equation method, which also is referred to as the chemical master equation or chemical Langevin equation approach. In this method, the probabilities of the various states being populated by different numbers of molecules evolve in time according to ODEs. Although this method is not necessarily more efficient or applicable to stiff systems, it serves as the starting point for several powerful approximations that are valid in certain limiting cases (20).

Methods for modeling spatially fluctuating reaction-diffusion systems, wherein molecules are treated as discrete particles whose coordinates are tracked and moved with time, also are available. These methods account for stochastic transitions as motivated above and are appropriate when concentration gradients are expected, as in the case of diffusion-controlled reactions. They also allow one to generate movies of the simulations to visualize the constituent processes in action. In the lattice Monte Carlo method, particles occupy nodes on a regularly spaced grid and move to adjacent nodes according to probabilities determined by the diffusion coefficient (and by long-range intermolecular potentials, if applicable). Changes in state occur probabilistically, and intermolecular interactions are subject to specific rules. For example, when the molecules are treated as point particles, interactions might occur in the next time step if the two species occupy adjacent nodes. In this limit, the method is very efficient computationally but suffers in accuracy because of its inability to resolve off-lattice events, which is especially significant when concentration gradients are steep. The Brownian dynamics or dissipative particle dynamics method is similar to the lattice Monte Carlo except that the particles adopt coordinates that are continuous throughout the domain. Thus, this method is accurate but also more computationally intensive. Both methods have been used in the context of signal transduction, particularly to study processes at cell membranes (26, 27).

Construction of biochemical networks

The multiplicity of modifications and interactions that signaling molecules engage in constitutes a large number of potential activity states. For instance, a protein or protein complex with 10 phosphorylation sites has \(2^{10} = 1,024\) distinct combinations of modified or unmodified sites, and this number does not even include the binding status of each modified site. This issue, termed combinatorial complexity, presents a major challenge for detailed kinetic modeling, and hence rule-based modeling methods and accompanying software tools have been developed (28-30). In this approach, the individual interactions and reactions in the network are enumerated and the specific contexts in which they are allowed to occur (rules) are specified. The network of possible species, which could number in the hundreds or thousands, is generated automatically in the form of a system of ODEs; alternatively, the rules can be used in a stochastic simulation wherein the species of the network are formed spontaneously. Efforts are underway to extend this approach to the spatial domain as well.

Applications of Mathematical Modeling in Cell Signaling

With the availability of computational tools and, more importantly, the ability to judiciously apply physico-chemical principles to the biologic realm, what systems are and have been ripe for quantitative modeling? In this section, signal transduction processes that have been modeled successfully from the standpoint of yielding biologically meaningful insights are surveyed. Limitations on the number of references preclude a comprehensive coverage of this literature.

Models of specific signaling pathways and processes

For several reasons, no pathway has received more attention from modellers recently than the mitogen-activated protein kinase (MAPK) cascade, which involves the successive activation of three enzymes (Fig. 3). MAPKs are conserved in eukaryotes from yeast to man and are critical for cell functions that range from proliferation to differentiation to stress responses, among others. MAPKs are considered master integrators of upstream signals and master controllers of transcription factors and other downstream effectors. Aside from their obvious importance in cell biology, MAPK cascades have achieved “über-pathway” status among modellers because of their potentially interesting dynamical properties. Building on pioneering work by Goldbeter and Koshland (31), who examined a hypothetical sequence of reversible enzymatic steps, the first model of a MAPK cascade was offered by Huang and Ferrell (32), who showed that the pathway was capable of a sensitive, switch-like relationship between the input (upstream of the cascade) and output (activation of MAPK); the basis for this response was the assumed biochemical mechanism, whereby MAPK and the upstream kinase are each activated by dual phosphorylation steps in a nonprocessive fashion. Building on this finding, more recent modeling efforts have focused on the consequences of the negative feedback regulation of the cascade (33-35) and the possibility that MAPK cascades operate as bistable switches, wherein two stable states (low and high MAPK activation) can be achieved at the same input strength (36, 37). Because of these modeling efforts, a solid understanding exists of what the cascade is capable of and how it apparently is modulated. Other models have been used to study the dynamics of MAPK cascades in the context of a full pathway, initiated by the activation of a specific receptor (38, 39). This approach is equally insightful albeit less general.

A rather proving ground for mathematical modeling in cell signaling has been the dynamics of intracellular calcium. Spurred by the nonlinear nature of the calcium release mechanisms, which exhibit cooperativity and both positive and negative feedbacks, and by the availability of fluorescent dyes for
Pathway models related to specific cell responses

Signaling pathways drive phenotypic responses, but the precise details of how this occurs at the molecular level remain elusive. Consider cell migration. From the signaling literature, we know which receptor-mediated pathways are more or less important in controlling migration, and from the literature on cytoskeletal and adhesion dynamics we have a good understanding of how migration is coordinated, but currently the interface between those two fields is poorly understood by comparison. To move forward with a mathematical model capable of yielding mechanistic insights, the prudent course of action is to focus on the molecular details of the upstream signaling and use a coarse-grained, phenomenological model of the cell response or vice versa. In the context of cell migration and chemotaxis signaling, both approaches have been adopted in recent models (66–68). In other cell response-specific models, understandably more attention is paid to the execution of the response and less to the dynamics of the upstream regulation, as in the modeling of the cell cycle (69) and programmed cell death (70, 71). It is expected that such models will be refined as the details of how upstream signaling pathways regulate the execution of cell responses are revealed, and quantitative modeling and analysis could play a key role in elucidating those mechanisms.

Prospects and Challenges

Quantitative models of signal transduction processes, in conjunction with quantitative experimentation, are being used to evaluate biochemical mechanisms, predict the outcomes of novel experiments, and generate nonintuitive insights and hypotheses warranting additional study. One challenge we now face is how best to integrate such models to analyze complex intracellular and cell–cell communication systems. Although it is envisioned that quantitative models of cell physiology ultimately will progress in lock step with our knowledge of biochemical mechanisms, several hurdles loom on the horizon. Arguably the most important of these hurdles is the specification of parameter values and parameter estimation from data sets. With a few notable exceptions, mining the literature for “known” parameters is a dicey prospect, and this practice should be used only to provide reasonable, order-of-magnitude guesses for parameter values. Estimating parameters in their particular biochemical context would be ideal, and algorithms for doing so are applied readily, but a significant amount of quantitative data is needed. This exercise is only tractable for models with a modest number of parameters, and hence the challenge arises: a very detailed signaling network model might require hundreds of free parameters. At the pathway/network level, the formulation of simplified, coarse-grain models with lumped parameters offers a reasonable compromise. Even as experimental techniques for quantifying cellular processes become increasingly refined, the formulation of a useful model will continue to rely on biological and mathematical intuition, which, at its core, is the art of modeling.
References


Further Reading


See Also

Cell Cycle, Computation and Modeling of Chemical Receptor-Ligand Interactions
Signal Cascades, Protein Interaction Networks in Systems Biology
Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s
Wei Chen, Yu-Chi Juang and Melanie Cobb, UT Southwestern Medical Center—Pharmacology, Dallas, Texas
doi: 10.1002/9780470048672.wecb348

Mitogen-activated protein kinases (MAPKs) are significant mediators in signal transduction pathways from the membrane to intracellular compartments including the nucleus. They regulate the functions of many gene products and therefore affect cell growth, differentiation, and death. Three main MAPK subfamilies have been identified and are studied widely. All function in cascades that include at least three tiers of protein kinases. Selective kinase inhibitors have been developed; they are powerful tools to study the physiologic functions of MAPKs and, in some cases, are promising and effective drugs. In this article, the regulation of the three main MAPK pathways is sketched, and the status of MAPK inhibitors and their inhibiting mechanisms are highlighted.

Signal transduction pathways play important roles in transducing environmental changes to the regulatory machinery in the cell, which allows cells to alter their behavior rapidly to respond appropriately to those changes. Protein kinases are major and critical components of signaling pathways. More than 500 genes that encode protein kinase domains have been identified in the human genome. They share conserved domains in sequence and in structure, but they have notable differences in regulatory mechanisms that are often similar within subfamilies (1). Mitogen-activated protein kinases (MAPKs) are protein Ser/Thr kinases that are involved in a broad range of cellular events, such as cell proliferation, cell death, homeostasis, acute hormonal responses, and the morphologic changes of embryogenesis and cell differentiation (2–9). Perturbation of these pathways results in diseases such as cancers, diabetes, inflammatory disorders, and autoimmune disorders. Therefore, MAPKs are important targets for drug development. The regulation of MAPK pathways has been studied widely. Several model systems have been established, including activating those pathways by various stimuli and identification of in vitro and possible in vivo substrates. However, novel strategies and tools for studying the physiologic functions of MAPKs are still needed. With the identification of the first inhibitors for the ERK1/2 MAPK pathway, PD098059 and U0126 (10, 11), pathway-specific inhibitors have proven indispensable in determining MAPK functions in biologic systems. The lack of absolute specificity is a well-known disadvantage of protein kinase inhibitors and will be discussed in the context of MAPK pathways. The advantages are many and include the possibility of investigating pathways in many cells and in whole animals. Moreover, these inhibitors may have therapeutic potential as anticancer or anti-inflammatory drugs, for example. Thus far, several dozen protein kinase inhibitors are in clinical development currently, and many more are in preclinical studies (12–14). In this article, the main MAPK subfamilies and their regulation are introduced and the status of selected small molecule inhibitors of MAPK pathways will be discussed with a focus on inhibitory mechanisms. For the sake of space, references are made to more extensive reviews that cite the primary literature.

Background
The MAPKs are protein kinases activated by growth factors, hormones, cytokines, and environmental stresses. One or more MAPKs are activated by almost every cell stimulus. The first MAPKs sequenced, Kss1p and Fus3p, are also the kinases most similar to mammalian ERK1/2 and were found in the pheromone response pathway of the budding yeast nearly 20 years ago (8). Subsequently, activities found in mammalian cells that favored Ser/Thr residues followed by Pro in substrates were purified and shown to be mammalian MAPKs. More than a dozen mammalian MAPKs have now been identified. The hallmark of the MAPK family is the tri-peptide motif (Thr-Xxx-Tyr) located within the activation loop (T-loop) of the kinase domain, which contains the two sites phosphorylated to activate the kinases (2–9). Based on the canonical TXY motif and other features of the primary sequence, MAPKs are classified even more into three major subgroups: the extracellular...
Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s

Figure 1

ERK pathways

To date, six MAPKs termed ERKs (ERKs 1, 2, 3, 4, 5, 7 also called ERK8 depending on species) have been identified (7-9). All but ERKs 3 and 4 contain the highly conserved TEY (Thr-Glu-Tyr) motif in the activation loop. The atypical ERKs 3 and 4 possess a SEG (Ser-Glu-Gly) activation loop sequence instead. ERK1/2 are the archetype MAPKs and also are the best-studied kinases in this subgroup. To some extent, ERK1/2 respond to most ligands and other cellular stimuli; nevertheless, the most pronounced responses are to growth factors, serum, phorbol esters, and cytokines (frequently reviewed, e.g., (8), and other articles in that volume). ERK1/2 are 43 and 41 kDa proteins with 83% sequence identity. The MAP3Ks that activate ERK1/2 include Raf-1 (or c-Raf) and B-Raf, most typically, A-Raf where it is expressed, and Mos and Tpl2 under very specific circumstances that have been reviewed in detail (2). Raf-1 is the best studied MAP3K for this pathway and is expressed ubiquitously. Once stimulated, it is phosphorylated and then activates the MAP2Ks MEK1/2, which in turn phosphorylate and activate ERK1/2. B-Raf can associate with and activate Raf-1, and this association can be involved in the transforming events caused by B-Raf in certain tumors (13).

ERK5, which is also well studied, is known as big MAP kinase 1 (B MK1) because it is twice the size of ERK1/2. Although ERK5 shares a TEY motif in its kinase domain, it contains a unique long C terminus that may have important regulatory functions (5). In the ERK5 pathway, the upstream kinase of ERK5 is MEK5 but not MEK1/2. MEK5 may be phosphorylated by the MAP3Ks MEKK2, MEKK3, Tpl2, and mixed-lineage kinases (MLKs). However, scaffolding is thought to distinguish settings in which these MAP3Ks act on ERK5 compared with other MAPKs discussed below (16). In common with ERK1/2, ERK5 is activated in response to serum and growth factors such as nerve growth factor and epidermal growth factor. ERK5 is more sensitive than ERK1/2 to many stress stimuli, such as oxidative stress and hyperosmolarity, although all three may be activated (5).

Compared with ERK1/2 and 5, the other ERKs are studied much less. Their regulatory mechanisms are less understood and may not involve dedicated MAP2Ks (7, 9).

JNK pathways

JNK (c-Jun N-terminal kinase) was first identified as the UV-induced activity responsible for phosphorylating, and thereby activating the proto-oncogene c-Jun (5). At the same time, they were found as SAPKs (stress-activated protein kinases), which are proline-directed kinases activated by growth factors and biosynthetic inhibitors such as anisomycin. Common stimuli that activate JNKS include inflammatory cytokines; fatty acids; and environmental stresses such as UV, osmotic shock, heat
Obesity results in JNK activation, which suggests that JNK inactivation and desensitization of insulin receptor substrate 1 (IRS-1) contribute to obesity (18). JNK inhibits insulin signaling through phosphorylation and dephosphorylation of IRS-1, and it is involved in JNK activation by different stimuli. JNK has important roles in determining cell fate during metazoan development, as well as in the regulation of tumor growth, inflammation, and obesity (18). JNK inhibits insulin signaling through phosphorylation and dephosphorylation of IRS-1, and it is involved in JNK activation by different stimuli. JNK has important roles in determining cell fate during metazoan development, as well as in the regulation of tumor growth, inflammation, and obesity (18).

Knockout studies have suggested that different MAP3Ks are involved in JNK activation by different stimuli. JNK has important roles in determining cell fate during metazoan development, as well as in the regulation of tumor growth, inflammation, and obesity (18). JNK inhibits insulin signaling through phosphorylation and dephosphorylation of IRS-1, and it is involved in JNK activation by different stimuli. JNK has important roles in determining cell fate during metazoan development, as well as in the regulation of tumor growth, inflammation, and obesity (18).

Many MAP3Ks have been reported to activate MEK4/7, including MEKK1-4, ASK1, TAK1, MLKs, and JNK3/SAPKγ. These proteins, which exist in 10 or more alternatively spliced forms, share more than 85% identity in the core kinase domain. JNK1 and JNK2 are ubiquitous, whereas JNK3 is expressed primarily in neuronal tissues and in the cardiac muscle. The JNKs are activated by dual phosphorylation on the activation loop (Thr-Pro-Tyr) motif by the MAP3Ks MEKK4 and MEK7. Interestingly, MEK4 displays a preference for tyrosine and MEK7 for threonine, which suggests that MEK4 and MEK7 activate JNKs synergistically (5). Many MAP3Ks have been reported to activate MEK4/7, including MEKK1-4, ASK1, TAK1, MLKs, and JNK3/SAPKγ. These proteins, which exist in 10 or more alternatively spliced forms, share more than 85% identity in the core kinase domain. JNK1 and JNK2 are ubiquitous, whereas JNK3 is expressed primarily in neuronal tissues and in the cardiac muscle. The JNKs are activated by dual phosphorylation on the activation loop (Thr-Pro-Tyr) motif by the MAP3Ks MEKK4 and MEK7. Interestingly, MEK4 displays a preference for tyrosine and MEK7 for threonine, which suggests that MEK4 and MEK7 activate JNKs synergistically (5).

Preferential effects on different p38 isoforms. p38α is stimulated by growth factors; stresses such as UV light, osmotic shock, and ionizing radiation; as well as inflammatory cytokines such as tumor necrosis factor (TNFα) and interleukin 1 (IL-1) and 2 (2-5). Once stimulated, p38α is activated by dual phosphorylation on the TGY (Thr-Gly-Tyr) motif by the upstream MAP2Ks MEK5 and 6. These upstream kinases have preferential effects on different p38 isoforms. p38α, p38β, and p38δ are phosphorylated by both MEK3 and MEK6, and p38β is phosphorylated by MEK6. MEK3 and 6 can be activated by the MAP3Ks MEKK1-4, MLKs, ASK1, TAK1, MLKs, and TAK1. It is well known that p38 is involved in inflammation, apoptosis, and cell differentiation.

### MAPK Inhibitors

The functions of MAPKs have been studied primarily using dominant negative mutants. Constitutively active point mutants of MAPKs are not available, and other constitutive forms involve fusions or mutations of multiple residues. A class of MAPKs and MAP2Ks have also been identified, but overexpressed MAP3Ks in particular often activate other pathways in addition to those intended (7). The development of specific inhibitors for each subfamily of MAPKs and inhibitors that act at different levels in these pathways would facilitate our understanding of the complex interactions of these signaling cascades greatly. Inhibitors for some MAPK, MAP2K, and MAP3K family members have been developed. These inhibitors have assisted in identifying physiologic substrates and cellular functions of these enzymes. Selected inhibitors of MAPKs are shown in Fig. 1.

### Chemistry

Protein kinases have two substrates: target proteins and ATP-Rg5. Kinases transfer the γ-phosphoryl group of ATP to hydroxy acceptor groups of Tyr, Ser, and Thr residues within target proteins. Phosphorylation may control the activation, inactivation, protein interactions, stability, and localization of the substrate. Crystallographic studies have shown that protein kinases consist of two folding domains: a smaller N-terminal domain composed largely of anti-parallel β-helices and a larger C-terminal domain composed primarily of α-helices. The nucleotide is bound in a cleft formed at the interface of the two C-terminal domain. The ATP binding pocket, together with less conserved surrounding pockets, has been the focus of inhibitor design. Most inhibitors target the ATP binding site itself and are competitive with ATP. Thus, the potential for any compound to inhibit multiple kinases is considerable. Inhibitor specificity cannot be deduced from primary sequence similarity among kinases, however, as should become clear in the discussion below. Most JNK and p38 pathways inhibitors fall into this group. A few inhibitors have been developed that inhibit protein kinases by noncompetitive mechanisms. Generally, these inhibitors are allosteric inhibitors that bind outside the ATP pocket to conformations other than the active one or bind in a mode that prevents transition to the active conformation. MEK1/2 inhibitors are good examples of this type of inhibition. Inhibitors that are not competitive with ATP may be more effective than those that are ATP competitive in the cellular milieu, with ATP concentrations in the millimolar range. A common binding site for several allosteric inhibitors exists in protein kinases that have a conformational state in which the active site segment containing the conserved DFG (Asp-Phe-Gly) motif is moved out of the active site. This conformation is known as the DFG out state. The aspartate in this motif coordinates Mg2+ bound to ATP and is important for positioning ATP for phosphoryl transfer. Inhibitors of the three major MAPK pathways will be discussed below.
Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s

ERK pathway inhibitors

Thus far, no potent ATP competitive inhibitors of ERK1/2 have been reported. Because of the likely cross-reactivity of ATP competitive inhibitors, compounds that inhibit through interactions outside the ATP binding pocket are attractive. Substrates bind to MAPKs on the opposite face from the active site in a region, which are sometimes called the common docking or CD site; this site can also influence kinase conformation (21). Small molecule inhibitors that bind to the CD site of ERK2 have been identified recently but have not received extensive testing (23).

Several inhibitors of upstream kinases Raf and MEK1/2 are used widely in elucidating the physiologic roles of the ERK pathways in a variety of biologic processes in cell culture systems (Figs. 1 and 2 (12, 13). The MEK1/2 inhibitors U0126, PD98059, and PD184352 (IC-1040) are noncompetitive with respect to both MEK substrates, ATP and ERK1/2, which is consistent with an allosteric mechanism of inhibition (24). They are effective largely by preventing the activation of MEK1/2, in addition to inhibiting MEK1/2 activity directly at higher concentrations. The recently solved crystal structure revealed that MEK1/2 have a unique inhibitor binding site located in an interior hydrophobic pocket near but not in the Mg-ATP-binding site. The binding of inhibitors induces several conformational changes in unphosphorylated MEK1/2 that may lock them into non-functional species (22). PD0325901 is a recently reported MEK inhibitor, which was modified from PD184352 with improvement of several pharmaceutical limitations. It has a 50-fold greater potency against MEK1 than PD184352, and suppresses ERK1/2 activity longer.

The catalytic domains of MEK5 and MEK1 have just less than 50% amino acid sequence identity and the MEK1/2 inhibitors PD98059, PD184352, and U0126 will inhibit the ERK5
Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s

Pathway through effects on MEK5. Inhibition of MEK5 occurs at roughly 5-fold greater concentrations (7). Recently, a benzimidazole derivative ARRY-142886 (AZD-6244) has been reported as a highly potent and selective inhibitor of MEK1/2. The IC50 was determined to be 14 nM against purified MEK1. This inhibitor is also not competitive with ATP, which is consistent with the high specificity of this compound for MEK1/2. It is currently in phase II clinical development (12). A novel compound XL518 was announced by Exelixis (South San Francisco, CA) as a potent and specific inhibitor of MEK1/2, which has highly optimized pharmacokinetic and pharmacodynamic properties.

BAY 43-9006 (Sorafenib; Bayer/Onyx Pharmaceuticals, Emeryville, CA) is a biaryl urea that inhibits Raf-1 kinase activity in vitro with IC50 value of 6 nM, and B-Raf with IC50 value of 22 nM (25). This compound also inhibits some receptor tyrosine kinases, which include VEGF receptor family members, PDGF receptor, FGF, and c-KIT, at close to the same potency.
and other protein Ser/Thr kinases including p38α, p38β, and RIP2 kinase (27, 14). The crystal structure of B-Raf with BAY 43-9006 shows that it binds in an allosteric site near to and partially overlapping the ATP pocket of B-Raf in a DFG out conformation, and it interacts with the residues of the kinase activation loop. The interaction prevents the activation loop and the catalytic residues from adopting a conformation that is competent to bind and phosphorylate substrates. Several other Raf inhibitors exist in different stages of clinical development including Chir-265, PLX-4032, GSK5074, and ZM 336372 (26).

JNK pathway inhibitors

SP600125 and CEP-1347 (KT-7515) are the two most commonly used inhibitors to probe JNK pathways (33). Both are ATP competitive and have demonstrated efficacy for use in vivo, with the successful intervention to decrease brain damage (CEP-1347) or to ameliorate some symptoms of arthritis (SP600125) in animal models (9). Both SP600125 and CEP-1374 inhibit the JNK pathway; however, their targets are different. SP600125 is an anthrapyrazolone that inhibits JNK1, 2, and 3 directly. The IC50 values are 40 nM for JNK1, 2, and 90 nM for JNK3. Testing against a broad panel of protein kinases has shown that SP-600125 inhibits at least 13 others with a similar potency as the JNKs, including p38 (27). Therefore, the activity of SP-600125 may not be attributed to the selective inhibition of JNKs. Despite this, SP-600125 has been useful to assess the role of JNK in cell culture and disease models, particularly in combination with p38 inhibitors described below. CEP-1347 inhibits MAP3Ks of the MLK group, which inhibits activation of the JNK pathway in those contexts in which MLKs are the MAP3Ks. The intra and cellular potentials are similar: Because of the role of JNK in diseases such as diabetes and obesity, it has been important to find additional novel small molecules that could inhibit the JNK pathway. At least two pan-JNK inhibitors have been reported. The Celgene (Summit, NJ) compound CC-401 has successfully completed a Phase I trial for acute myelogenous leukemia. The Merck compound A5602801 (Merck & Co., Whitehouse Station, NJ) may have therapeutic potential in multiple sclerosis and fibrosis.

p38 pathway inhibitors

Because the activation of p38 plays essential roles in the biosynthesis and release of proinflammatory cytokines such as TNF-α and interleukin-1β, blocking its activity may offer an effective therapy for treating many inflammatory diseases such as rheumatoid arthritis and inflammatory bowel disease (2, 5-7).

Many p38 kinase inhibitors have been developed and have been evaluated extensively in preclinical models of arthritis (12, 13). Among them, SB 203580, which is a pyridinyl imidazole compound, has been used as the template for many p38 inhibitors, and it has been extremely useful to delineate the function of p38. SB 203580 inhibits the catalytic activity of p38 by binding to the ATP-binding site, but it does not prevent its inhibition by MAP2Ks, based on ATP competitive experiments and crystallographic studies (30). SB 203580 inhibits only p38α and β isoforms, but not γ or δ isoforms. VX-745 (Vertex-745; Vertex, Cambridge, MA), another p38γ/δ inhibitor is a modified pyridinyl imidazole compound with better pharmacologic characteristics. Although it progressed to a Phase II trial for rheumatoid arthritis, additional development has ceased because it crossed into the central nervous system.

A diaryl urea compound BIRB796 is a potent and selective p38 α/β/γ/δ inhibitor that bears little structural similarity to SB 203580. Structural analysis shows that BIRB796 interacts with a DFG out conformation of p38 in which the activation loop has been reorganized exposing a critical binding site; this structure is incompatible with ATP binding (31, 32). The specific binding mode of BIRB796 suggests it will be a potent reagent in the treatment of chronic autoimmune diseases. In contrast to SB 203580, BIRB796 inhibits not only p38α and β, but also the γ and δ isoforms. Other p38 inhibitors in clinical phase II trials include VX-702 (Vertex), Socrs 469 (Socrs Inc., Mount View, CA), and PH-797804 (Pfizer Inc., New York, NY). These compounds hold promise for the treatment of rheumatoid arthritis and cardiovascular diseases.

Tg12 is a MAP3K that can lie upstream of all three MAPK pathways. It activates the translation of TNFα messenger RNA and TNFα production through activation of MAPK pathways. Therefore, Tg12 inhibitors could be of value for treatment of certain inflammatory diseases. A series of 1,7-naphthyridine-3-carbonitriles and the related quinoline-3-carbonitrile (cyano-quinoline) have been found to inhibit Tg2 activity and are promising for treatment of rheumatoid arthritis (33).

Key Experiments and Observations

X-ray crystallography is a powerful and direct tool to understand kinase inhibitory mechanisms. Key experiments demonstrate that inhibitors interact with multiple regions of kinases that lead to mechanisms that may or may not be competitive with substrates. As examples, p38 inhibitors fall into both groups.

Noncompetitive mechanism

All MEK1/2 inhibitors described above, U0126, PD098059, PD184352, and ARRY-142886 are noncompetitive inhibitors. From kinetic analysis, it was deduced that these inhibitors...
Mitogen-Activated Protein Kinases (MAPKs): ERKs, JNKs, and p38s

Competitive mechanism

Most kinase inhibitors compete with ATP. SB203580 is a pyridinyl imidazole with structure shown in Fig. 2c. Young et al. (35) first showed that SB203580 could bind to the inactive, unphosphorylated form of p38 with a Kd of about 40 µM. Binding in the active site was consistent with competitive inhibition of p38 by SB203580. They showed that SB203580 could also bind to activated, phosphorylated p38 and inhibit its activity. Inhibition was ATP competitive with a Kd of 21 nM. The KIC, ATP inhibition constant was determined as 2.100 ± 0.006, in the presence of 100 nM SB203580. Frantz et al. (36) compared the affinity of SB 203580 directly for both active and inactive p38 with a radioligand binding assay. They showed that the inhibitor could bind equally well to inactive and active p38, which demonstrates even more that binding of SB 203580 is independent of phosphorylation state. More direct evidence came from the crystal structure of unphosphorylated p38 that showed a pyridinyl-imidazole in the ATP pocket (37).

Ongoing and Future Research Efforts

Selective kinase inhibitors are invaluable tools to dissect the physiologic and pathophysiologic roles of protein kinases, to identify new substrates, to identify model systems that allow evaluation of potential clinical use, and importantly, to develop new therapeutic agents. Selectivity is critical for the use of protein kinase inhibitors in basic research. Protein kinases are a large class of enzymes related by sequence, all of which share the common capacity to bind ATP (1). Most inhibitors developed to date are competitive with ATP and bind in the active site. Thus, the potential for inhibitor cross-reactivity is great. Hence, establishing whether drug effects are caused by actions on the expected target or to previously unrecognized targets is an important step to validate their use.

We now understand that the entire kinase domain, not just the active site, may provide binding sites for inhibitory interactions. Inhibition may occur through selection of conformations that prevent a functional active state. Allosteric modifiers offer the possibility of high specificity and efficacy with reduced or different cross-reactivities relative to ATP competitive inhibitors. With more and more kinase structures solved, the inactive forms of kinases are becoming attractive targets for drug design, as the inactive kinases may have more distinct active sites compared with their active forms. Targeting the diverse inactive conformations might improve the inhibition specificity. A rather different approach may come from the development of nonphosphorylatable substrate analogs to interfere with the kinase-protein substrate association. These kinds of substrate-competitive inhibitors may also provide high specificity inhibitors with distinct cross-reactivities because of the varied binding modes of protein substrates.

Some MAPKs have several isoforms that may have different specific roles. Gene knockout studies have suggested isoform-specific roles. Thus, the development of inhibitors with the ability to interfere with one or a few isoforms as well as those that inhibit all isoforms will be desirable to understand the biology of these enzymes.

References


Nitric oxide (NO) is an essential signaling molecule for many eukaryotic organisms. NO is produced in vivo by the enzyme nitric oxide synthase (NOS) from the amino acid L-arginine. The apolar gas readily diffuses across cell membranes, where it binds to the heme of soluble guanylate cyclase (sGC), the principle NO receptor. Once activated, sGC converts GTP to cGMP at a rate that is several-hundred-fold above the basal level. This NO/cGMP signaling cascade modulates several physiologic processes including vasodilation, platelet aggregation, and neurotransmission. Although the cGMP-dependent affects of NO remain active areas of research, additional cGMP-independent responses to NO also are being investigated. Endogenous levels of NO can modulate protein function by S-nitrosation, a covalent modification that has been implicated in the transcriptional regulation of genes involved in the immune response and in apoptosis.

In biologic systems, nitric oxide (NO) functions as both a critical cytotoxic agent and an essential signaling molecule. The toxicity of the diatomic gas has long been accepted; however, nitric oxide was not known to be a physiologically relevant signaling molecule until it was identified as the endothelium-derived relaxing factor (EDRF) (reviewed in Reference 1). Since this discovery, the enzymatic synthesis of NO and the signaling pathways that it regulates have been the focus of many studies. In higher eukaryotes, nitric oxide synthase (NOS) produces NO from L-arginine (reviewed in References 2–4). Despite several years of research, the NOS catalytic mechanism remains a topic of investigation, but commonly it is accepted that NO is essential for several physiologic processes. Many signaling responses that NO modulates are mediated by the NO-induced activation of the heme protein soluble guanylate cyclase (sGC). NO binds to sGC at a diffusion-controlled rate and leads to a several-hundred-fold increase in the synthesis of the second messenger cGMP from GTP (5, 6). Other diatomic gases either do not bind (dioxygen) or do not activate sGC significantly (carbon monoxide). This characteristic provides selectivity and efficiency for NO even in an aerobic environment, which is critical because of the high reactivity of NO. The NOS/sGC pathway is important for maintaining homeostasis, and many diseases have been linked to the dysfunction in NO signaling (reviewed in Reference 7). Studies on NO-dependent NOS responses continue to expand, and other roles for the gas are emerging in both prokaryotic and eukaryotic organisms. In higher eukaryotes, protein S-nitrosation, an oxidative modification of cysteine residues, is implicated in an increasing number of cGMP-independent signaling systems (reviewed in Reference 8), whereas in bacteria, a class of potential heme-based NO sensors recently has been identified and proposed to participate in two-component signal transduction pathways (9).

Background and Significance

NOS is regulated highly to ensure that NO concentrations do not reach toxic levels and to control properly the processes that respond to the signaling molecule. At high concentrations (low mM), NO in an aerobic environment can damage DNA, oxidize critical heme proteins, and covalently modify essential biologic molecules (10–12). To deter these events, the expression, cellular localization, and activity of NOS is regulated highly (reviewed in Reference 3). All three isoforms, endothelial, neuronal, and inducible NOS (eNOS, nNOS, and iNOS), are regulated at the transcriptional level. Both eNOS and nNOS are expressed constitutively, and iNOS is induced with the appropriate immunostimulatory signals. The isoforms critical to signal transduction pathways (eNOS and iNOS) generate low nanomolar levels of NO and are regulated by the appropriate immunostimulatory signals. The isoforms critical to signal transduction pathways (eNOS and iNOS) generate low nanomolar levels of NO and are regulated in vivo by the binding of calcium and calmodulin. NO produced by NOS can diffuse rapidly across a cell membrane to activate sGC, a heme protein that has evolved to bind NO selectively even in the presence of oxygen (µM) (Fig. 3). In addition to NO, cGMP production by sGC is regulated by the nucleotides GTP and ATP.
Nitric Oxide signal transduction pathway. NO synthesized by NOS diffuses across cell membranes to a target cell. NO activates sGC, which leads to an increase in cGMP synthesis. The oxidation products of NO also can react with protein thiols, which leads to protein 5-nitrosation.

Figure 1  Nitric oxide signal transduction pathway. NO synthesized by NOS diffuses across cell membranes to a target cell. NO activates sGC, which leads to an increase in cGMP synthesis. The oxidation products of NO also can react with protein thiols, which leads to protein 5-nitrosation.

NO is important for the function of the cardiovascular system and is critical for blood pressure regulation. In vascular smooth muscle cells, cGMP can bind to and activate cGK, specifically the type Iα and Iβ isoforms. These isoforms are splice variants of the same gene that have different sensitivities to cGMP. During activation, cGK phosphorylates the large conductance Ca2+-activated K+ channel (23) and IRAK (IP), receptor-associated cGMP kinase substrate (24), which are involved in the regulation of extracellular Ca2+ entry and intracellular Ca2+ release, respectively. The release of Ca2+ into the cytosol leads to smooth muscle contraction by the activation of a myosin light chain kinase (MLCK) that phosphorylates the myosin light chain (MLC) (reviewed in Reference 25). Smooth muscle contraction also is regulated by myosin light chain phosphatase (MLCP), the protein that dephosphorylates MLC. cGKI phosphorylates and inhibits RhoA, a GTPase that activates Rho kinase. Rho kinase inhibits the activity of MLCP, and therefore the cGMP-dependent inhibition of RhoA contributes to smooth muscle relaxation (26). Vasodilation also is modulated by PDE5, a major cGMP-hydrolyzing PDE. This protein is important for inducing relaxation under low Ca2+ conditions and has become an important drug target because the inhibition of PDE5 leads to increased levels of cGMP after NO-induced sGC stimulation (19).

sGC activation also is important for the immune response. Human platelets generate cGMP after NO activation of sGC, which leads to the inhibition of platelet activation or aggregation. This effect is mediated primarily by cGMP activation of cGKI. Many targets for activated cGKI have been proposed, including the vasodilator stimulated phosphoprotein (VASPI). Phosphorylation of VASP correlates with the binding of NADPH oxidase to glycoprotein IIb/IIIa, expression of P-selectin, and platelet adhesion (reviewed in References 16 and 27). Small molecule sGC activators (28–30) have been shown to inhibit platelets and are potential antiplatelet agents that could be used to treat cardiovascular diseases.

cGMP-independent signaling

The oxidative addition of NO to a thiol, termed 5-nitrosation, is a posttranslational modification that can modulate protein function. With high concentrations of NO, this modification can alter protein function indiscriminately; however, only a limited number of proteins are 5-nitrosated in vivo (8). This selectivity of nitrosothiol formation suggests that a mechanism of regulation of SNO formation and/or decay exists; however, the details of this regulation are unknown.
Nitric Oxide, Biological Targets of

The S-nitrosation of proteins has been implicated in regulating apoptosis, protein expression, and tissue oxygenation (31–33). For example, low levels of NO can inhibit apoptosis via the S-nitrosation of caspase proteases, which contain a cysteine residue that is essential for catalytic activity (33). Furthermore, this process may be regulated by the protein thioredoxin, the primary intracellular oxidoreductase that may function as a nitrosotransferase (34). S-nitrosation also inhibits the DNA binding activity of NF-kB transcription factors, which effects protein expression (31), and S-nitrosohemoglobin has been implicated in the regulation of blood flow and tissue oxygenation (reviewed in References 32 and 35).

Interestingly, both NOS and sGC have been shown to be S-nitrosated by low levels of NO (36–38). In NOS, this nitrosation occurs at zinc tetrathiolate cysteines that are critical for maintaining a functional dimer. Modification of these cysteines leads to the formation of inactive monomers, which could be a means of regulating NO production in vivo (37). S-nitrosation of sGC results in the inhibition of NO-stimulated activity (38). This mechanism of desensitization may account for the clinical condition known as NO tolerance, which is an ongoing problem in the treatment of heart disease.

No Chemistry

NOS

Mammalian NOS is a P-450-like enzyme that catalyzes the oxidation of L-arginine to L-citrulline and NO. This process is a two-step reaction that leads to a five-electron oxidation of L-arginine. The enzyme requires NADPH and O2 as substrates for both reaction steps, and iron protoporphyrin IX (heme), FMN, FAD, and tetrahydrobiopterin (H4B) as protein-bound cofactors. NOS is active as a homodimer and contains an N-terminal oxygenase (or heme) domain, a C-terminal flavoprotein reductase domain, and a central calmodulin binding region (4) (Fig. 2a). The heme domain of NOS (NOS_heme) can be isolated, and it binds heme, H4B, and L-arginine. This domain is functional if provided with reducing equivalents such as sodium dithionite (39, 40). The crystal structures of eNOS_heme (41) and iNOS_heme (42) show how substrate and cofactors bind within the active site and identify residues that are important for H4B binding and dimerization, including a zinc tetrathiolate at the bottom of the dimer interface that stabilizes subunit binding and is involved in maintaining the integrity of the H4B binding site (43). The reductase domain binds to NADPH, FMN, and FAD and provides electrons to the heme active site for catalysis, a process that is controlled by Ca2+/calmodulin binding.

In the first step, L-arginine is hydroxylated to form N-ω-hydroxy-L-arginine (NHA) (Fig. 2b). This reaction mechanism is analogous to those catalyzed by cytochrome P-450s, which involves a proposed high-valent oxo-iron intermediate that could transfer an activated oxygen species to a substrate. In the second reaction step, the 3-electron oxidation of NHA produces L-citrulline and NO. It has been proposed that this step involves the attack of a ferric peroxide intermediate on the guanido carbon; however, experimental evidence is not sufficient to distinguish between this and other proposed mechanisms (reviewed in Reference 2). The most controversial questions about the NOS mechanism concern the source of the electrons in each reaction step and cofactor stoichiometry.

NO can have a short half-life in aqueous solution, which may seem problematic for it to reach its intracellular target. A second-order dependence exists on NO autoxidation shown in the rate law below (reviewed in Reference 44).

\[ v = k[N02][O2] \]

Consequently, at nanomolar signaling concentrations, the lifetime of NO is sufficient for it to reach sGC. The end-products of NO decomposition are nitrite (NO2−) and nitrate (NO3−). NO and reaction intermediates along the decomposition pathway can react with several intracellular molecules, but reactions

Figure 2 Nitric oxide synthase. (a) Domain architecture of NOS. The heme domain binds Zn2+ (gray box), heme (gray parallelogram), and H4B (white box). The reductase domain binds FMN, FAD, and NADPH (white boxes). CaM (white box) is between the heme domain and the reductase domain. (b) Two-step reaction scheme for NO synthesis by NOS.
with heme cofactors and cysteines are the most relevant to its function as a signaling agent and also contribute to its toxicity.

Reactions with heme

Perhaps the best-characterized interactions are between NO and the heme proteins hemoglobin and myoglobin. NO binds to Fe\textsuperscript{II}-unligated globins on the order of 10\textsuperscript{7} M\textsuperscript{-1} \textsuperscript{-1} s\textsuperscript{-1} and leads to the formation of a stable 6-coordinate Fe\textsuperscript{II}-NO complex (45). These 6-coordinate complexes can be very stable and thereby inhibit the function of heme proteins by blocking the coordination site. NO also can react with Fe\textsuperscript{II}-O\textsubscript{2} complexes, which often leads to heme oxidation (Fe\textsuperscript{III}-heme) and the formation of NO\textsuperscript{3-}. With hemoglobin and myoglobin, this reaction occurs on the same order of magnitude as simple NO binding to the ferrous heme (30).

The affinity of NO for ferric heme is significantly lower than for the ferrous heme (46), but reductive nitrosylation of proteins has been observed (47). In this reaction, one equivalent of NO reduces the heme to Fe\textsuperscript{II}, and a second equivalent of NO rapidly binds to the unoccupied coordination site. This reaction also can generate the nitrosating agent NO\textsuperscript{2-}, which can react subsequently with free thiols. In fact, the 5-nitrosylation of hemoglobin and nitrophorin, a protein involved in NO storage and delivery in some bloodsucking insects, has been observed (48, 49). As the ferric heme is the physiologically relevant state for nitrophorins, it is likely that this reaction occurs in vivo.

It is evident from these reactions that O\textsubscript{2} binding to sGC would reduce the ability of the enzyme to function as a selective NO sensor. The mechanism of sGC activation by NO and the ability of NO to discriminate against O\textsubscript{2} currently are under investigation.

sGC activation and ligand discrimination

sGC is a heterodimeric protein that consists of two homologous subunits, \(\alpha\) and \(\beta\). The most commonly studied isofrom is the \(\alpha\)2\(\beta\)1 protein; however, the \(\alpha\)2 and \(\beta\)2 subunits also have been identified (50, 51). sGC contains an N-terminal heme binding region, a predicted PAS-like region, and a C-terminal catalytic domain (reviewed in Reference 52). The crystal structure of the sGC heme domain as well as sGC-like homologues discovered in bacteria (9) have facilitated the study of the heme environment and ligand binding (53–55).

NO binds to the heme of sGC at a diffusion-controlled rate to form an initial 6-coordinate complex, which rapidly converts to a 5-coordinate ferrous nitrosyl complex (Fig. 3a) (50). The breaking of the Fe-His bond is thought to be critical to the activation of sGC by NO; however, recent data has shown that the NO coordination to the heme is not sufficient for full activation (13, 56). A low-activity Fe\textsuperscript{II}-NO complex can be formed in the presence of stoichiometric amounts of NO, and this species is identical spectroscopically to the highly active form of the enzyme that is formed in the presence of substrate or excess NO. Based on these observations, two mechanisms of NO activation have been proposed. One proposal is that excess NO activates the ferrous nitrosyl complex by binding to nonheme sites on the protein (13). The second proposal involves excess NO binding to the heme to form a transient dinitrosyl complex, which then converts to a 5-coordinate complex with NO bound in the proximal heme pocket (56).

The ability of sGC to select against O\textsubscript{2} binding is important for it to function as a NO sensor because O\textsubscript{2} is present at much higher levels than NO in vivo and Fe\textsuperscript{II}-O\textsubscript{2} and Fe\textsuperscript{III}-O\textsubscript{2} proteins react rapidly with NO. Interestingly, some bacterial sGC-like homologues also bind O\textsubscript{2} and NO. These proteins were named heme nitric oxide/oxygen binding (H-NO\textsubscript{X}) proteins (reviewed in Reference 57). The crystal structure of the O\textsubscript{2}-binding H-NO\textsubscript{X} from Thermoaerobacter tengcongensis (TL) with O\textsubscript{2}...
bound was reported recently (54, 55). This structure shows that a distal pocket tyrosine interacts with bound O₂ though an H-bond. Whereas the crystal structure of the O₂-excluding H-NOX from Nostoc sp. shows that no hydrogen bond donor exists in the distal heme pocket (53), sequence analysis predicts that polar residues capable of interacting with O₂ are absent in sGC. Mutagenesis studies that introduced a Tyr into the distal pocket of the pL H-NOX domain produced a protein that was capable of binding O₂ (58); however, the same mutation in full-length sGC did not facilitate O₂ binding (59, 60). This finding indicates that the presence of a distal pocket Tyr may be involved in stabilizing Fe²⁺–O₂ complexes in H-NOX proteins, but other factors are involved in ligand discrimination in sGC. The size and overall polarity of the heme distal pocket and the strength of the proximal Fe-His bond have been proposed as mechanisms of discriminating against O₂ binding (57). The crystal structures of the H-NOX proteins also have facilitated the study of sGC activation. Specifically, the differential pivoting and bending in the H-NOX heme during NO or CO binding may account for the varying degree of activation induced by the two ligands (200-fold versus 4-fold, respectively) (53). However, details about how movement in the H-NOX domains affects the catalytic domain may remain unresolved until the full-length structure is elucidated.

Reactions with cysteine

Many pathways exist to generate a nitrosobrad in vitro by the 1-electron oxidation of NO. Nitrosothiols can be formed via the reaction of a thiol with N₂O₃, a nitrosating agent that is an inter- mediate in the decomposition of NO in aerobic solution, or via the reaction of NO with a thiol to form an addition complex (SNO⁻) followed by a 1-electron oxidation. S-nitrosation of a protein thiol also can occur by a trans-S-nitrosation reaction, from a low molecular weight nitrosothiol, such as S-nitrosoglutathione, or from a nitrosated protein thiol (8). Whereas the in vivo mechanism of protein S-nitrosation is unknown, a protein-mediated trans-S-nitrosation mechanism is an attractive possibility because of the specificity it could impart on the reaction. Additionally, the same protein could catalyze both the nitrosation and denitrosation of a specific cysteine. A report showing that the protein thiodexin can transnitrosate caspase 3 selectively and reversibly lends support to this proposal (34).

Chemical Tools and Techniques

Enzymology

The study of an enzyme requires a functional assay, or a quan- titative method for measuring the conversion of substrate to product. To develop this method the enzyme substrate(s) and product(s) must be identified. Isotopically or radiolabeled compounds can facilitate this identification. NO was identified first as an intermediate in the L-arginine to NO₂⁻/N₂O⁻ pathway by incubating macrophage cells with L-[guanido-¹⁵N₂]arginine (61). Similar experiments showed that L-arginine was the precursor in this pathway (62, 63), that L-citrulline was an additional product (63), and that L-arginine conversion to L-citrulline was coupled with NO formation (64). Chromato- graphic methods to separate substrate from product, radioim- munassays (RIA), or enzyme-linked immunosorbent assays (ELISA) can be used for the sensitive detection of reaction prod- ucts and/or substrate. Reaction stoichiometry, turnover number (kcat), and Kₘ for substrate can be determined. Additionally, reaction intermediates and possible transition states can be inves- tigated by rapid quench methods, design of rational based inhibitors, and isotope exchange experiments.

Spectroscopy of heme proteins

Iron porphyrin complexes, or heme, have diverse functions in biologic systems. Specifically the iron(III) complex of protopor- phyrin type IX is a cofactor critical to the production of NO by NOS and the subsequent activation of sGC by NO. The highly conjugated π-system of these porphyrins gives these proteins their characteristic color and facilitates the study of the biologic systems they regulate. Several methods can be used to study the ligation state and heme environment of these proteins. Briefly, electronic absorption, resonance Raman, and electron paramag- netic spectroscopy will be discussed; however, several useful techniques are used to investigate heme proteins (reviewed in Reference 65).

Electronic absorption spectroscopy can be determined to the general structure of porphyrins and their derivatives. The oxidation and coordination state of the iron and the identity of the amino acid that ligates the heme can be examined by com- paring the absorption spectrum of the protein of interest with the spectra of known heme proteins (45, 65). General charac- terization with electronic absorption spectroscopy indicated that NO and CO, but not O₂ bind to the sGC heme moiety (66). Dynamic studies of ligand binding and dissociation also can be examined with this technique on μsec-msec time scales with standard stopped-flow systems.

Heme proteins have been studied extensively with resonance Raman (RR) spectroscopy, a method that can be used to ex- aminate the environment around a heme cofactor and to confirm the atoms that coordinate to the metal center. A Raman spec- trum contains peaks of scattered light where the observed fre- quency shifts correspond to the various vibrational frequencies of the scattering molecules. Specifically, the oxidation state of the heme is indicated by the electron density marker, νₑ, in the 1350–1380 cm⁻¹ region, and ν₁, ν₂, and ν₃ are the spin and coordination state markers (65). Different ligation states of sGC (67) and NOS (68, 69) have been characterized with this technique. Interestingly, the resonance Raman spectra of sGC Fe²⁺-NO and Fe²⁺-CO complexes are influenced by the presence of GTP and known sGC allosteric activators (70, 71). This in- fluence indicates that conformational changes exist at the heme pocket during substrate and activator binding that may correlate with activation.

Electron paramagnetic resonance (EPR) is a powerful tool for studying radicals such as NO. This method specifically detects molecules with unpaired electrons. The g value, a dimensionless parameter determined from an EPR spectrum, is influenced by the spin and orbital angular momentum of the unpaired...
proteins from obligate aerobic bacteria bind NO but not O₂ (58). The nitrosative stress response (78). Interestingly, the H-NOX WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY
6
References
9. Iyer LW, Anantharaman V, Arazlud. Ancient conserved dom-
15. Trout TW. Physiological concentrations of purines and pyrim-
Nitric Oxide: Biological Targets of Oxidation

38. Sayed N, Baskaran P, Ma X, van den Akker F, Beuve A.


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY

7 - 1766 -
Nitric Oxide: Biological Targets of


See Also

Enzyme Kinetics
Hemes in Biology
Post-Translational Modifications to Regulate Protein Function
Signal Cascades, Protein Interaction Networks in Spectroscopic Techniques: Proteins
Nuclear Receptors, Chemistry of

Lioudmila A. Campbell, Graduate Program in Biological Sciences, Department of Physiology, University of California, San Francisco, California

Holly A. Ingraham, Departments of Cellular and Molecular Pharmacology and Physiology, University of California, San Francisco, California
doi: 10.1002/9780470048672.wecb402

Nuclear hormone receptors are integral players in endocrine networks that lie at the interface between biology and chemistry. Unlike most other classes of transcription factors, these proteins are designed uniquely to bind small molecules and, thus, affect gene expression in response to the cellular and organismal chemical environment. After several decades of research, it is now appreciated that nuclear receptors bind very diverse lipophilic small molecules with a wide range of specificity and affinities. Recent nuclear receptor structures coupled with large-scale screening efforts challenge the dogma that all nuclear receptors, especially the large subset of constitutively active receptors, will have ligands and will represent tractable drug targets. As such, the "pharmacologic future" for such orphan nuclear receptors may reside outside of the ligand-binding pocket.

Nuclear hormone receptors are classically defined as ligand-regulated transcription factors. The transcriptional programs affected by these proteins are linked to metabolic pathways, endocrine homeostasis, and organ development; thus, both the loss and the gain of function of these receptors are associated closely with a variety of human diseases that include developmental and metabolic defects, cardiovascular disease, diabetes, reproductive failure, and cancer. Forty-eight nuclear receptors have been identified in the human genome and are subclassified into seven distinct subfamilies that consist of NR1, NR2, NR3, NR4, NR5, NR6, and NR0 based largely on sequence similarity in their two signature domains (1). These two domains are present in almost all nuclear receptors and consist of the N-terminal DNA-binding domain (DBD) and the C-terminal ligand-binding domain (LBD). The DBD interacts with specific DNA elements located in promoters of target genes, whereas the LBD binds hormones or other lipophilic molecules (2). Additionally, receptors include two highly variable domains: the N-terminal DNA-binding domain (DBD) and the C-terminal ligand-binding domain (LBD). The DBD interacts with specific DNA elements located in promoters of target genes, whereas the LBD binds hormones or other lipophilic molecules (2). Additionally, receptors include two highly variable domains: the N-terminal DNA-binding domain (DBD) and the C-terminal ligand-binding domain (LBD). The ligand-dependent feature of some nuclear receptors has been exploited successfully for therapeutic intervention against diseases such as breast cancer, type 2 diabetes, and hypertension (Table 1, 3–39).

Overview

Nuclear receptors are sophisticated homeostatic sensors that function in the endocrine network of vertebrate organisms and allow for communication between or within different tissues and organs, often over large distances. These receptors can detect a constantly changing environment by binding small lipophilic hormones and metabolic intermediates. The ligand dependent feature of some nuclear receptors has been exploited successfully for therapeutic intervention against diseases such as breast cancer, type 2 diabetes, and hypertension (Table 1, 3–39).

The use of nuclear receptors to mediate hormone signaling seems to have developed late during metazoan evolution. Indeed, genome-wide comparisons reveal that nuclear receptors are absent in some eukaryotic genomes. However, in those organisms, other signaling pathways have been adapted to meet their endocrine needs and respond to small lipophilic molecules. For instance, no nuclear receptors have been identified in the yeast genome. Interestingly, a protein fold similar to the nuclear receptor LBD was identified by structural prediction in the budding yeast S. pombe and by structural prediction in the budding yeast S. pombe.
Table 1: Nuclear hormone receptors and their ligands

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Examples of members</th>
<th>Endogenous ligands</th>
<th>Examples of synthetic ligands</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR1</td>
<td>TR</td>
<td>Thyroid hormone</td>
<td>GC-1 (PPAR α), Rosiglitazone (PPAR γ)</td>
<td>(3-5)</td>
</tr>
<tr>
<td></td>
<td>PPAR</td>
<td>Fatty acids</td>
<td>GW6471 (PPAR α), Rosiglitazone (PPAR γ)</td>
<td>(6-9)</td>
</tr>
<tr>
<td></td>
<td>LXR</td>
<td>Oxyesters</td>
<td>GW3965, T0901317</td>
<td>(10, 11)</td>
</tr>
<tr>
<td></td>
<td>PXR</td>
<td>Not known</td>
<td>Rifampicin, SR12813, Hyperforin</td>
<td>(12-14)</td>
</tr>
<tr>
<td></td>
<td>CAR</td>
<td>Not known</td>
<td>Androstanol, Phenobarbital, CITCO</td>
<td>(15-19)</td>
</tr>
<tr>
<td>NR2</td>
<td>RXR</td>
<td>Retinoidic acid</td>
<td>GW0791 (RXR α)</td>
<td>(17, 20, 21)</td>
</tr>
<tr>
<td></td>
<td>HNF4</td>
<td>Fatty acids (?)</td>
<td>None to date</td>
<td>(22, 23)</td>
</tr>
<tr>
<td>NR3</td>
<td>ER</td>
<td>Estradiol</td>
<td>Tamoxifen, IC1164.384</td>
<td>(24-26)</td>
</tr>
<tr>
<td></td>
<td>ERR</td>
<td>Not known</td>
<td>None to date</td>
<td>(27, 28)</td>
</tr>
<tr>
<td>NR4</td>
<td>NGFI-B</td>
<td>Not known</td>
<td>None to date</td>
<td>(29)</td>
</tr>
<tr>
<td>NR5</td>
<td>SF-1</td>
<td>Phospholipids (?)</td>
<td>None to date</td>
<td>(29, 30)</td>
</tr>
<tr>
<td></td>
<td>LRH-1</td>
<td>Phospholipids (?)</td>
<td>GSX4716 (ERR γ)</td>
<td>(27, 28)</td>
</tr>
<tr>
<td>NR6</td>
<td>GCNF</td>
<td>Not known</td>
<td>None to date</td>
<td>(31, 33-35)</td>
</tr>
<tr>
<td>NR0</td>
<td>DAX-1</td>
<td>Not known</td>
<td>None to date</td>
<td>(36)</td>
</tr>
</tbody>
</table>

Nuclear Receptors, Chemistry of

cerevisiae. These transcription factors heterodimerize and bind the fatty acid oleate, reminiscent of the mammalian retinoic X receptor (RXR, NR2B)/peroxisome proliferators-activated receptor (PPAR, NR1C) signaling pathway (40). Similarly, hormone signaling in multicellular plants is not mediated by nuclear receptors despite the fact that sterols mediate many analogous functions in plant biology. Instead, plants seem to use other ligand binding motifs. For example, the growth promoting plant phytohormone brassinosteroid binds a cell surface receptor that activates downstream kinases and ultimately Myc family transcription factors (41, 42). Another large family of homeodomain-START (star-related lipid-transfer) proteins is hypothesized to affect gene expression directly after binding sterols and lipids selectively via the START domain (43, 44). Collectively, these examples suggest conserved signaling by lipophilic molecules using evolutionarily distinct binding proteins.

Ligand Activation of Nuclear Receptors

To carry out the transcriptional programs that require both activation and repression of target genes, nuclear receptors interact with numerous coregulators, which nucleate the assembly of macromolecular protein complexes that remodel chromatin and modulate transcription initiation or silencing (45, 46). For ligand-dependent receptors, the presence or absence of ligand determines the nature of the assembled protein complex. Given the importance of the LBD in binding ligand and interacting with coregulators, collective efforts of academia and industry have now elucidated LBD crystal structures for nearly all seven subfamilies (47). The nuclear receptor LBD structure is conserved and consists of an α-helical bundle (α1-α12), one to five β-strands, with three to four antiparallel layers and a hydrophobic ligand-binding pocket that occupies the core of the bundle.
Nuclear Receptors, Chemistry of

The volume of this pocket varies greatly among the receptors, which enables these proteins to accommodate ligands of varying shapes and sizes (Fig. 1) (47, 48). In addition, the LBD contains a dimerization interface that allows receptors to bind DNA as homodimers or heterodimers (49). And, for some receptors, it is also a key site for interaction with the heat-shock proteins (50).

Based on the first crystal structures of liganded nuclear receptors, the “mouse trap model” was proposed to account for ligand-initiated activation (51). Ligand was proposed to complete and stabilize the hydrophobic core of the receptor in an active conformation of the LBD. Concomitant with binding of the ligand, helix H12 that contains the activation function 2 (AF2) undergoes a dramatic rearrangement, docking across the ligand-binding pocket and trapping the ligand inside (51). This repositioning of helix H12 creates a new hydrophobic surface that is bound by the LXXLL motif within coactivator proteins (Fig. 2a) (54, 55). Interestingly, corepressor proteins compete with coactivators for binding to the same hydrophobic groove but form a slightly extended surface that eliminates the need for ligand (56). This mechanism allows the ligand to dictate nuclear receptor action by repositioning AF2 and thus shifting the equilibrium between coactivator and corepressor binding (Fig. 2a) (57). However, many nuclear receptors, especially orphan receptors, are constitutively active in the apparent absence of a ligand. Moreover, structural studies point to seemingly small receptor-specific differences within the LBDs that must underlie the diversity of receptor action in controlling distinct biologic processes (47).

Thus far, about half of all nuclear receptors have been paired with physiologic ligands. The other half remain orphaned, and either await identification of their native ligands or alternatively will never be bound by a ligand. For the most part, matching ligands with their cognate receptors has followed traditional drug discovery approaches using both cell-based assays and biologic clues. Although nuclear receptors are found readily in tractable genetic model organisms, such as flies and worms (58), hunting for ligands by standard genetic screens has proven difficult and may reflect an overrepresentation of receptors that belong to the so-called “orphan receptor” subfamilies in these invertebrate species. Exceptions include the discovery of the ecdysone hormone receptors and heme receptors in Drosophila.
surprising that some of the first LBD crystal structures were those of the steroid receptors (52). The steroid receptors were also the first to be targeted by pharmaceutical compounds, even before the availability of the high-resolution LBD structures that paved the way for structure-based drug design. The estrogen receptor (ER, NR3A) is the best example of successful manipulation of a nuclear receptor with synthetic ligands. Crystal structures of the ER LBD bound by several distinct ligands reveal the exquisite specificity with which these ligands manipulate ER into active and inactive conformations. Each ER–ligand complex presents a distinct set of structural changes in the position of the AF2 relative to the core LBD, which suggests that standard approaches can be used to design specific agonists or antagonists for this receptor. When bound by the natural ligand estradiol (E2), ER possesses a relatively small and well-defined ligand-binding pocket, and multiple contacts between the receptor and the ligand result in high specificity of interaction (24). These features allow one to design ER modulatory ligands that range from selective ER modulators (SERMs) such as tamoxifen, which exhibit mixed agonist/antagonist properties depending on the tissue or promoter, to complete antagonists such as (ICI 164,384 (25, 26). In the latter case, the ER/ICI structure revealed how addition of bulky constituents to an agonist scaffold results in a protrusion from the ligand-binding pocket and movement of the AF2 helix into nonproductive conformation, which provides a paradigm for designing steroid nuclear receptor antagonists (Fig. 2c) (26). Regrettably, this approach has not worked for other ligand-dependent receptors. Indeed, in a search for a thyroid hormone receptor (TR, NR1A, Fig. 3a) antagonist for treatment of hyperthyroidism, adding bulky constituents onto the endogenous TR ligand triiodothyronine (T3) does not create a true antagonist as would be predicted from studies on synthetic ER ligands (3, 4). On the other hand, novel synthetic TR agonists have emerged based on the structure of T3 complexed with the LBD (5).

The existence of SERMs raises some intriguing questions: What does the inactive LBD structure mean at a cellular level? Do only active nuclear receptors interact with the genome? Based on the ER LBD structures with tamoxifen and relaxifene, no productive interactions with coactivator proteins are possible because the AF2 helix adopts an inactive conformation (24, 25); however, paradoxically, SERM-bound ER receptors retain transcriptional activity in certain tissues and on certain promoters. Thus, the small overlap in tamoxifen and relaxifene regulated genes when profiled in an osteosarcoma cell line (65) illustrates how diverse the transcriptional outcomes can be for different SERMs. Similarly, an extremely small overlap was noted between groups of genes regulated by tamoxifen and E2 in a uterine cell line, despite the fact that tamoxifen is thought to be a partial agonist in this tissue (66). Although additional studies are needed, these results illustrate how ligands can alter gene expression dramatically. With the advent of new genome-wide technologies, one can begin to examine how promoter occupancy is affected by ligands. Recent studies that use chromatin immunoprecipitation combined with microarray analyses (CNP/CHIP) reveal that many ER binding sites are located at a great distance from the proximal promoters and that some sites could be bound by the receptor even in the absence of ligand.
of E2 (67). For receptors fortunate enough to have high affinity ligands, as found for steroid receptors (NR3A, NR3C), the collective information gathered from these genome-wide approaches is likely to shed new insights into the physiologic consequences of drug and provide for additional refinement of drug structure.

**PPAR and LXR: Orphans Adopted by Pharmaceuticals**

Peroxisome Proliferator-Activated Receptor (PPAR, NR1C) and Liver X Receptor (LXR, NR1H3) represent two clear examples in which the lack of structural information did not hinder the development of efficacious high affinity pharmaceutical compounds. PPAR and LXR are responsive to glucose and lipid levels, and play important roles in inflammation, cholesterol and lipid metabolism, and energy balance (68–70). Despite the fact that natural ligands for PPAR remain controversial, with fatty acids and eicosanoids as the proposed low affinity endogenous ligands for PPARα (71), highly specific synthetic agonists and antagonists have been developed (Fig. 2a [6, 7]). Indeed, thiazolidinediones and the structurally related fibrates are used widely to treat diabetes and cardiovascular disease (8, 9).

Oxysterols are the proposed endogenous LXR ligands and can bind the ligand-binding pocket of LXR and activate its transcription in cellular assays (10, 72). Additionally, genetic disruption of oxysterol biosynthesis in mice attenuates LXR function greatly (73). Existing synthetic LXR agonists show potential to treat cardiovascular disease, although their collective role in controlling liver and gut metabolism may impose unwanted, off-target effects (10, 73). Oxysterols may not be the only endogenous ligands for LXR. Remarkably, a recent study reports that LXR also acts as a glucose sensor, in which high concentrations of glucose (2 mM) displace oxysterols from the ligand-binding pocket, bind directly to the LBD, and also seem to act synergistically with the synthetic LXR ligand to affect endogenous target gene expression in the liver (74). If true, LXR would be the first intracellular glucose sensor to be discovered and could provide a molecular explanation for the prominent linkages of diabetes with cardiovascular disease. Mechanistically, the authors suggest that glucose binds directly to the LXR LBD, perhaps in combination with oxysterol, or alternatively binds elsewhere in the pocket or on the solvent exposed surface of the LBD to modulate LXR activity allosterically (74, 75). If structural analysis upholds the latter, it would raise an interesting dilemma—how would a hydrophilic molecule, such as glucose or the cellular glucose-6-phosphate, bind tightly into the hydrophobic pocket of LXR? Nonetheless, this finding is provocative and potentially provides a new paradigm for targeting nuclear receptors.

**NR5A Receptors: Large Pockets in Search of Large Ligands**

The NR5A subfamily of nuclear hormone receptors includes LRH-1 and SF-1, as well as the Drosophila nuclear receptor Ftz-F1. SF-1, is required for endocrine tissue development and sexual differentiation, and it is a major regulator of steroid biosynthesis (76). LRH-1 is essential in embryonic development, and, in adults, it regulates bile acid production, cholesterol transport, and ovarian function (77). All LBD crystal structures of murine and human members of this subfamily revealed large ligand-binding pockets and structural inflexibility as evidenced by the minimal changes observed with or without ligand or coactivator peptide (31–3335, 78). The overall stability of the NR5A subfamily can be explained partially by the presence of an additional stabilizing layer caused by a well-formed and elongated helix H2.

Phosphatidyls were found in the ligand-binding pockets of mouse and human SF-1 and human LRH-1 and are relatively large (~750 Da) compared with other ligands such as steroid derivatives. The lipid tails fit exceptionally well into the ligand-binding pocket and make several specific contacts with helix H12 and the hydrophobic cavity (Fig. 3b [31, 32]). In addition to being integral membrane components, phospholipids also bind in the ligand pocket of START domain and in phosphatidyl inositol transport proteins (79). All NR5A receptors exhibit constitutive activity in cells, thus it is unclear whether these ligands serve simply to stabilize the LBD helical bundle or whether they act as regulatory ligands. Notably, filling the pocket with bulky residues diminishes ligand uptake in biochemical assays (31) and also attenuates transcriptional activity in cells (32, 33, 35). The challenge in designing synthetic ligands for NR5A receptors is two-fold. First, finding a ligand that recapitulates the positioning of the acyl chains and the phosphate head group might be problematic and, second, whereas the SF-1 LBD protein readily exchanges the bacterial phosphatidyl glycerol with PI3P or PI2P2 (HAI, unpublished data), and might be bound naturally by phosphatidic acid (80), displacing the endogenous phospholipid with a small molecule in a cellular environment might prove difficult. However, a recent report describes a small molecule that at nanomolar concentrations promotes coactivator peptide recruitment to SF-1 and LRH-1, displaces the phospholipid ligand, and evokes a modest increase in endogenous target gene activation in human hepatocytes (34). These studies suggest that, perhaps, this family of nuclear receptors is still tractable for drug discovery.

**PXR and CAR: Too Much Receptor for a Single Ligand**

Pregnane X receptor (PXR, NR12) and the constitutive androstane receptor (CAR, NR1I2) are highly promiscuous nuclear receptors that bind a variety of structurally diverse compounds. Thus, no difficulty exists in identifying ligands for these receptors—finding highly specific ligands seems to be the challenge, which is true especially for PXR and most likely reflects its role in the xenobiotic response. Several different compounds that range from small hydrophobic drugs to the large anti-epileptic rifampicin are accommodated in its ligand-binding pocket (12-14). Five expandable β-sheets, unique to PXR, allow for this dramatic increase in size of the ligand-binding pocket.
(12). Interestingly, a similar structural feature is also found in START domain proteins and may represent a critical structural arrangement for binding a wide variety of lipophilic molecules (79). CAR exhibits a large but empty ligand-binding pocket and high constitutive activity that results from two structural features: an α helix that precedes helix H12 that stabilizes AF-2 in an active conformation and an extended helix H2, similar to the NR5A receptors (15–18, 21). For an organism, the promiscuity of PXR and CAR activation is an indispensable feature because it assures protection from a variety of harmful xenobiotics and metabolites. However, this characteristic also presents a formidable challenge to rational drug design. Once again, and as found with TR, bulky constituents added onto existing PXR agonist scaffolds fail to yield suitable antagonists (81). For CAR, it appears that a significant mode of regulation occurs by shuttling between the nucleus and cytoplasm rather than by ligand activation (82). Interestingly, inverse agonists or ligands that reduce the constitutive activity of CAR have been reported (15, 19). Whether natural ligands exist for PXR and CAR remains unclear, and it may be more likely that these receptors are designed to sample their chemical environment constantly, and protect the organism from harmful cellular metabolites or from environmental toxins.

True Orphans Without Pockets

Finally, structural information on other receptor subfamilies reveals some receptors either to be complemented with “structural non-exchangeable ligands” or to simply have inadequate capacity in their pockets to accommodate the smallest of ligands. To date, hepatocyte nuclear factor 4 (HNF4, NR2A2) is an example of a receptor with a structural ligand. Structures of the rat HNF4α (NR2A1) and human HNF4γ (NR2A2) LBDs showed a mixture of bacterial fatty acids that occupy the ligand-binding pocket (22, 23). Although HNF4 is found complexed with only a small selection of fatty acids among an assortment of many, this fatty acid ligand is entrenched completely in the ligand-binding pocket and is dislodged only after complete denaturation of the protein, which suggests that in vitro approaches to ligand identification may not be feasible. Similar to this finding, the phosphorylpid ligand in human LRH-1 is also resistant to in vitro exchange with other phospholipids perhaps, presenting another case of a structural ligand for a nuclear receptor (35).

As mentioned above, the lack of a conventional hydrophobic cavity makes the ligand hunt extremely difficult. Two receptor subfamilies appear to be “pocketless,” including members of NR4A [(NGFI-B/Nur77, NR4A1; Nur1, NR4A2), (NOR1, NR4A3)] and their fly ortholog SHH-R8, and members of the NR0 subfamily including Dax-1 (NR0B1). All three structures of NR4 LBDs adopt a canonical protein fold but lack any ligand-binding pocket due to obstruction by bulky side chains (Fig. 3e). These LBDs also lack a hydrophobic coactivator cleft that is instead replaced with a charged surface (29, 30, 83). Another case of an empty pocket is the new structure of the atypical orphan nuclear receptor Dax-1 complexed with LRH-1. Both Dax-1 and SHP (NR0B2) lack a LBD altogether and thus rely on interactions with other NRs and transcription factors to be recruited to the DNA, but both are potent repressors in cellular reporter assays (36, 84). From the crystal structure, it is evident that the ligand-binding pocket of Dax-1 (80 Å³) cannot accommodate even the smallest ligand (E. Sablin and R. J. Fletterick, personal communication). Based on this structure and given the high identity with Dax-1, SHP is also predicted to be refractory to ligand regulation.

Finding pharmacological ligands for receptors with very small pockets still remains a feasible option as illustrated by recent discovery of a synthetic agonist for the estrogen-related receptor γ (ERRγ, NR3B3). ERRγ is a constitutively active nuclear receptor with no known natural ligand, and the crystal structure of the ERRγ LBD revealed an extremely small ligand-binding pocket (220 Å³) (27). Remarkably, in a new crystal structure of ERRγ-LBD with a synthetic agonist, GSK4716, the ligand-binding pocket expanded to a notable 610 Å³ (28). This result underscores the ability of the LBD to accommodate ligands of varying size, and suggests that continuing the hunt for ligands might yield some future surprises.

Alternative Surfaces for Regulation

Despite the fact that ligand discovery has focused historically on the LBD, emerging data suggest that alternative surfaces might be targeted to regulate receptor activity. Alternative binding surfaces have been suggested by structural studies on the NR4A subfamily member, Nur1. Nuclear magnetic resonance footprinting studies of Nur1 LBD with peptides derived from the nuclear receptor coactivator INO80 and a related coactivator SMRT identified a hydrophobic binding site on the surface of the LBD between helices H11 and H12 (85). Mutational disruption of this interaction surface abolished transcriptional activity of Nur1 underscoring its importance in Nur1 function. Because the canonical coactivator groove is absent in the NR4A subfamily, this additional LBD surface is possibly the major site for interaction with the coregulators. On that note, it is of interest that crystal structures of the rat farnesoid X receptor LBD (FXR, NR1H4) and the human LXR-1 LBD revealed two coactivator LXXLL peptides bound to the receptor (31, 33, 86); in these cases, the relevance of this additional bound peptide remains to be determined.

New pharmaceuticals might act by covalent modification of a key protein-protein interaction surface, by blocking an interaction surface, or by affecting the ligand-binding pocket allosterically, as suggested for glucose binding to LXR. Presumably, for most interactions, one would disrupt the assembly of receptor-coregulator complexes and in essence mimic conventional antagonists (87, 88). Rodriguez et al. (89) synthesized a small molecule inhibitor of coactivator binding that structurally mimics key contacts of a coactivator LXXLL motif with the hydrophobic binding groove of the nuclear receptor. The authors used a crystal structure of agonist-bound ERα in a complex with a coactivator peptide to guide small molecule design followed by a screen to identify molecules that abolish peptide recruitment but do not directly compete with ligand binding. A similar high-throughput approach was used to identify novel covalent inhibitors of TRβ (NR1A2), ß-arrestins. These inhibitors
invariably react with a cysteine residue located in the coactivator groove of TRβ-LBD thus disrupting the interaction between TRβ and an LXXLL-containing coactivator peptide (90). TRβ has multiple solvent-exposed cysteines on the LBD, yet these compounds show high selectivity towards a single residue, unique to the TR family of receptors. Additionally, some of the tested compounds appear to be isomor-specific, demonstrating vastly different affinities for TRα and TRβ. Similar to these findings, 4-hydroxylamidoxifen (OHT) was found to inhibit coactivator recruitment to ERα and surprisingly, the crystal structure of the ERα-LBD revealed two bound OHT molecules (91). One molecule was bound in the ligand-binding pocket, and another molecule was revealed in the coactivator groove, displacing the AF2 away from the LBD, into inactive conformation (25). While the exact contribution of this external OHT binding site to the antagonistic effects of OHT on ERβ function is unclear, this binding event could be uncovering a subtle structural difference between the two ER isoforms. Finally, another allosteric inhibitor compound has been identified for the structural difference between the two ER isoforms. It shows reversible binding at a novel hydrophobic LBD surface, conserved in other steroid receptors, and this binding allosterically moves the AF2 helix into an inactive conformation (E. Estévez-Pérfils and R. J. Flitterick, personal communication). Collectively, these studies raise the possibility that new drugs may emerge that target additional surfaces other than the hydrophobic ligand-binding pocket.

Summary and Future Directions

The ability of nuclear hormone receptors to bind small molecules with high affinity and high specificity places them squarely at the interface between biology and chemistry. As such, the nuclear receptor field has historically been focused primarily on the identification of regulatory ligands. Now, an alternative approach is needed for those receptors that fail to exhibit classic ligand dependency, but instead appear to be ligand-independent. Domains outside the DBD and the LBD, especially the Activation Function 1 (AF1) offer a regulatory platform for multiple posttranslational events and coregulator interactions (92, 93). Positioning of the AF1 varies among nuclear receptors, suggesting that it has hopped around throughout evolution, and can be found in the variable N-terminal extension preceding the DBD (NR4) for steroid receptors or in the hinge region close to the LBD for the NR4 and NR5 subfamilies (98-101). Both the N-terminal extension and the hinge region are highly variable in length and sequence, and most likely are disordered and flexible, thus making their structural determination elusive. Multiple sites of posttranslational modifications are found in these variable regions, such as phosphorylation, sumoylation, acetylation, and ubiquitination, and these sites of modification often cluster closely together. Additionally, the AF1 appears to be a major surface for interaction with numerous coregulator proteins (102-106). Considering the importance of AF1 in nuclear receptor activity, its structure, function, and interaction with the DBD and the LBD are still poorly understood, and there are no pharmacologicals available for direct manipulation of AF1 function. For the ligand-independent receptors, such as the NR4 subfamily, posttranslational modifications might be crucial in regulating their activity (107, 108).

The interplay between posttranslational modifications and the ligand potentially leads to a myriad of functional outcomes for the nuclear receptors. We are only beginning to map out the relationships between individual posttranslational events and to understand the specific effects of their combinations on receptor activity. Numerous studies highlight the importance of “the histone code” or how posttranslational modifications of histone proteins affect transcriptional state of the chromatin and dictate transcriptional competency of genes. The abundance of posttranslational modifications on nuclear receptors suggests a similar idea of regulation.

From the extensive cellular, biochemical and structural studies carried out on nuclear hormone receptors it is now appreciated that their ability to be “classically” regulated by ligands is no longer taken for granted. Indeed, we now know that over half of these receptors are not regulated by ligands as discovered for the steroid receptors many decades ago. In the last ten years, intensive research has focused on the “orphan receptors” with the goal of finding their high affinity ligands. Now, it is realized that many receptors cannot be bound by a ligand or have a non-exchangeable “structural” ligand embedded in their pockets. For ligand-dependent receptors, the challenge for the next decade will be to refine the specificity of the existing known ligands or identify allosteric modulatory ligands. For ligand-independent receptors, research will have to take a new direction to identify other regulatory sites that can then be targeted by small molecules. Given the importance of nuclear receptors in human biology and disease, they are likely to remain a primary focus for both academia and industry for years to come.

References


75. Lazor MA, Wilson TM. Sweet Dreams for LXR. Cell Metab. 2007;5:159-163.


Nuclear Receptors, Chemistry of


See Also

Crystallization of Proteins: Overview of Applications in Chemical Biology
Liver X Receptors (LXR), Chemistry of
Orphan Nuclear Hormone Receptors
Peroxisome Proliferator-Activated Receptor (PPAR), chemistry of
Steroid Hormones, Biology and Biochemistry of
Quorum Sensing

Pratibha Singh, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan; Department of Advanced Interdisciplinary Studies, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

Jun Igarashi, The Otsuka Protech Laboratory, Komaba Open Laboratory, The University of Tokyo, Tokyo, Japan

Hiroaki Suga, Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan; Department of Advanced Interdisciplinary Studies, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan; The Otsuka Protech Laboratory, Komaba Open Laboratory; The University of Tokyo, Tokyo, Japan
doi: 10.1002/9780470048672.wecb642

Introduction

Bacterial pathogenesis can cost in terms of loss of man-hours, loss of livestock, and damage to cash crops. Quorum sensing (QS) is a mechanism by which bacteria regulate the expression of their virulence factors and the resultant pathogenicity. It presents itself as an attractive target for the design of small-molecule inhibitors. An added advantage of targeting QS is that it does not threaten the viability of the bacteria as opposed to the traditional antibacterial agents. Thus, when compared with the traditional antibacterial agents, these small-molecule inhibitors promise to maintain their efficacy over a longer period of time. In this article, the mechanism of bacterial QS systems will be discussed along with a survey of the small molecules/strategies designed to combat them.

Mechanism of the luxIR System in Gram-Negative Bacteria

In the luxIR system (Fig. 1a), AI is an acylated homoserine lactone (AHL), and the family of AHL signal molecules is generally called A-I (Fig. 1a and Fig. 1b) (3). This name refers to the chronology of QS signals discovery. LuxI is the synthase of A-I, which is synthesized from S - adenosylmethionine (SAM) and acyl-acyl carrier protein (acyl-ACP) (Fig. 1a). Activation of the QS circuit(s) relies on the synthesis, accumulation, and subsequent sensing of A-I. The minimal concentration of A-I is constitutively synthesized by its cognate synthase in Vibrio fischeri, for instance, the chemical structure of A-I is...
Quorum Sensing

Figure 1  \( P.\ aeruginosa \) QS system. Its mechanism. (a) Biosynthesis of acyl-homoserine lactone (AHL). Abbreviations: SAM, S-adenosyl methionine; ACP, acyl carrier protein. (b) General chemical structure of AHL molecules, generally called autoinducer-1 (AI-1). (c) Chemical structure of \( V.\ fischeri \) AI-1. (d) Chemical structure of \( P.\ aeruginosa \) 3-oxo-C12-HSL and (e) C4-HSL. (f) Pseudomonas quinolone signal, PQS. 3-oxo-C6-HSL (Fig. 1c), which is synthesized by LuxI (4, 5). As the cell density increases (i.e., more bacteria are present) AI-1 accumulates in the area surrounding the bacteria. When the concentration of AI-1 reaches a critical threshold, AI-1 reenters the cells and is specifically sensed by its cognate transcription factor called LuxR. The LuxR–AI-1 complex binds to a region of DNA (lux box) causing the activation of downstream genes (6). In addition, the LuxR–AI-1 complex also causes an increase in the expression levels of LuxI, which results in more production of AI-1. Thus, the AI (autoinducer) is named after the function of “auto-induction” of its own synthesis. In \( V.\ fischeri \), activation of QS circuit results in activation of luminescence genes; thus, when \( V.\ fischeri \) exists at high cell densities, its colony becomes luminous.

Why can bacteria only sense exogenous the AI signal, not the endogenously existing AI signal? The most recent biochemical studies have suggested that LuxR associates with the cell membrane, and its AI receptor site is not likely exposed to the interior of the cell when it is in an unbound form. Thus, LuxR would not see the AI signal inside of the cell. Moreover, some bacteria actively transport AI signals via an efflux pump. Such mechanisms generally operate to sense the exogenous AI signal, and thus bacteria can monitor the surrounding environment for its own species or for other species of bacteria.

The QS system of \( Pseudomonas\ aeruginosa \) has been studied extensively because of its importance as a human pathogen. \( P.\ aeruginosa \) has a unique system of QS that consists of las, rhl, and \( Pseudomonas\ quinolone\ signal \) (PQS) circuits (Fig. 2a). The first two circuits, las and rhl, have a set of proteins (synthase and receptor) and AI-1 like the typical luxR circuit. The las circuit is composed of LasI and LasR along with 3-oxo-C12-HSL (Fig. 2a) as its cognate AI, whereas in the rhl circuit RhlI and RhlR exist with C4-HSL (Fig. 2c) as their cognate AI. The fundamental mechanism of the \( P.\ aeruginosa \) QS system is similar to that of \( V.\ fischeri \), but activation of the rhl circuit relies on activation of an upstream lasI circuit (7); for example, accumulation of LasR–3-oxo-C12-HSL activates the expression of the rhlI gene, which results in the activation of the rhl circuit. Significantly, activation of the rhl circuit is connected to the transcription of genes that express secondary metabolites and virulence factors, such as pyocyanin and elastase (8). Recently, more attention has been focused on the fact that the QS system controls biofilm formation/maturation (9).

AIs isolated from different Gram-negative bacteria differ in the N-acyl side-chain length (from C4 to C14) or degree of substitution (3-oxo, 3-hydroxy, saturated, or unsaturated). Generally, assumed to be freely diffusible in bacterial cells, radio-labeled \( V.\ fischeri \) AI (3-oxo-C6-HSL) (10) has been shown to be freely diffusible into and out of \( V.\ fischeri \) and \( Escherichia\ coli \) cells. However, in \( P.\ aeruginosa \), whereas C4-HSL can diffuse freely into and out of \( P.\ aeruginosa \) cells, 3-oxo-C12-HSL is actively transported by an efflux pump (MexAB-OprM) (11) outside the cell, which results in three times higher levels of 3-oxo-C12-HSL inside the cell, which suggests that the length and/or degree of substitution of the N-acyl chain determines whether it diffuses freely or is actively pumped out from the cells.

In the PQS circuit, it has been shown that PQS (12) is involved in controlling genes required for virulence factor expression and biofilm formation. PQS regulates the expression of lasI, encoding for elastase, an important virulence factor, and both the las and rhl QS systems of \( P.\ aeruginosa \) affect the synthesis and bioactivity of PQS. LasR positively regulates the synthesis of PQS, whereas RhlR represses it. Both
3-oxo-C₁₂-HSL and C₄-HSL compete for pqsABCDE (PQS operon) regulation, and the levels of expression are thus dependent on the ratio of the two autoinducers (13). Furthermore, PQS also induces the rhl circuit (14). PQS is hydrophobic and is conveyed between P. aeruginosa cells via a specialized vesicular transport mechanism (15). The vesicles package the PQS and other quinolones/quinolines and traffic them between P. aeruginosa cells that exist in biofilms in cystic fibrosis sputum. Mutants that do not produce these vesicles also do not show PQS-mediated QS. Interestingly, PQS behaves as an antibiotic against Gram-positive cells (16, 17), which suggests that mechanisms similar to those of other known quinolones/quinolines antibiotics are likely to be in place.

**Cross-Species Communication between Bacteria Using LUXR Signals**

P. aeruginosa is a soil bacterium and shares its habitat with antibiotic-producing bacteria such as Streptomyces tenebrarius (the source of tobramycin, an antibiotic used commonly against P. aeruginosa) (18) and the pathogenic filamentous yeast Candida albicans (19). It thus may use its AI molecules to restrain antipseudomonal compound(s) that produce environmental pathogens. Recently, it has been shown that the degradation product of 3-oxo-C₁₂-HSL resembles a class of antibacterial compounds called tetramic acids (Fig. 3) (20). Such a tetramic acid analog is produced through the enol form of the 3-oxo-carbonyl group on the fatty acid side chain undergoing intramolecular alkylation of the lactone ring of HSL, which gives rise to a HSL open-ring structure. This rearranged molecule turns out to resemble tetramic acid structurally. It has been shown that this tetramic acid-like molecule acts like an antibiotic toward several Gram-positive bacteria, while leaving the Gram-negative bacteria unaffected. Although the mechanism of bactericidal activity of this tetramic acid-like molecule derived from 3-oxo-C₁₂-HSL is as yet unclear, it has been postulated that it chelates metal ions such as Fe³⁺ and that this complex somehow may increase its bactericidal potency (Fig. 3) (20). Alternatively, it simply may exhibit the antibiotic activity via a mechanism similar to that of reutericyclin, which is known to act as a proton ionophore that dissipates the transmembrane change in pH and leads to the cell lysis of Gram-positive bacteria (21).

4-hydroxy-2-heptylquinoline-N-oxide (HQNO), a quinolone family member, similarly has antistaphylococcal activity, which suppresses the growth of many Gram-positive bacteria. Paradoxically, it also allows some Gram-positive bacteria to grow, albeit slowly, in the presence of aminoglycoside antibiotics like
Quorum Sensing

Why is Quorum Sensing Important as a Therapeutic Target?

Traditional antibacterial agents target processes crucial to the survival of a bacterium. These processes include cell wall biosynthesis, replication, and protein synthesis, which endanger its survival and thus lead to the development of drug resistance in bacteria. Bacteria eliminate drugs by exporting them out via multidrug pumps or by circumventing/degrading the drug molecule. QS inhibitors (QSI(s)), however, do not threaten bacterial survival, and therefore, in principle, bacteria would not be forced to evolve resistant strains if QSI(s) are used. Because QS regulates the production of virulence factors, QSI(s) potentially can attenuate the virulence of pathogenic bacteria (Fig. 3). Thus, QSI(s) may offer us a new class of antimicrobial agents that would not suffer from the emergence of resistant strains.

P. aeruginosa is an example of one such attractive target for therapeutic drug design. It is the leading cause of mortality among cystic fibrosis patients. It also affects immunocompromised individuals with HIV infections, cancer, burns, or organ transplants. Through knockout studies, it has been firmly established that P. aeruginosa, incapable of QS, is also incapable of virulence. For instance, analysis of a lasI mutant, rhlI mutant, and lasI-rhlI double mutant in a neonatal mouse model revealed markedly decreased virulence (24), with the most notable reduction in the double mutant. Other infection models studied (25, 26), firmly establish that it is the QS mechanism that enables P. aeruginosa to form biofilms and cause sustained infections in the host system (Fig. 3). Thus, strategies geared toward elimination of QS can aid in clearing the infection because they retard the formation of biofilm and the resulting sustained virulence, which makes the QS system an attractive therapeutic target.

Small Molecular QS Modulators

The discovery of AI antagonists is challenging and therapeutically significant. A potent antagonist was recently discovered by modification of natural AI antagonists, halogenated furanones (Fig. 4a), produced by a marine alga to prevent colonization by QS bacteria (27–31). These natural products disrupt the Serratia liquefaciens SwrR-C4-HSL (32) interaction and also inhibit the LasR-3-oxo-C12-HSL (33) and CarR-3-oxo-C6-HSL (27) interactions but have little activity against the LasR-3-oxo-C12-HSL interaction. Givkov et al. synthesized an analog that lacks the alkyl side chain of the natural furanones and found that this compound had considerable inhibitory activity against the P. aeruginosa QS system (Fig. 4b). They showed that the molecule inhibited QS-controlled reporter genes and virulence factors in a QS mutant with exogenous AIs but not in wild-type P. aeruginosa with natural levels of AI. This compound did not inhibit biofilm formation, but it affected biofilm architecture and enhanced the process of bacterial detachment. Interestingly, the same group recently reported that the QS inhibition by the synthetic furanone was not because of the interaction with LuxR demonstrated in V. cholerae; more likely, it accelerates the turnover, for example, degradation, of LuxR via an unknown mechanism(s). Although this finding might not be totally surprising because the synthetic furanone and natural furanones are structurally quite different from natural AIs, it is significant that the molecules somehow associating with QS inhibition can attenuate bacterial virulence.

Natural furanones are not alone in having anti-AI activity. They are joined in nature by garlic, vanilla, and phytochemicals present in dietary fruits and spices, and so forth. Garlic extracts have been shown to inhibit QS sensing in P. aeruginosa and also render it sensitive to tobramycin and phagocytosis by polymorphonuclear leukocytes (PMN) in a mice model of P. aeruginosa infection (34). However, the garlic-extract dose used to treat the mice equates to approximately 50 bulbs of garlic per day for an average human being. Thus, more work to identify and isolate the pure compounds responsible for inhibition of QS is awaited. Vanilla extract, however, mainly contains vanillin (85%). Vanillin, despite having no structural similarity to furanone derivatives or AIs, shows significant inhibition of QS in Chromobacterium violaceum (35). In Chromobacterium violaceum also is inhibited by phytochemicals in dietary fruits, herbs, and spices (36). These phytochemicals also affect the swarming mobility of P. aeruginosa through a possible impact on synthesis and activity of AI molecules. Macrolides like azithromycin also repress the las and rhl in P. aeruginosa albeit at sublethal concentrations (37).

Molecular insights into the LasR-3-oxo-C12-HSL interactions, available from crystal studies, may provide an important tool into the process of antagonist design. Over the past decade, tremendous efforts have been made to generate a soluble form of various LuxR family proteins and their cognate AIs. Three LuxR proteins, Erwinia chrysanthemi ExpR (38), Agrobacterium tumefaciens TraR (39), and Erwinia carotovora CarR (40) have been expressed successfully as a complex with the cognate AIs and purified for in vitro characterization. Particularly, the recent success in solving the crystal structure of the dimer of TraR-AI (3-oxo-C6-HSL) complex interacting with the target DNA (41, 42) has provided the first visual information regarding the molecular interactions between a LuxR family protein and its cognate AI. It should be noted that the dimerization of TraR (as well as ExpR and CarR) has been predicted by the in vitro biochemical studies, and thus the X-ray structure most likely represents the active form in vivo.

Despite the fact that many attempts to overexpress the full-length LasR failed, Bottommy et al. (43) have recently succeeded in crystallizing a construct of LasR consisting of only the N-terminal domain.
the predicted ligand binding domain (LasR-LBD), which consists of the amino acids from Met-1 to Ile-173) in a complex with 3-oxo-C12-HSL. The complex was observed as a symmetrical dimer of LasR-LBD, and 3-oxo-C12-HSL was buried deeply inside the binding site. The monomer exhibits structural features similar to TraR, which consists of an α-β-α sandwich structure with three α-helices flanking a five-stranded antiparallel β-sheet. 3-oxo-C12-HSL lies parallel to the β-sheet and lies buried in a pocket formed between the β-sheet and α3–α5 helices. Also, all the polar groups of 3-oxo-C12-HSL, except the oxygen of the lactone ring, make hydrogen bonds with the amino acid residues of LasR–LBD. The acyl chain extends into a hydrophobic residue-lined cavity. Some of these residues are observed only in LasR, thus providing high specificity and a minimum crosstalk between bacteria of different species. In silico, modeling of the interaction of available QS inhibitors with the LasR-LBD provides a clue to the mechanism of action of the inhibitors. This modeling also provides a scaffold for the de novo design of better inhibitor molecules.

Suga et al. (44, 45) have reported that the screening of a library of synthetic AI analogs with substituted HSL has yielded a novel class of QS inhibitors. Interestingly, two of these antagonists found in the screening, 3-oxo-C12-(S) 2 and 3-oxo-C12-(S,S) 1, are structurally related to the synthetic agonist 3-oxo-C12-(S,S) 1 and the antagonist 3-oxo-C12-(S) 2. Small structural changes of the agonist HSL substitute (1) altered activity dramatically from agonist to antagonist. In light of the strong agonist activity of 1, these antagonists most likely maintain binding to LasR but fail to activate it and hence act as potent inhibitors. These QS inhibitors inhibit expression of virulence factors, such as elastase B and pyocyanin. Moreover, 3-oxo-C12-(S) 2 inhibited biofilm formation in both QS mutant (in the presence of exogenous AIs) and wild-type P. aeruginosa under static conditions. The latter compound 3-oxo-C12-(S) 2 did not inhibit biofilm formation but significantly altered biofilm architecture. Thus, these antagonists are promising leads to derive more potent antagonists by the screening of focused libraries.

In fact, a resurvey of focused libraries based on the antagonists found in the initial screening has given several new strong antagonists. 3-oxo-C12-(S) 2 and its related molecule 3-oxo-C12-(S,S) 1 (which also is related to the 3-oxo-C12-(S) 2 ) have shown a strong inhibitory effect on las circuit and also show inhibition on the downstream rhl circuit (Fig. 4c). Likewise, 3-oxo-C12-(S) 2 exhibited a strong inhibitory effect on the QS circuits. Significantly, 3-oxo-C12-(S,S) 1 and 3-oxo-C12-(S,S) 6 showed remarkable abilities in the inhibition of biofilm formation of wild-type P. aeruginosa.
Quorum Sensing

In flow-cell experiments (Hiroaki Suga et al., unpublished data), these compounds to wild-type P. aeruginosa that had formed mature biofilms resulted in the detachment of biofilms and the removal of bacteria from the glass surface in a flow cell. Along similar lines, Greenberg et al. (46) reported two new substrates of HSL that can inhibit P. aeruginosa QS circuits (Fig. 4d). These QS inhibitors are promising candidates for drug development, and more results from in-depth studies in various models of infection are awaited.

A deletion of the fatty acid side-chain based on the structure of AHL is an obvious alternative approach to antagonist(s) discovery. Earlier works devoted to changing the length of the fatty acid or 3-functional group unfortunately did not yield notable antagonists. However, more recent investigations independently reported by Doutheau (47, 48) and Blackwell (49) successfully showed that more drastic derivatizations of the fatty acid side chain generated potent antagonists. Doutheau et al. has shown that AI analogs with C6-acyl, C7-sulfonyl, or C6-ureido chains (Fig. 4d, 7-10) bearing aromatic groups such as phenyl at the end of an alkyl chain antagonize the QS circuit of V. fischeri. Blackwell et al. later observed that substitution of the aromatic ring with an indole ring on the C6-acyl chain gave approximately two-orders of magnitude stronger inhibitory effect on P. aeruginosa QS system than the parental compounds reported by Doutheau. This molecule also could block the biofilm production under static conditions. It must be noted, however, that these AHL-based analogs likely suffer from an instability problem because of the hydrolysis of the lactone ring when the analogs are subjected to an in vivo environment (vide infra).

In conclusion, it is to be noted that although the current QSI candidates can disrupt the interaction of the A-HL molecule with its cognate receptor, they are not potent enough to halt the QS process and the resultant pathogenesis. Thus more studies geared toward the discovery of molecules that antagonize the QS system are awaited, in particular, those that target not only the receptors but also the AHL synthases.

Modulating the QS System Outside of Bacterial Cells

Because the QS system is triggered by the exogenous AI signal entering the cell, if the QS signals were to be degraded (50) so that the signaling function is lost, then the QS system could be modulated. In fact, some prokaryotes and eukaryotes do have such defense systems.

The aiiA gene (51) encoding an AHL lactonase, recently discovered in a Gram-positive bacillus bacterium isolated from soil, enables it to compete against Gram-negative bacteria in the soil. Expression of aiiA in a transformed Erwin carotovora strain reduced the release of AI significantly and attenuated the pathogenic effects on important crops like potato, eggplant, Chinese cabbage, carrot, celery, cauliflower, and tobacco. Such genetically engineered crops would be expected to fare better against bacterial infections.

Paraxonases (PONs) (52) are a family of mammalian lactone hydrolases, expressed in liver and various tissues, and they can deactivate 3-oxo-C12-HSL by the hydrolysis of the lactone ring. Similarly, arachidonic acid cells can inactivate 3-oxo-C12-HSL by an enzymatic mechanism. (53) The inactivation is selective for A-HLS with certain carbon chain lengths, and C6-HSL is reportedly immune to such inactivation. This inactivation activity is cell associated and not mediated by a secreted factor. Also, the inactivation is shown to occur in cell-free lysates, and the ability of the cell-free lysate to inactivate 3-oxo-C12-HSL depends on the amount of lysate used.

An alternative approach to the inhibition of QS is the use of antibodies directed toward the A-HL molecule. Recently, Kaufmann et al. (54) has reported the generation of anti-AHL monoclonal antibodies (mAbs). Their hapten design initially focused on the synthesis of A-HL analogs with a carboxylic acid functionality to facilitate the binding of BSA or KLH. Because of the instability of the hapten, a lactam moiety replaced the lactone ring. Three haptenes thus were synthesized, and after conjugation to KLH (18–23 hapten molecules per carrier protein), Balb/c mice were immunized for generation of hybridomas.

The monoclonal antibodies generated against the 3-oxo-hapten had a good affinity (Kd = 15 nM to 5 nM) for 3-oxo-C12-AHL and the lactam analog and high specificity because short-chain 3-oxo-AHLs were not recognized. Also, these antibodies, in particular RS2-1G9, demonstrated strong inhibition of QS signaling in both wild-type and mutant P. aeruginosa PAO cells concomitant with inhibition of pyocyanin production. The crystal structure of FabRS2-1G9 in complex with the lactam analog has revealed that the polar lactam moiety is encapsulated completely in the antibody-combining site (55). This study provides insight into the immune recognition of a quorum-sensing molecule by an antibody. Furthermore, this structure can be used for protein engineering that leads to an enhanced interaction of an antibody with the A-HL molecule. Lactonase activity could be added into the antibody through site-directed mutagenesis of the antibody. Lending additional credence to the antibody approach is the study on 3-oxo-C12-HSL-BSA conjugates (56). Mice immunized with 3-oxo-C12-HSL-BSA conjugate showed significant amounts of antibody in the serum. When challenged intranasally with P. aeruginosa, 36% of the mice survived for 4 days post challenge as compared with the control mice that died in 2 days. Interestingly, the bacterial numbers in the lungs were similar in the two groups. Thus, specific antibodies to 3-oxo-C12-HSL confer a protective advantage against acute P. aeruginosa infections.

Conclusion and Perspective

In this article we have seen how QS presents an alternate route to combating bacterial pathogenesis. Although more studies are awaited to develop efficient drug molecules to inhibit QS system, it nonetheless proves to be an attractive target for drug development. Also, the attenuation of virulence observed with the use of QSI is not accompanied with a loss in viability of the bacterial pathogens. In this respect, QSI are different from traditional antibacterial agents that act through affecting the viability of the bacteria in various ways. It is hoped that QSI will not suffer from the drawback of resistance development, which
tradiotional antibacterials have to battle against. Thus, more studies geared toward the design of QSIs, including both synthetic analogs of AI molecules and inhibitors of AI synthases, may lead to the availability of drug molecules capable of maintaining their efficacy over longer periods of time compared with the traditional antibacterial agents.

**References**


Quorum Sensing


Further Reading


Further Reading


- 1785 -
See Also

Antibacterial Drugs, Design of
Chemical Signals in Sensing
Quorum Sensing
Ras Oncoproteins, Structure, Biochemistry and Biology of

Emily J. Chenette, Duke Institute for Genome Sciences and Policy, Duke University Medical Center, Durham, North Carolina

Channing J. Der, Department of Pharmacology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
doi: 10.1002/9700470046k72.wenb001

The Ras small GTPase is well known for its role in regulating normal cellular proliferation, as well as in promoting human oncogenesis when activated mutationally. In signal transduction, Ras functions as a guanine nucleotide-regulated on-off switch. Positioned at cellular membranes, Ras relays signals initiated by diverse extracellular stimuli to a complex network of cytoplasmic signaling cascades to affect changes in gene transcription, cell-cycle progression, survival, and differentiation. The intense research into the biologic and chemical nature of Ras has prompted the development of a variety of biologic, pharmacologic, and genetic tools to study Ras signaling. Also, Ras is the founding member of a superfamily of Ras-related and Ras-like proteins. The technical approaches and chemical concepts that have been generated from the study of Ras have aided greatly the studies of Ras superfamily proteins, which revealed the versatile and divergent biologic roles of small GTPases in cell physiology.

The connection between mutations in RAS genes and in human carcinogenesis was first established nearly 25 years ago (1). Research efforts since then have demonstrated that aberrant Ras signaling plays an integral role in the development and malignant growth of many types of human cancers. Ras proteins are simple binary switches that are controlled by a regulated GDP/GTP cycle (2) (see Fig. 1). A diverse spectrum of extracellular stimuli promotes the transient formation of active, GTP-bound Ras. Then, activated Ras interacts with downstream effectors that regulate cytoplasmic signaling cascades to promote changes in cell-cycle progression, actin cytoskeletal organization, cell survival, and gene transcription (3). Whereas wild-type Ras cycles between the GTP-bound and GDP-bound states in a regulated manner, cancer-associated mutations lock Ras in the “ON” position, which renders the protein constitutively activated. Given its role in oncogenesis, the biologic and chemical nature of Ras and the consequence of persistent Ras activation has been the subject of intensive biologic and pharmacetical research. This article will provide a brief overview of the chemistry, structure, and biology that underlie Ras signal and will summarize the key assays and reagents that have been developed and applied to study Ras signal transduction and biologic activity.

Biology

Ras proteins function as nodal points in signal transduction. Diverse extracellular stimuli act on plasma membrane-bound receptors (which include receptor tyrosine kinases and G protein-coupled receptors) and cause the activation of Ras (see Fig. 1). A accumulating evidence suggests that once activated, Ras may transit from the plasma membrane to endomembranes (4, 5). Alternatively, endomembrane-bound Ras may be activated by a discrete set of upstream activating proteins (4, 6, 7). Activated Ras then regulates myriad cytoplasmic signaling cascades to cause changes in normal cell physiology, which includes the regulation of cell morphology, growth, survival and differentiation, vesicular transport, and gene expression.

In addition to its role in promoting normal cellular growth and differentiation, Ras is perhaps best known as an oncogenic protein for its involvement in human carcinogenesis. Approximately 30% of all human cancers express a mutationally activated Ras protein (1, 8), with the highest incidence of somatic Ras mutation observed in pancreatic (90%), lung (50%), and colon (30%) cancers (9). An extensive body of research has implicated a strong causal role for mutated Ras in cancer development and growth. This relationship has prompted considerable
Ras Oncoproteins: Structure, Biochemistry, and Biology of Ras

The three human RAS genes encode four highly related 188-189 amino acid RAS proteins (H-Ras, N-Ras, and two K-Ras-encoded splice variants, K-Ras4A and K-Ras4B) that share significant sequence identity (90%) and common structural elements (see Fig. 2). RAS genes are conserved in vertebrate and invertebrate evolution, with homologous genes found in C. elegans, Drosophila, and yeast but not in plants or bacteria. RAS genes are the founding members of a greater than 150-member superfamily of RAS-related genes that encode small GTP-binding and hydrolyzing proteins (GTPases) (12, 13). RAS proteins are composed of two discrete sequence elements: a catalytic G domain and a hypervariable membrane-targeting domain (see Fig. 2). The amino-terminal 166 amino acids of RAS compose the G domain, which contains the consensus sequence motifs found in classic GTP-binding proteins. The carboxyl-terminal 23 to 24 amino acids compose the hypervariable region. The hypervariable region terminates in a consensus CAAX tetrapeptide motif that signals for posttranslational modifications that increase the hydrophobic nature of RAS. This region also contains additional sequences that promote membrane association and direct subcellular localization of Ras.

The three-dimensional structure of the G domain consists of five α-helices, six β-strands, and five loop regions (G1–G5) (see Fig. 3). The five loop regions are involved in protein–protein interaction and nucleotide binding (14). Within the G domain, two switch regions (switch I and II, residues 30–38 and 59–76, respectively) form exposed flexible loops that adopt different conformations depending on the identity of the bound guanine nucleotide. The effector loop or effector domain, which corresponds to G2, is embedded in switch I and forms critical contacts with effector molecules. Therefore, the stereochemical arrangement of the switch I and switch II regions translates Ras conformation into a downstream response by diminishing effector binding to GDP-bound Ras and favoring effector binding to GTP-bound Ras. The other G loops contact the guanine base, phosphate groups, and Mg2+ ion (15). In addition to mediating effector binding, the two switch regions of Ras interact with proteins that regulate the GDP/GTP cycle: guanine nucleotide exchange factors (GEFs) and GTPase activating proteins (GAPs).

Ras possesses a weak GDP/GTP exchange activity. Because the cellular levels of GTP are 20-fold greater than that of GDP, the dissociation of GDP from Ras favors the formation of Ras-GTP. Ras also possesses low intrinsic GTPase enzymatic activity, which stimulates hydrolysis of bound GTP to GDP to cycle Ras back to its inactive GDP-bound state. The intrinsic exchange and hydrolytic activity of Ras is not sufficient to promote rapid GDP/GTP cycling. Ras-specific GEFs and GAPs are critical to accelerate guanine nucleotide cycling and hence to promote a response to extracellular stimuli. RasGEFs (e.g., Sos, RasGRF, or RasGRP) promote the exchange of GDP for GTP and activate Ras signaling activity (16, 18). RasGAPs (e.g., p120 RasGAP, or NF1 neurofibromin) stimulate hydrolysis of GTP to GDP and terminate signaling activity (19). The activity of GEFs and GAPs enable Ras proteins to signal downstream to their effector molecules in a regulated manner.

Subcellular localization of Ras

In addition to nucleotide binding, the signaling activity of Ras is governed by subcellular localization. Ras proteins terminate in a tetrapeptide CAAX motif, where C is cysteine, A is any aliphatic amino acid, and X is serine or methionine. Ras proteins are synthesized initially as cytoplasmic pro- teins that have no signaling activity (see Fig. 3). Ras then undergoes a series of CAAX-directed posttranslational modifications. Farnesyltransferase (FTase) catalyzes a covalent addition of a 15-carbon farnesyl (isoprenyl) group to the cysteine of the CAAX motif, followed by Ras converting enzyme 1 (Rce1)-catalyzed proteolytic cleavage of the CAAX tetrapeptide motif, which signals for posttranslational modifications that increase the hydrophobic nature of Ras. This region also contains additional sequences that promote membrane association and direct subcellular localization of Ras.
Figure 2: Schematic of Ras small GTP binding and hydrolyzing proteins. The five G boxes (G1-G5, yellow) comprise the catalytic domain of Ras. The hypervariable region (red) at the carboxyl terminus contains the second signal that is critical for proper subcellular localization. The CAAX box (blue) is post-translationally modified by a 15-carbon isoprenyl lipid and is required for membrane association. The sequence of each motif for H-, N-, and K-Ras4B is given, as well as the consensus sequence found in all G domain-containing proteins. Mutation of residues highlighted in green (at G12, G13, or Q61) results in a constitutively activated protein. Conversely, mutation of the residue highlighted in red (S17) creates a dominant negative protein. Alternative names for some motifs, as well as their main function in Ras biochemistry, are listed as well.

residues (20), and isoprenyl/cysteine-directed carboxy methyltransferase (Icmt)-catalyzed carboxylmethylation of the now terminal prenylated cysteine residue. The prenyl residue inserts into cellular membranes, which bthers Ras to the cytosolic face of cellular membranes (21, 23). Point mutations engineered in the carboxyl terminus of Ras proteins (such as a cysteine to serine substitution in the CAAX box) as well as pharmacologic inhibitors (e.g., FTase inhibitors or FTIs) disrupt this posttranslational processing and cause mislocalization of Ras (24, 27).

The three CAAX-signaled modifications increase the overall hydrophobicity of the carboxyl terminus of Ras and are necessary to promote Ras membrane association. However, these modifications alone are not sufficient to direct full plasma membrane association and signaling activity. At least two additional motifs exist at the carboxyl terminus of Ras proteins, which serve to facilitate plasma membrane association and direct Ras proteins to discrete membrane subdomains. These motifs function as secondary signals and are composed either of a stretch of basic amino acids (K-Ras4B) or of cysteine(s) that are palmitoylated (H-Ras, N-Ras, K-Ras4A) and positioned immediately upstream of the CAAX motif (28, 30) (see Fig. 3). The addition of a palmitate fatty acid is catalyzed by a protein acyltransferase, which forms a reversible thioester bond between the cysteine and the palmitoyl group (31, 32).

Effector interaction

The biologic function of Ras is mediated by its ability to activate downstream cytoplasmic signaling networks through binding to a panel of effector proteins. The Ras-effector interaction is mediated through interactions between the core effector domain (amino acids 32–40) of Ras (see Fig. 2) and the specific residues within the Ras-binding domain (RBD) or Ras-association domain (RA) found in most Ras effectors. The RBDs and RA domains of effector proteins do not exhibit primary sequence homology, but instead they share a common tertiary structure that consists of an ubiquitin superfold, which forms critical contacts with Ras (33).

Figure 3: Ras membrane targeting is a multi-step process. A native Ras protein is found in the cytoplasm of the cell. The cysteine of the CAAX motif is posttranslationally covalently modified by a farnesyl (red) lipid in a reaction catalyzed by farnesyltransferase (FTase). A series of posttranslational processing steps catalyzed by the indicated enzymes at the endoplasmic reticulum cleave the AAX residues and modify the now terminal prenylated cysteine with a carboxymethyl group. A secondary signal, either lysine residues in the case of K-Ras4B or cysteine bound to palmitate in the case of N-Ras and H-Ras is required for final membrane location. Mutating the cysteine of the CAAX motif to a serine (Ras-SAAX) precludes the addition of an isoprenyl group and all subsequent modifications, resulting in a cytosolic, inactive mutant protein.

At least 10 distinct families of Ras effector proteins have been identified, with the Raf serine/threonine kinases, phosphatidylinositol 3-kinases (PI3K), and RasGEFs being the most extensively characterized (3) (see Fig. 4). Beyond their shared ability to bind preferentially to Ras-GTP, Ras effectors possess highly divergent biochemical and biologic functions.
members) can contribute to apoptosis (34).

RalBP1 RalGEF functions as an activator of the Ras-like (RalA and RalB) small proteins. The phosphatase and tensin homolog (PTEN) tumor suppressor promotes activation of the AKT serine/threonine kinase and other signaling proteins. The phosphoinositol 3,4,5-triphosphate (PIP3) second messenger regulates phosphoinositide metabolism and formation of the leading to subsequent phosphorylation and activation of the Elk-1 transcription factor and changes in gene expression. PI3K activation regulates phosphatidylinositol metabolism and formation of the phosphatidylinositol 3,4,5-triphosphate (PIP3) second messenger. PIP3 promotes activation of the AKT serine/threonine kinase and other signaling proteins. The phosphatase and tensin homolog (PTEN) tumor suppressor protein antagonizes the activity of PI3K by dephosphorylation of PIP3.

Ras proteins can be monitored through another fluorescence-based assay. The fluorescence of an N-methylanthraniloyl derivative of mant-GTP increases ∼20% when binding to small GTPases. The intrinsic and GAP-stimulated GTP hydrolysis can be measured with real-time fluorescence-based assays. The fluorescence of an N-methylanthraniloyl derivative of mant-GTP increases ∼20% when binding to small GTPases. The intrinsic and GAP-stimulated GTP hydrolysis can be measured with real-time fluorescence-based assays.

Various protein–protein interaction methods have been developed to study the GTP binding, exchange, and hydrolytic activities of Ras. The in vitro ultra rate of nucleotide exchange and GTP hydrolysis can be measured with real-time fluorescence-based assays. The fluorescence of an N-methylanthraniloyl derivative of GTP (mant-GTP) increases ∼20% when binding to small GTPases, and therefore can be used to monitor the rate of nucleotide exchange and incorporation of mant-GTP into Ras (38). The intrinsic and GAP-stimulated GTPase activities of Ras proteins can be monitored through another fluorescence-based technique.

### Ras structural mutants

The creation and application of Ras mutants that display gain- or loss-of-function phenotypes has provided the foundation for our knowledge of the basic function and regulation of Ras proteins (see Table 1). By analogy to Ras, similar mutations have been introduced in many Ras superfamily proteins as well as other unrelated GTPases. These mutant proteins, particularly mutants of the Rho family small GTPases, have proven to be powerful reagents for functional studies.

Constitutively active Ras proteins can be generated through substitutions of the preserved glycine at position 12 or glutamine at position 61, for example, with valine (G12V) or leucine (Q61L), respectively (see Fig. 2). Essentially, any amino acid substitution at these two positions impairs the intrinsic and GAP-stimulated GTP hydrolysis rate, which results in the formation of a constitutively GTP-bound and chronically active protein. These activated mutants can be used to determine the downstream signaling activities of Ras proteins.

Conversely, Ras with a substitution of serine at position 17 with asparagine (S17N) binds to and forms nonproductive complexes with RasGEFs (see Fig. 2). Because each RasGEF can activate several Ras proteins, Ras(S17N) acts as a dominant negative inhibitory protein that inhibits activation of all Ras proteins. This dominant negative reagent can be used to determine whether the biologic activity of a particular extracellular stimulus is mediated through RasGEF-dependent activation of Ras.

### Ras GEFS and downstream signaling

Figure 4 Ras activates multiple effector pathways. GTP-bound active Ras travels to and stimulates a myriad of cytoplasmic signaling cascades. The three most extensively characterized effector pathways (RalGEF, Raf, and PI3K) are shown. Signal propagation occurs via phosphorylation of Raf, leading to subsequent phosphorylation and activation of the ERK1/2 MAPK pathway. PI3K activation regulates phosphoinositide metabolism and formation of the phosphatidylinositol 3,4,5-triphosphate (PIP3) second messenger. PIP3 promotes activation of the AKT serine/threonine kinase and other signaling proteins. The phosphatase and tensin homolog (PTEN) tumor suppressor protein antagonizes the activity of PI3K by dephosphorylation of PIP3.

### Chemical Tools and Techniques

The research on Ras proteins has resulted in the development of an impressive collection of reagents, assays, and techniques to dissect the Ras function. These techniques include mutants of Ras proteins, biochemical assays to monitor Ras function in vitro and in vivo, pharmacologic and genetic tools to dissect Ras signaling, as well as more recent approaches that apply RNA interference (RNAi) and other gene silencing techniques, and fluorescence-based protein assays to monitor the spatial and the temporal activities of Ras. Tables 1–3 provide a comprehensive list of the tools and reagents available to study Ras biology.
Table 1. Genetic reagents for activation of Ras and Ras effector pathways

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ras</td>
<td>Ras(G12V), Ras(Q61L) - GTPase-deficient, GAP-insensitive, constitutively activated mutants. Based on mutations detected in RAS alleles found in human cancers. RasGRP1 - RasGEF. Activated by treatment with phorbol esters.</td>
</tr>
<tr>
<td>Raf-MEK-ERK Pathway</td>
<td>H-Ras(G12V/T35S) - Effector domain mutant of activated H-Ras that binds to and activates Raf but not PI3K or RalGEF. Raf-CAAX - Human c-Raf-1 chimeric protein with carboxyl-terminal 18 amino acid membrane-targeting sequence of K-Ras(4B). Persistent membrane location and signaling activity.</td>
</tr>
<tr>
<td>B-Raf(V600E)</td>
<td>Human B-Raf with the V600E (formerly V599E) missense mutation observed in most mutated BNAF alleles found in human cancers.</td>
</tr>
<tr>
<td>Raf-22W</td>
<td>Truncation of 305 amino-terminal amino acids in human c-Raf-1. Lacks the negative regulatory sequences that lie upstream of kinase domain.</td>
</tr>
<tr>
<td>MEK(S218D/S222D)</td>
<td>Missense substitutions at Raf serine phosphorylation sites with charged amino acids to mimic persistent phosphorylation.</td>
</tr>
<tr>
<td>MEK-ED</td>
<td>N-terminally truncated MEK, with missense substitutions at Raf serine phosphorylation sites (S218/S222D) with charged amino acids to mimic persistent phosphorylation.</td>
</tr>
<tr>
<td>PI3K-AKT Pathway</td>
<td>H-Ras(G12V/Y40C) - Effector domain mutant of activated H-Ras that binds to and preferentially activates PI3K but not Raf or RalGEF.</td>
</tr>
<tr>
<td>p110CAAX</td>
<td>p110γ chimeric protein terminating with the C-terminal 18 amino acids plasma membrane targeting sequence of K-Ras(4B). Persistent membrane location and signaling activity.</td>
</tr>
<tr>
<td>p110α(E545K), p110α(H1047R)</td>
<td>Corresponding gene missense mutations observed in most mutant alleles of the gene encoding p110α (PIK3CA) detected in human cancers.</td>
</tr>
<tr>
<td>Interfering RNA for PTEN</td>
<td>Suppresses PTEN expression, prevents PTEN-mediated conversion of the PI3K product phosphoinositol 3,4,5-triphosphate (PIP3) to the PI3K substrate phosphoinositol 4,5-diphosphate (PIP2).</td>
</tr>
<tr>
<td>Mγ-AKT</td>
<td>Fusion protein of AKT with an amino-terminal myristoylation signal sequence. This fatty acid modification promotes persistent plasma membrane association and signaling activity.</td>
</tr>
<tr>
<td>RalGEF-Ral Pathway</td>
<td>Ral(G23V), Ral(Q71L) - Human GTPase-deficient mutants of RalA and RalB. Analogous to the Ras(G12V) and Ras(Q61L) mutants.</td>
</tr>
</tbody>
</table>
Analyses of Ras subcellular localization

When expressed in quiescent cells where Ras is inactive, the GFP-Raf-RBD probe is distributed homogeneously throughout the cell. When cells are stimulated with a growth factor or through a novel method that detects free thiol groups in proteins (50, 51). This assay involves blocking all available free thiol groups, presumably only at the sites of palmitate linkage. The thiol group can then be detected using a biotinylated thiol group, presumably only at the sites of palmitate linkage. The thiol group can then be detected using a biotinylated thiol reagent.

Analyses of Ras subcellular localization and posttranslational processing

Proper localization of Ras is essential for function, and several pharmacologic and genetic tools are available to study the posttranslational processing and subcellular localization of Ras. Note that epitope or fluorescent tags must be added at the amino terminus of Ras so as not to disrupt posttranslational processing of the carboxyl terminus. The localization of GFP-Ras can be monitored in live cells through confocal or widefield microscopy. Alternatively, anti-Ras antibodies can be applied to detect the location of endogenous protein in fixed cells. Similarly, antibodies that recognize epitope tags can be used to evaluate the subcellular localization of ectopically expressed, amino-terminal epitope-tagged Ras proteins. These antibodies include the commercially available antibodies that recognize peptide sequences found in the influenza hemagglutinin antigen (YPYDVPDYA), the human c-Myc transcription factor (LDEESILKQE), and the FLAG epitope (DYKDDDDK).

The treatment of cells that express GFP- or epitope-tagged Ras with various pharmacologic inhibitors is also useful to determine which posttranslational modifications are required for Ras membrane association. For example, treatment of cells that express GFP-H-Ras with FTIs causes GFP-H-Ras to mistarget to the cytoplasm, which indicates that farnesylation is required for membrane association. Compounds that block protein palmitoylation, such as 2-bromopalmmitate, can also be used to determine whether palmitoylation is required for Ras-related proteins to associate with the plasma membrane (46–48).

Various assays are available for direct analyses of the posttranslational modification of Ras by lipids. Recombinant FTase together with recombinant Ras can be used to evaluate the farnesylation of Ras. Rabbit reticulocyte lysate can be used for in vitro transcription/translation analyses of all CAAX-signaled modifications. Direct measurement of palmitoylation can be assessed by evaluating incorporation of a tritiated palmitate analog (49) or through a novel method that detects free thiol groups in proteins (50, 51). This assay involves blocking all available free cysteines in a protein and then chemically cleaving thiolester bonds that link palmitate to cysteine. Cleavage generates a free thiol group, presumably only at the sites of palmitate linkage. The thiol group can then be detected using a biotinylated thiol reagent.

Pharmacologic inhibitors of the enzymes involved in AAX proteolysis and carboxymethylation steps are available. However, as with all pharmacologic inhibitors, these compounds are likely to have off-target activities (52). Indeed, genetic approaches have provided more specific means to evaluate the role of these modifications in promoting Ras membrane association. In particular, mouse embryo fibroblast cell lines derived from mice that are deficient in Rce1 (RCE1−/−) or Icmt (ICMT−/−) are useful to determine the role of these two enzymatic modifications for Ras membrane association, signaling, and biologic activity. As mentioned, after protein prenylation, the AAX residues of Ras are cleaved by Rce1 (20). This step is critical for proper localization; exogenous H-, N-, and K-Ras are mislocalized in RCE1−/− mouse embryo fibroblasts (53). After proteolytic cleavage of the AAX residues, the prenylated cysteine residue is carboxymethylated by Icmt. As in RCE1−/− mouse embryo fibroblasts (53), after proteolytic cleavage of the AAX residues, the prenylated cysteine residue is carboxymethylated by Icmt. As in RCE1−/− mouse embryo fibroblasts (53), after proteolytic cleavage of the AAX residues, the prenylated cysteine residue is carboxymethylated by Icmt.

### Table 2: Reagents for biochemical detection of Ras and Ras effector pathway activation

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ras GST-Ras-RBD</td>
<td>Glutathione-S-transferase (GST) fusion protein with the isolated Ras binding domain (RBD) of c-Raf-1. Binds preferentially to activated, GTP-bound Ras proteins.</td>
</tr>
<tr>
<td>Raf-MEK-ERK Pathway</td>
<td>Antibody for Western blot detection of phosphorylated and activated MEK when phosphorylated by Raf at S221. Antibody for Western blot detection of phosphorylated and activated ERK when phosphorylated by MEK at T202 and Y204. Phosphorylation of these residues activates ERK catalytic function.</td>
</tr>
<tr>
<td>PI3K-AKT Pathway</td>
<td>Antibody for Western blot detection of phosphorylated and activated AKT when phosphorylated by PDK1 at T380</td>
</tr>
<tr>
<td>RaGEF-Ral Pathway</td>
<td>GST fusion protein with the isolated Ral binding domain (RBD) of the RaBPI (amino acids 397–518) effector. Binds preferentially to activated, GTP-bound RaL, and RaB proteins.</td>
</tr>
</tbody>
</table>
cells, all three Ras proteins are mislocalized when expressed in ICMT−/− cells. Interestingly, a recent study suggests that the membrane association of CAAX-containing Ras family GT-Pases that are modified by geranylgeranylation do not require these modifications (S3).

Ras trafficking to cellular membranes can be measured by fluorescence recovery after photobleaching (FRAP) and fluorescence loss in photobleaching (FLIP) (S4). Both techniques rely on the expression of fluorescent-labeled Ras proteins to monitor different parameters of Ras movement across and between cellular membranes. FRAP involves photobleaching a membrane subdomain and measuring the kinetics of fluorescence recovery—and hence Ras trafficking—into the bleached area. With FLIP, a cellular membrane is photobleached repeatedly and the subsequent intercellular movement of the photobleached area is monitored.

**Analyses of Ras effector activation and function**

In addition to the use of Ras effector domain mutants, structural mutants of Ras effectors and pharmacologic inhibitors of effector signaling are valuable to study the role of specific effector signaling pathways in Ras function. Tumor-derived, constitutively activated mutants, as well as lab-designed and generated mutants, have been particularly useful reagents for these studies (see Tables 2 and 3). For example, the addition of the carboxyl-terminal CAAX-contiguous sequences that target K-Ras4B to the plasma membrane onto c-Raf-1, the Raf-1 GEF, Raf, and the p110α catalytic subunit of PI3K has created effector fusion proteins that are constitutively membrane-bound and active. Membrane localization activates effector function because Ras activates effectors, in part, by promoting their association with the plasma membrane. More recently, human disease-derived mutants have provided physiologically relevant activated versions of these effectors, with the identification of missense mutational activation of B-Raf and p110α in human tumors, and activated MEK1 and MEK2 in cardio-facio-cutaneous syndrome (S1). Novel activating mutations in Ras and B-Raf, as well as the Sos RasGEF, have also been identified in developmental syndromes. These genetic variants may also be useful reagents to study Ras signaling and function (S5, S6). Activated effectors can be used to determine whether the activation of a specific effector pathway alone is sufficient to mediate a specific Ras function.

Various approaches have been useful to block the function of a specific effector and to determine its necessity and role in Ras function. First, catalytically dead mutants of Ras effectors or their substrates function as dominant negative mutants that, when ectopically expressed, block the function of the endogenous effector (see Tables 2 and 3). These mutants include kinase-dead mutants of c-Raf-1, MEK, and ERK that block the Raf effector pathway. Second, selective pharmacologic inhibitors of the MEK (PD98059 and U0126) and PI3K (LY294002) kinase pathways have been useful to study Raf and PI3K effector signaling. Recently, RNAi has been applied to impair specific effector function selectively (S7, S8). However, this approach may be limited by the fact that a majority of effectors also have highly related isoforms that are expressed broadly and may have overlapping functions. Finally, the use of mice that are deficient in effector function, as well as cells derived from these animals, has been very effective to determine the role of specific effectors in Ras-mediated oncogenesis. For example, although genetic loss of Tiam1, PLC epsilon, or Raf-GEF does not impair mouse development, each knockout mouse has demonstrated elegantly that this specific PI3K isoform was necessary for Ras-mediated lung and skin oncogenesis (S2). Hence, the development of similar mouse models will be useful to evaluate the role of Ras effectors that are normally essential for development, because of their roles in oncogenesis.

**Conclusions**

The decades of research on Ras proteins have yielded impressive insight into the biologic and chemical properties of Ras, and they have established many important paradigms that have defined key concepts in signal transduction. Concurrent with these studies has been the development and application of an armamentarium of reagents, genetic and pharmacologic tools, and assays with which to study Ras. These technologic developments have also aided studies greatly to evaluate the biophysical, structural, and biologic properties of other Ras superfamily proteins. These studies reveal that despite a shared simple fundamental biochemical function as GDP/GTP-regulated binary switches, Ras family proteins display an incredible diversity in the mechanisms by which they are regulated and in how their subcellular localization is controlled. These tools have allowed additional delineation of Ras function and an appreciation of their involvement in virtually all facets of cell physiology and behavior. Our earlier depiction of Ras as a relay switch at the plasma membrane in a simple linear signaling pathway has been revised to reflect the current model: Ras is a central component of a highly complex and dynamic signaling network with discrete functions at several subcellular locations. In summary, whereas great strides have been made to understand Ras activation, localization, and signaling, much need still exists for focused research in each of these areas. New methodologies and technologies that are developed will undoubtedly facilitate additional advances. Because we have been surprised by many recent new discoveries regarding Ras, one must expect the unexpected in Ras signaling and biology.

**References**

Table 3 Pharmacologic and genetic inhibitors of Ras and Rho effector pathways

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Description</th>
</tr>
</thead>
</table>

**Raf-MEK-ERK Pathway**
- Sorafenib (BAY 43-9006): Cell-permeable ATP-competitive inhibitor of Raf kinase activity. Also a potent inhibitor of a variety of other protein kinases.
- MEK1(K97A) mutant: Dominant negative MEK1 mutant.
- MKP-1: ERK dual-specificity protein phosphatase; removes phosphorylation of ERK1 and ERK2 at Y204 residues.

**PI3K-AKT Pathway**
- Wortmannin, LY294002: Cell-permeable inhibitors of PI3K family lipid kinases.
- PTEN: Truncation of p85 regulatory subunit; does not bind to p110.

**Ras Oncoproteins, Structure, Biochemistry and Biology of**
- Kim E, Arntztulek P, Otto JC, Taylor B, Ahtz S, Shannan K, Casey PJ, Young SG. Disruption of the mouse rclA gene results in...


29. Hancock JF, Paterson H, Marshall CJ. A polybasic domain or palmitoylation is required in addition to the caax motif to localize p21ras to the plasma membrane. Cell 1996;63:133–139.


Further Reading


See Also

Fluorescence Techniques: Proteins
Gene Silencing Techniques
Mitogen-Activated Protein Kinases (MAPKs): ERK
Protein-Protein Interactions, Tools to Study
Ras-Like Proteins
Signal transduction in cells has been an area of intense study for many years. However, until recently, the focus was on interactions between individual components rather than on the global behavior of the cell signaling networks. New experimental technologies, such as protein and DNA microarrays, high-throughput screening instrumentation, and diverse compound libraries, allow for many interactions to be examined simultaneously. Along with new experimental methods, quantitative models with mathematical methods, such as deterministic, stochastic, and statistical analyses and graph theory, have been useful in the mapping and analysis of intracellular signaling networks. Here we describe current experimental and computational approaches that are used to develop a global understanding of the complex behavior of intracellular signaling networks. We also discuss important applications of signaling network analysis, including the discovery of new drug targets, the identification of the signaling components responsible for the side (off-target) effects of drugs, and the development of combination therapies.

Currently it is estimated that the human genome contains 25,000 protein-encoding genes (1), which correspond to several hundred thousand possible protein species resulting from alternative splicing and posttranslational modifications. Of these presumed protein species, the functions and interactions of only a fraction are well understood. Among the well-studied proteins, many interactions result in complex networks that often form the basis for the complex physiologic functions. Moreover, protein expression levels vary among individuals and different cell or tissue types, and the distribution of proteins varies in the subcellular regions of each cell. Therefore, it is important to develop tools to analyze these systems as a whole.

Despite the growing need for global analysis of signaling networks, mapping individual pathways has been the main approach in studying signal transduction. Studies focused on defining how these pathways interact to form networks are relatively new. By mapping out signaling networks within cells, we can get an initial description of the rules that are applied in the development of these networks and the discovery of new trends or network properties. This review article discusses the interactions that comprise intracellular networks, the mechanisms underlying them, their conserved trends, and the emerging tools being developed to understand networks as a whole.

### Cell Signaling Networks

When a cell receives stimuli, the signals are propagated from one component to another, eventually forming out to form a network of signals that produces physiologic responses (phenotypes). Interactions within a set of interconnected signaling pathways can be represented as a network composed of components and connections. Many natural and engineered systems can be described as networks. These systems include the World Wide Web (2), networks of coauthors of scientific papers (3), and other social networks (4).

Biochemical interactions within cells can also be described as networks (5-7), and in this review we consider two intracellular networks, one involved in the stimulation of cell proliferation by the epidermal growth factor (EGF) and the other involved in the initial immune response to pathogens through the toll-like receptor (TLR) (Fig. 1a and 1b, respectively).

---

*Both authors have contributed equally and share credit as first authors.*
Mechanisms of Cellular Signaling

Signal initiation

The initial stimulus for a signaling pathway can be intracellular or extracellular. Intracellular stimuli typically are the result of changes in the cellular environment, such as the formation of reactive oxygen species or DNA damage. For extracellular signaling events, the process is initiated by a ligand binding to a receptor, often transmembrane proteins, such as the EGF receptor (EGFR) or the TLR. (Fig. 1a and 1b). Ligands can either be small molecules (hormones and neurotransmitters), peptides (insulin), or small proteins (growth factors). The naturally occurring ligands that have an affinity for EGFR are EGF, transforming growth factor (TGF-α), and β-cellulin (βC), all of which are small proteins, whereas the natural ligand for TLR is a macromolecule, lipopolysaccharide (LPS). Extracellular ligands can be secreted by other cells or generated by the cleavage of a larger precursor protein within the signaling cell. For example, EGF is generated by the proteolytic cleavage of the membrane-bound precursor, heparin-binding EGF-like factor (HB-EGF), by a matrix metalloproteinase (MMP) (9).

Biochemical reactions involved with signal propagation

Signal propagation within cells occurs by binding and enzymatic interactions that involve proteins, ions, and small molecules. Binding interactions typically occur between two or more proteins, proteins and lipids, or proteins and small molecules. Enzymatic reactions involved in signal propagation include phosphorylation, dephosphorylation, ubiquitination and sumoylation, controlled protein cleavage, and the generation of small signaling molecules such as cyclic adenosine monophosphate (cAMP), diacylglycerol (DAG), and inositol triphosphate (IP3). Many signaling reactions in intracellular networks are reversible chemical reactions, catalyzed in the forward direction by the enzyme and in the reverse direction by a different enzyme. One example of a reversible signaling reaction is the phosphorylation of a protein by a kinase and the subsequent dephosphorylation by a phosphatase. For instance, in the TLR pathway (Fig. 1b), the activated kinase TAK1 (TGF-activated kinase 1), phosphorylates Iκκ (Iκκ kinase), whereas the protein phosphatase 2C (PP2C) dephosphorylates phospho-Iκκ. In nonenzymatic signaling reactions, adaptor proteins act as mediators between the activated receptor and signaling enzymes by recruiting other signaling molecules to form a complex, which increase the activity of the receptor by cross-linking it to other adaptors. For example, Grb2 (growth factor receptor bound protein 2) binds to the phosphorylated EGFR and, in turn, recruiting adaptor proteins Gab1 (growth factor receptor bound protein 2-associated protein), mediates the kinase Aki (a-akt murine thymoma viral oncogene homolog 1) that mediates cell survival signaling (5, 10).

Patterns of signal flow: propagation, amplification, and loops

Signal transduction from the membrane to the nucleus occurs via interactions between signaling components. For example, one TLR pathway (Fig. 1b) starts with the activation of TLR3 by the binding of viral dsRNA. Next, the adaptor protein, TRIF (Toll/IL-1 receptor domain-containing adapter inducing IFN-β), is recruited, and TRIF mediates the phosphorylation of TANK-binding protein kinase 1 (TBK1) by an unidentified kinase. Phospho-TBK1 in turn phosphorylates the transcription factor, IFN regulatory factor-3 (IRF3), at five distinct sites. The phosphorylated IRF3 homodimersizes and translocates to the nucleus to regulate transcription (11, 12).

Cellular signals generally are not propagated through single, linear pathways. Instead, the signals have the tendency to branch or fan out into several or many pathways, which lead to one or several possible outcomes. An activated protein may have more than one protein that it activates or interacts with, such as Grb2 (Fig. 1a and 1b) and TRAF6 (tumor necrosis factor receptor-associated factor 6) and Tak1 (Fig. 1b, 3b). Nonlinear increases the complexity of the network and introduces crosstalk between pathways or networks. Networks for the EGFR pathway (5) and the TLR pathway (6) have been constructed and have resulted in highly detailed, complicated schematics that represent the networks in their entirety, hence, a portion of each network is depicted in Fig. 1 to create a meaningful visual representation of the network. It is notable in complete networks that patterns of signal propagation and amplification exist.

Signal flow within networks often is controlled and/or amplified by regulatory loops (Fig. 1c). The connectivity within regulatory loops has multiple configurations. For example, consider an upstream protein A that activates downstream protein B that in turn activates protein C. In a feed forward loop (FFL), protein A activates protein C. In a positive feedback loop (PFBL), protein C in turn activates protein A. In a negative feedback loop (NFBL), protein C inhibits protein A. In the TLR pathway, a FFL is found within the MAPK cascade that activates ATF2 (activating transcription factor 2) through both
Figure 1: Signaling networks and loops. Schematic interaction maps that represent the major components in the signaling networks emanating from (a) the epidermal growth factor receptor and (b) the toll-like receptor are shown. (c) Schematics that represent several types of regulatory loops that are found in intracellular signaling networks.

(a) Phosphate group
(b) the epidermal growth factor receptor and (b) the toll-like receptor are shown.
(c) Schematics that represent several types of regulatory loops that are found in intracellular signaling networks.
by facilitating Ras-mediated GTP hydrolysis to GDP (5). (Ras-GTPase activating protein), which inhibits Ras signaling example of a more distal NFBL is EGFR activation of Ras-GAP signaling by dephosphorylation of the receptor (5, 20). Another recruited to EGFR through Gab1 and Grb2 but inhibits EGFR signaling in cells under different conditions, such as stress, native ligand specificity. Some TLR ligands are involved in the TLR network; whereas, epithelial cells have relatively high expression level of TLRs, whereas epithelial cells have relatively high expression levels of growth factor receptors (25, 26).

Multiplicity in signaling pathways forms a complex network

Families of signaling receptor proteins most often include several isoforms that have varying affinities for ligands and adaptor proteins. Usually several to many ligands have affinity for one or more receptor. Once all possible ligand–receptor combinations and the multiplicity of recruited adaptor and signaling protein combinations are taken into account, the complexity of signaling networks becomes apparent. Take for example the ErbB signaling network. Four members exist: ErbB1 (EGFR), ErbB2 (Her-2/Neu), ErbB3, and ErbB4 (Fig. 1a). As depicted in Fig. 1a an abundance of adaptor proteins signal in the ErbB network. Ligand binding induces a conformational change that allows the ErbB receptors to homodimerize and crossphosphorylate at the C-terminal region of the intracellular domains (23, 24). The potential of heterodimerize and crossphosphorylate at the C-terminal region of the intracellular domains (23, 24). The potential of the receptors to homo- or heterodimerize with one another varies because of the expression level of the receptor, their affinities, and the amount and identity of ligand(s) present. The phosphorylated region provides a docking site for adaptor proteins, which have varying affinities for the different ErbB receptors. Specifically, the Src homology-2 (SH2) or phosphotyrosine binding (PTB) domains of adaptor proteins including Vav, Shc, Grb2, small heterodimeric partner (SHP), and phospholipase-Cγ (PLCγ), have affinity for the intracellular tyrosine-phosphorylated EGFR region (Fig. 1a). Many possible combinations of ligands, receptors, and recruited adaptor proteins can form the final signaling complex. Therefore, complexity in the signaling network is introduced early, at the level of the ligand. Further, the ligand governs the nature of the dimerized receptors which, in turn, modulate the mode and magnitude of the signal. At the ligand-receptor level, the TLR network (Fig. 1b) is similar to the EGFR network with respect to ligand specificity. Some TLR ligands are specific for one TLR, and others are not. In contrast to the EGFR network, one main adaptor protein MyD88 (myeloid differentiation primary response gene 88) has similar affinity for all TLR family members, with the exception of TLR3. Therefore, in the TLR pathway the nature and magnitude of the signal depends highly on the expression level of the receptors and the nature and relative amount of ligand(s) present. Because of the variation in the identity and level of expressed receptors across various cell types, different cells will respond differently to stimulation with particular ligands. The distribution of different receptors reflects the interplay of the signaling capacity of the receptor and its significance in the specialization of the cell type. For example, immune cells, such as dendritic cells, have a relatively high expression level of TLRs, whereas epithelial cells have relatively high expression levels of growth factor receptors (25, 26).

Experimental Approaches to Understanding Signaling Networks

The majority of published, reconstructed cellular signaling networks are built on the data that describe individual interactions and loops through classical biochemical techniques, such as immunoblot analysis, coimmunoprecipitation, binding, and enzymatic assays. However, to obtain global perspectives on interaction networks, new experimental approaches to obtain high-content data sets on interactions that make up cellular signaling networks recently have been developed. These fields serve two distinct but complementary purposes: First, to catalog and inventory genes, mRNAs, and proteins and their functions; and second and more pertinent, to develop signaling networks, to parse, on a global scale, the properties of signaling networks in cells under different conditions, such as stress, native ligand treatment, non-native agonist/antagonist treatment, or other. Also, to observe differences in different cell types, including healthy and normal cells and those in a disease state. Often, because of the heightened possibility of false positive or negative results because of their high-content nature, results from these high-content experiments are validated by single experiments that use more classical techniques. The experimental methodologies and examples of experiments in these fields will be discussed here; whereas, the mathematical and computational approaches including the analyses of the resultant large data sets will be discussed in the section entitled “Computational approaches that facilitate the understanding of intracellular networks.”
Genomic and proteomic methods

Genomics entails cataloging genes and genetic mutations and observing changes in gene expression levels under various conditions. Genomics can be used to catalog genetic mutations, polymorphisms, or amplifications that may play a role in disease by comparing gene sequences or levels from disease and normal genetic samples. The overall changes in gene expression or genetic mutations in a disease can help identify a critical signaling protein or pathway as a therapeutic target. Genetic analysis of cells that have undergone treatment to activate a specific signaling pathway can lead to signaling information, which provides a basis for the construction and analysis of the cell signaling network. For example, after treatment with an agent known to affect exclusively a pathway of interest, all the genes that are upregulated or downregulated by the pathway can be analyzed by genomics. This type of analysis may lead to information on the genetic programs that are outcomes of signaling pathways.

The technology to detect gene expression, coined DNA microarrays or DNA chips, was developed over the past decade (27). Recently, methods such as ChIP-chip (chromatin immunoprecipitation-microarray) (28), DamID (DNA adenine methyltransferase ID) (29), and PBMs (protein binding microarrays) (30), allow for a more mechanistic understanding of gene regulation (and chromosome structure). With transcription factor microarrays, one can detect transcription factors that are activated or inhibited under various conditions and/or different cell types.

Proteomics focuses on obtaining an inventory of proteins and their functions, as well as unfolding interaction networks and posttranslational modifications, such as phosphorylation (phospho-proteomics). Further, proteomic experiments can yield quantitative information such as rates and relative concentrations. A range of methodologies are used to generate proteomic data, ranging from low-throughput methods such as Western blotting and immunoprecipitation to high-throughput methods such as protein microarrays, yeast two-hybrid screens (31), and SILAC (stable isotope labeling by amino acids in culture) combined with quantitative tandem mass spectrometry (32). Important advances in the technologies involved in proteomic studies allow for a more detailed, global understanding of interaction networks.

A protein microarray study by Jones et al. (33) was designed to test the affinity of all proteins that contain SH2 and PTB domains, which specify binding to the phosphorylated domains of all four ErbB family members. The study uncovered that some ErbB family members were more promiscuous than others, which has important implications for the ErbB signaling network in general and specifically in that the promiscuous ErbB family members are much more commonly overexpressed in several cancer cell types. Furthermore, 116 new ErbB interaction partners were discovered. A further study by Schürr et al. (34) was also designed to identify all interaction partners for the phosphorylated ErbB family members but using a novel methodology that combines SILAC and LC-MS/MS. This study defined the specific ErbB sites where the interaction partners bind.

Many of these studies have resolved the temporal and spatial aspects of signaling networks. The dynamics of EGF stimulation was examined by Blagoev et al. (35) in a phospho-proteomics study by examining the phosphorylation pattern of EGFR after different periods of EGF treatment. The spatial aspects of signaling can be captured by analyzing different fractions of the cell such as the nucleus, organelles, vesicles, membrane-bound proteins, or cytoplasmic proteins. Several recent studies have analyzed holistically the signaling components of purified cellular organelles, such as synaptic vesicles (36) and the ER-Golgi apparatus (37).

Chemical genetic approaches

Chemical genetics (38, 39) can be considered a combination of chemistry and genetics or proteomics because it is based on the use of small molecules to modulate protein or gene function. This method is similar to pharmacology, except that the experiments are designed in a similar fashion to "omics" experiments, in which an entire set of genes, proteins, or phenotypes is examined simultaneously. The effects of modulating protein function with a small molecule or library of small molecules are examined either phenotypically or at the signaling level. Chemical genetic experiments are instrumental in understanding how signaling networks are affected by the down- or upregulation of one or more signaling proteins by a drug or other small molecule(s).

The two types of chemical genetic methods, "forward" and "reverse," lead to different results. In a "forward" experiment, small molecules are used to probe for a desired phenotype, and the protein that binds the small molecule and induces the phenotype can be subsequently identified. In an example of a "forward" chemical genetic experiment, the small molecule necrostatin-1 was discovered by screening a 15,000-compound library for an inhibitor of a novel type of cell death that is unique from apoptosis or necrosis, coined "necroptosis" (40). Other forward chemical genetic experiments have led to information on the genetic basis of neural stem cells (41) and on the role of copper and lysyl oxidase in notochord morphogenesis during development (42). Furthermore, forward chemical genetic experiments have been used to identify unknown targets of clinically approved drugs and can lead to the identification of new potential drug targets (38, 39, 43).

In a "reverse" chemical genetic experiment, a protein with a known function is modulated with small molecules, either to simply discover a small molecule modulator of the protein or to examine the downstream effects of modulating the protein to further dissect the mechanism of its function. In an example of a "reverse" chemical genetic experiment, a protein kinase of interest is mutated to be specifically inhibited by a bulky, synthetic ATP analog, which cannot bind all other protein kinases. These "reverse" chemical genetic experiments have led to the elucidation of numerous signaling pathways including the cell cycle pathway by inhibiting CDK-activating kinase 7 (cdk7) (44), c-Jun N-terminal kinase (JNK) signal transduction pathway (45, 46), G protein-coupled receptor kinases (GRKs)-regulated receptor regulation pathway (47), the actin assembly process by inhibiting vinculin (48), and mitosis and cell division by inhibiting Plk-1 (49). The consequences or potential off-targets of modulating a protein with a small molecule or drug can be elucidated by using reverse chemical genetics.
Collaborative research initiatives for the use of chemical genetics to facilitate discovery of potential cancer (or other) therapeutics by the Initiative for Chemical Genetics at the National Cancer Institute (50) and to facilitate the discovery of new targets and drug candidates by the Chemical Genomics Center at the National Institutes of Health (51) have been initiated. These initiatives will result in databases that contain information that most likely will be valuable to the study of cellular signaling networks.

The design and synthesis of chemical compound libraries

The synthesis, collection, and development of chemical libraries are essential for the global analysis of intracellular signaling networks, chemical genetics, and drug discovery. Commercial sources of compound libraries are available and often categorized by the type of compounds they contain. Library categories include natural product, drug-like, various molecular weight ranges, various inhibitor classes, and peptide or peptide-like (38, 52). Many institutions have a screening facility where purchased libraries and/or libraries obtained from archives of compounds synthesized in-house are available for assaying.

Most commonly, a commercial library, containing a subset of compounds that matches a desired set of properties (if known), is screened in an initial study. Once hits are obtained and verified, a small library of compounds is synthesized to produce a set of compounds that are chemically different from the original hit structure. This process allows for hits with higher potency and elucidates information of the structure-activity relationship within the system of interest. Such synthetic libraries of chemically diverse compounds have been made possible through combinatorial chemistry (52-56) and diversity-oriented synthesis (52).

There continues to be a strong motivation in the chemical biology field to synthesize new libraries and build on existing ones, as well as to improve combinatorial chemistry approaches. The data collected from chemical genetic experiments adds to the delineation of the proteome and a complete cell signaling network and promotes the discovery of new therapeutic targets. To make a compilation of the data generated, Chembank (affiliated with the Initiative for Chemical Genetics by the NCI) was created. Chembank is an initiative to create a public database containing all data obtained from screening compound libraries for various purposes.

Computational Approaches that Facilitate the Understanding of Intracellular Networks

In the past decade, computational analyses, based on mathematical methods and experimental data, have been recognized as powerful tools to understand the complexity that is inherent to biological systems. The integration of computational and experimental approaches to construct and analyze cell signaling networks generally is a multistep process involving collaborative efforts between the two fields. The major steps involved in the development of mathematical models that yield systems-level insights are outlined in Fig. 2.

The first step in constructing a cell signaling network model is to generate an in silico interaction network. Signaling components of interest are identified, and data on binary interactions are extracted from the experimental literature. Often times, the creation of an entire interaction map for a large network is beyond the scope of one laboratory; therefore, public databases have been created in which newly discovered proteins and/or protein interactions are deposited (57-63). However, often data are included from studies that cover a broad range of protein interactions, such as proteomic studies or yeast two-hybrid screens, and frequently contain many potential false positives or negatives. Thus, it becomes necessary to critically examine each reference to verify the interaction data reported, as this is in silico network is the basis of further study.

Once the reaction network is created, an appropriate mathematical approach is identified according to the information available and the biological question(s) to be addressed. In this section, some mathematical approaches that have been applied successfully in the analysis of biological signaling networks will be discussed. These methods are summarized in Table 1 (64-85).

Statistical approaches

High-throughput experiments such as those involved with genomic and proteomic studies (e.g., microarray studies and SILAC-mass spectrometry studies, discussed in the section entitled “Experimental approaches to understanding signaling networks”) are useful for understanding the interdependence of signaling components in the network. However, these types of experiments generally yield large-scale data sets that often are analyzed by computational algorithms, based on statistics. The basic concept behind this computational statistical analysis is to identify clusters or groups of signaling components that show similar trends. This analysis is used to extract the relationships between different signaling components in a network from these large-scale data sets. These kinds of analyses have been used for analyzing microarray data to identify patterns that correlate with distinct physiologic and pathophysiologic states. For example, an hierarchical algorithm has been used to identify different types of cancers in human soft tissue tumors (86), and K-mean clustering has been used to identify molecular subtypes of brain tumors (87). However, these types of statistical clustering analysis do not exclude the noise or artifacts induced by the experimental techniques. To overcome this limitation, singular value decomposition (a modified type of principal component analysis) has been used to identify which clusters of signaling components have a significant amount of influence in changing patterns of biological behaviors (88-91). Another statistical technique, partial least square regression analysis on multidimensional experiments (signaling components, time, and ligands), has been used to predict new interactions between components in a network, regulating apoptosis (92).
Signaling Networks in Biology

Identify key signaling components

Construct interaction network based on experimental data

Identify the appropriate mathematical representation

Analyze network using simulations and make experimental predictions

Test predictions with experiments

New systems level understanding

Graph theory

Statistical analysis is a powerful tool for understanding interdependence between groups of self-similar behaviors, especially when a large data set is being analyzed. However, it does not yield information about the topology of the interaction network. Graph theory-based approaches have been developed to study the topological properties of a large biological network. In graph theory, each signaling component in the network is represented as a node and identified by its degree, which represents the number of interacting partners. Within the context of the entire signaling network, the relationship between the degree of individual nodes and the degree of distribution of the whole network can be characterized. This kind of analysis has shown many intrinsic topological properties, such as scale-free configuration, small world configuration, and modular organization of networks across different cell types and species (2, 7, 22, 93, 94).

Boolean analysis

Although it is possible to analyze the topological properties of the network using graph theory, the states of the signaling components (phosphorylated/dephosphorylated, bound/unbound) often are not accounted for in such analyses. Boolean analysis can be used to understand such dynamics. In Boolean analysis, the signaling components in the network are identified by strings of 0’s and 1’s. The activated state of the protein is denoted by ‘1’, and the inactivated state is denoted by ‘0’. Changes in states of components can be computed as a function of various perturbations to the network.

To illustrate the concept of Boolean analysis in a biological context, an example is taken from the TLR pathway (Fig. 1c). To initiate the induction of the IFNβ gene in TLR signaling pathway, it is necessary for three transcription factor complexes, ATF2-c-jun heterodimer, two IRF3 homodimers, and NF-κB, to bind simultaneously at the promoter region. This biological condition can be represented in the Boolean form in the following way: Each of the four transcription factor complexes, ATF2-c-jun, two IRF3 homodimers and NF-κB, and the gene IFNβ is expressed as strings of five 0s and 1s that lead to 2^5 (or 32) combinations. For instance, the representation ‘11111’ denotes that all transcription factor complexes are bound to the promoter region and IFNβ gene expression occurs; ‘01000’ denotes that only one of two IRF3 homodimers are bound to the promoter region and, hence, IFNβ gene expression does not occur. A simulation is performed to analyze the effect of the change of state of an upstream signaling component on the downstream component in the signaling network, and rules, based on Boolean operators, are applied (95, 96).

Boolean analysis has been used to understand the effect of different states of signaling components on their downstream signaling components under various conditions. For example, Albert and Othmer (64) used Boolean analysis of a Drosophila melanogaster network to validate results from previously published microarray data and predicted the crucial role of the two genes, wingless and sloppy paired, in segment polarity. Gonzales et al. (65) mutated different genes in silico and used Boolean analysis to predict that the delay in the activation of two genes, Apterous and Notch, is essential to maintain the dorsal-ventral boundary of Drosophila melanogaster. Sarkar and Franz (97) predicted that the costimulation of T-cell receptors (TCR) and CD28 in T cells influences cell proliferation by providing a greater diversity of paths in the network. Maayan et al. (67) has used a combination of graph theory and Boolean analysis to obtain a distribution of redundant pathways in a cell signaling network.
Table 1 Various computational methods that have been used in analyzing signaling networks

<table>
<thead>
<tr>
<th>Computational methods</th>
<th>Descriptions</th>
<th>References</th>
</tr>
</thead>
</table>
| Deterministic analysis | Coupled biochemical systems
  - Reaction kinetics are represented by sets of ordinary differential equations (ODEs).
  - Rates of activation and deactivation of signaling components are dependent on activity of upstream signaling components.
  Spatially specified systems
  - Reaction kinetics and movement of signaling components are represented by partial differential equations (PDEs).
  - Useful for studies of reaction–diffusion dynamics of signaling components in two or three dimensions. | (64–70)     |
| Stochastic analysis   | Based on Monte Carlo theory and probabilities of reactions.
  - Facilitates modeling signaling components in small volumes and capturing stochastic fluctuations within the network. | (71–73)     |
| Graph theory          | Describes topology of networks and subnetworks, based on quantification of the number of nodes (signaling components) and links between them.
  - Dynamic properties of networks through Boolean analysis.
  - Network analysis based on probabilities (Markov chain and Bayesian) to identify paths and relationships between different nodes in the network. | (75–81)     |
| Statistical analysis  | Clustering based on correlations to identify group relationships.
  - Partial least square regression analysis to track pathways of signal flow.
  - Principal component analysis to identify significant signaling components. | (82–85)     |

Deterministic analysis

Although large-scale network analysis using graph theory or Boolean analysis can be very powerful in understanding the overall topological properties of the network, these tools do not take into consideration the rate of change of states of the signaling components with respect to time, explicitly. To understand the dynamics of activation or deactivation of a component in a signaling network, the most commonly used approach is deterministic analysis, which involves solving ordinary differential equations (ODEs). For deterministic analysis of biological systems, the initial concentrations and the kinetic parameters (K_m, K_d, and V_max) for enzymatic reactions and binding reactions (K_a or K_d) are required and need to be determined experimentally. Numerous examples exist in which deterministic analysis has provided critical insights into the dynamic behavior of a protein in a network under various conditions. One early effort in computational modeling, which uses a system of ODEs, is the EGF-EGFR reaction kinetics and internalization of the receptors (98–102). Bhalla and Iyengar (103) predicted the importance of feedback loops in a signaling network by illustrating bistable behavior of output proteins. Moreover, the sensitivity of the MAPK activity in response to a stimulus has been studied in detail by using systems of ODEs (104, 105). Bentele et al. (106) have studied apoptosis induced by the CD95 receptor by using deterministic analysis.

Stochastic analysis

The deterministic analysis provides the average behavior of signaling components in a network but does not include the results of fluctuations in the activation states of the signaling components (noise). Such analysis becomes essential when the volume of interest becomes in the scale of femtoliters because of the limitingly small number of molecules contained in such a small volume (1 fL of a 0.1 µM solution is equivalent to 60 molecules). Small volumes such as these are biologically relevant and present in cellular systems quite often, for example, in the spines of dendrites and endosomal regions. Moreover,
stochastic analysis becomes important in studying gene regulation where many genes have less than 10 copy numbers. Stochastic analyses, based on Monte Carlo theory and probabilities, have been used to capture the effects of fluctuations of an upstream signaling component on the kinetics of a downstream signaling component in the network. The most popular method for stochastic analysis is the Gillespie method (107), which is based on the probability of which reaction will take place within a defined time period. This type of simulation can give insights into the characteristics of signaling events at the single-molecule level. For example, in the analysis of an individual cell, noise introduced in enzymatic futile cycles can be strong enough to switch the system from one state of the cell to another (108). This technique has not gained much popularity over deterministic approaches, even though it is the only appropriate mathematical representation for certain cellular phenomena. This lack of popularity primarily is because of the following reasons: 1) Stochastic models need information at the molecular level as building blocks to construct the model properly and to validate simulation results obtained from the model. Experimental methods that yield molecular-level data have not been developed for biological systems, and 2) Stochastic simulations of a relatively small model, which contains less than 10 signaling components and their interactions, can pose large computational demands. However, application of the Gillespie method in small volume can reduce computation time to be comparable to that of deterministic simulation.

The Importance of Signaling Networks in Understanding Mechanisms that Underlie Disease and Drug Discovery

Cellular signaling networks are useful for drug discovery in three ways: 1) to develop an understanding of the biochemical mechanisms that underlie disease, which aids the discovery of novel drug targets, 2) to examine how current drugs affect the on-target (leading to therapeutic effect) and off-target (causing side effects and/or drug resistance) signaling pathways, which may lead to more informed prescribing methods and improved therapies, and 3) to lay out the interplay between signaling pathways in such a way that explains or facilitates the success of combination therapies. A dynamic relationship exists between the understanding of the biochemical mechanisms that underlie disease and drug discovery (Fig. 3). The compilation of data into a detailed signaling network is central in properly choosing a drug target (109). Many network models are in place, such as those described for the EGFR and TLR networks in this review, which can be applied to drug discovery (110). Before pursuing a drug target, it is important to understand as thoroughly as possible the consequences of regulating the target of choice because of the sheer expense involved with drug development and clinical trials. This understanding may be gained by analyzing a computational model of the signaling network involved, generating a knockout mouse, examining the chemical genetics of the target, and critically examining the current data on the drug target in the literature and databases (i.e., Chembank).

Of the total number of proteins that have a function we currently understand, only a small fraction is drug targets. The pool of drug targets is evolving with the introduction of new drug discovery technologies, such as the cell signaling network analyses described in this review, combinatorial chemistry, chemical genetics, molecular informatics, and advanced high-throughput screening technologies. With a new direction in the medical field toward personalized therapy regiments, it is increasingly important to expand the “druggable” network, and the enhanced global knowledge of signaling networks and the new drug discovery technology are indispensable in achieving this goal.

A thorough understanding of signaling analysis is essential to uncover new drug targets and aid in subsequently validating their potential, it also plays a role in studying roadblocks for approved pipeline drugs, such as developed drug resistance or off-target effects. Network analysis combined with carefully designed proteomic experiments aid in understanding the mechanisms of resistance and in developing methods to overcome the resistance. For example, a study by Chen et al. (111) combined signaling network analysis and a carefully designed proteomic experiment to illuminate the signaling network involved with the resistance of ovarian cancer cells to cisplatin treatment.

In the light of the abundant data on drug resistance in disease, many therapies are dosed in combinations to affect several pathways at once and/or gain a therapeutic advantage that stems from synergistic effects between drugs (112). An assembly of the data on the pathophysiology and biochemistry of cancer will allow treatment regimens to evolve toward targeting a network rather than a single protein (113). Efforts are currently underway to use a network analysis approach in understanding the effects of treatment with more than one drug (114). Alternatively, the combination of effects of two or more drugs can be used to uncover information about how the networks that each drug targets are connected (114). Combined with advances in high-throughput screening technologies, these approaches will be the powerhouse in the discovery of numerous novel therapeutic drug combinations and the most effective dosing schedules.

Perspective and Future Directions

Cell signaling networks are a rich source of targets for drug. A vast majority of drugs today are targeted to receptors or signaling components inside the cell such as protein kinases. The understanding that individual signaling pathways come together to form networks is very useful in looking for new drug targets not solely as individual components but also as pairs and triplets in interacting pathways for combination therapy for complex diseases. The convergence of computational approaches with high-throughput experimental analyses and chemical libraries offers substantial new opportunities not only to understand in depth the regulatory complexities within cells but also to design
and identify drugs that can be used as therapeutic agents for diseases that arise from alterations in the regulatory capabilities of cell signaling networks.

Acknowledgment

This work was supported by National Institutes of Health Grants numbers GM54508, DK38761 and a contract from National Institute of Allergy and Infectious Diseases.

References


84. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 
83. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis 
82. Belacel N, Wang Q, Cuperlovic-Culf M. Clustering methods for 
81. Sanchez L, Thieffry D. Segmenting the fly embryo: a logical 
80. Sanchez L, Thieffry D. A logical analysis of the Drosophila 
79. Sarkar A, Franza BR. A logical analysis of the process of T 
78. Sachs K, Perez O, Peer D, Lauffenburger DA, Nolan GP. 
77. Friedman N. Inferring cellular networks using probabilistic 
75. Gonzalez A, Chaouiya C, Thieffry D. Dynamical analysis of the 
73. Resat H, Ewald JA, Dixon DA, Wiley HS. An integrated model 
72. Bhalla US. Signaling in small subcellular volumes. II. Stochastic 
70. Ma L, Wagner J, Rice JJ, Hu W, Levine AJ, Stol ovitskiy G. A 
69. Lee S E, Hori Y, Chakraborty A. Low T cell receptor expression 
67. Hoffmann A, Levchenko A, Scott ML, Baltimore D. The 
66. Hayer A, Bhalla US. Molecular switches at the synapse emerge 
64. Sarkar A, Beard DA, Franza BR. Effect of binding in cyclic 
61. Waters CM, Oberg KC, Carpenter G, Overholser KA. Rate con- 
60. Starbuck C, Lauffenburger DA. Mathematical model for the 
59. Schoeberl, B., Eichler-Jonsson C. Gilles, E.D. and Muller, G. 
58. Shen-Orr S, Milo R, Mangan S, Alon U. Network motifs in the 
57. Lauffenburger DA. Receptors: models for binding, trafficking, 
56. Friedman JR. Ones and Zeros: Understanding Boolean Algebra, 
54. Shiva SG. Introduction to Logic Design. 1998. CRC Press, Boca 
53. Mischel P, SF N. Gene expression profiling identifies molecular 
52. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, 
51. Shai R, Shi T, Kremen T, Horvath S, Liu, L, Cloughesy T, 
50. Wang J, Delabie J, Aachem H, Smeland E, Myklebost O. Clus- 
49. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, 
48. Elenbaas B., Eichler-Jonsson C. Gilles, E.D. and Muller, G. 
47. Shiva SG. Introduction to Logic Design. 1998. CRC Press, Boca 
46. Receptors: models for binding, trafficking, and signal- 
45. Feng L, Eichler-Jonsson C, Gilles, E.D. and Muller, G. 
44. Shiva SG. Introduction to Logic Design. 1998. CRC Press, Boca 
43. Markus H., Eichler-Jonsson C. Gilles, E.D. and Muller, G. 
42. Starbuck C, Lauffenburger DA. Mathematical model for the 
41. Schoeder, B., Eichler-Jonsson C. Gilles, E.D. and Muller, G. 
40. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 
39. Bakalov Z. Molecular switches at the synapse emerge from 
38. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis 
37. Sanchez L, Thieffry D. A logical analysis of the Drosophila 
33. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, 
32. Wang J, Delabie J, Aachem H, Smeland E, Myklebost O. Clus- 
31. Holdt LS, Canizares-Perez J, Ouellette BM, Wang Q, Cuperlovic-Culf M. 
30. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
28. Shen-Orr S, Milo R, Mangan S, Alon U. Network motifs in the 
27. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
26. Shen-Orr S, Milo R, Mangan S, Alon U. Network motifs in the 
25. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
24. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
23. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
22. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
21. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
20. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
19. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
18. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
17. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
16. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
15. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
14. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
13. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
12. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
11. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
10. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
9. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
8. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
7. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
6. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
5. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
4. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
3. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
2. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N. 
1. Segal E, Pe’er D, Regev A, Sharan R, Koller D, Friedman N.


Further Reading


Attenuation, Control of Gene Expression by

Paul Babitzke, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania

Paul Gollnick, Department of Biological Science, State University at Buffalo, Buffalo, New York
doi: 10.1002/9780470048672.wecb030

Bacteria use a variety of strategies to regulate transcription elongation in response to changes in their environment. Several of these regulatory events are classified as transcription attenuation mechanisms, which involve regulated transcription termination to control expression of downstream genes. In most cases, two alternative RNA secondary structures, an antiterminator and an intrinsic terminator, share nucleotides in common such that their formation is mutually exclusive. Thus, the level of expression of the downstream genes depends on the frequency that the antiterminator forms in the nascent transcript. Transcription attenuation mechanisms in general are divided into two categories. The first category involves situations in which the action of a regulatory molecule promotes transcription termination, with the default situation being transcription readthrough. The second category involves situations in which the action of the regulatory molecule promotes transcription readthrough (antitermination), with the default situation being transcription termination. Transcription attenuation controls the expression of several amino acid biosynthetic and catabolic operons. These mechanisms are often controlled by the efficiency of leader peptide synthesis in Gram-negative bacteria, whereas Gram-positive organisms often use RNA binding proteins or tRNA molecules to determine whether the RNA will fold into the antiterminator or terminator structures. Several examples of transcription attenuation also exist in which an antisense RNA is responsible for controlling plasmid copy number. In addition, small metabolites participate in transcription attenuation mechanisms of a wide variety of bacterial genes by binding directly to the nascent transcript.

The ability of organisms to regulate gene expression in response to environmental signals is vital for their survival. Virtually every step involved in the synthesis, function, and degradation of macromolecules is a potential target for one or more regulatory events (reviewed in References 1–3). Several regulatory events that occur after transcription initiates are categorized as transcription attenuation mechanisms in which the extent of transcription termination is modulated to control the expression of downstream genes. Transcription attenuation allows the organism to modulate the extent of transcription readthrough past the terminator in response to changing environmental signals, thereby regulating expression of the downstream genes. Transcription attenuation mechanisms in general are divided into two categories. In one case, the default situation is transcription readthrough; the action of a regulatory (effector) molecule promotes termination. The default situation for the second category is transcription termination. In this case, the action of an effector molecule promotes transcription readthrough (antitermination) (1, 2). In each case, the genetic information required for transcription attenuation is encoded within a leader region located between the transcription start site and the first structural gene of the operon or between two genes within an operon (reviewed in Reference 4). This article is organized around the class of effector molecule that is used to sense environmental changes. These regulatory factors can be translating ribosomes, proteins, RNA molecules, or metabolites (Fig. 1). Examples are known in which each class of effector molecule promotes transcription termination or antitermination. One example of each of these
The discovery of transcription attenuation was the first demonstration that organisms can exploit an RNA structure to modulate gene expression. The first examples of transcription attenuation were discovered for the Escherichia coli trp (tryptophan biosynthetic) and Salmonella typhimurium his (histidine biosynthetic) operons, although many other amino acid biosynthetic operons in Gram-negative bacteria are regulated in a similar fashion. Together, these discoveries established an important paradigm for regulation of gene expression after transcription initiation. It is now apparent that translating ribosomes can serve as the effector molecule for both mechanistic categories of transcription attenuation.

Transcription Termination Mediated by a Translating Ribosome

TrpR, which is a DNA binding repressor protein, regulates transcription initiation of the E. coli trpEDCBA operon. Under tryptophan-limiting conditions, TrpR represses transcription initiation, whereas repression is relieved in the presence of excess tryptophan. Once transcription initiates the elongating transcription complex is subject to control by transcription attenuation (reviewed in References 5 and 6). The leader transcript can form three RNA secondary structures that are referred to as the pause hairpin, the antiterminator structure, and an intrinsic terminator hairpin. Because the antiterminator shares nucleotides in common with the terminator, their formation is mutually exclusive. The pause hairpin has two additional roles in this transcription attenuation mechanism; it serves as an anti-antiterminator structure that prevents antiterminator formation, and it codes for a leader peptide. A model of the E. coli trp operon transcription attenuation mechanism is presented in Fig. 2a.

Soon after transcription of the trp operon is initiated, the pause hairpin forms in the nascent transcript, which signals RNA to pause. The paused RNA complex allows sufficient time for a ribosome to initiate translation of a short open reading frame that encodes a 14-amino-acid leader peptide. The translating ribosome disrupts the paused RNA complex such that transcription resumes, which thereby couples transcription and translation. At this point two different outcomes can occur depending on the level of tryptophan in the cell. Under tryptophan-limiting conditions, the aminoaacylation of RNA is inefficient such that the level of tryptophanyl-tRNA (also called charged tRNA) is low. As a consequence, the translating ribosome stalls at either of two tandem Trp codons within the leader peptide coding sequence. Ribosome stalling at either Trp codon uncouples transcription from translation. As transcription proceeds, the antiterminator structure forms, which prevents formation of the overlapping terminator hairpin, resulting in transcription readthrough into the trp operon structural genes (Fig. 2a). Under tryptophan excess conditions, the concentration of tryptophanyl-tRNA is sufficient to allow efficient translation of the Trp codons within the leader peptide coding sequence. As a consequence, the ribosome continues translation to the end of the leader peptide unaltered and dissociates from the mRNA. In this situation, the anti-antiterminator forms and prevents formation of the overlapping antiterminator. Because formation of the antiterminator is prevented, the overlapping terminator hairpin can form, which results in termination of transcription before RNA reaches the trp operon structural genes. Thus, expression of the trp operon is decreased when the cell has an adequate supply of tryptophan.

Figure 1 Generalized model for sensing regulatory effectors by nascent mRNA leader transcripts. Transcription attenuation mechanisms have been identified in which the nascent transcript interacts with a translating 70S ribosome, a protein, an RNA molecule or a small metabolite. (a) Binding of the effector molecule promotes transcription termination. (b) Binding of the effector molecule promotes transcription readthrough (antitermination). See text for details.
Figure 2 (a) Transcription attenuation model of the E. coli trpEDCBA operon. RNAP pauses after formation of the pause hairpin (stem-loop 1:2). RNAP pausing provides time for a ribosome to initiate translation of the leader peptide. In tryptophan-limiting conditions, the ribosome stalls at one of the tandem Trp codons, which thereby allows formation of the antiterminator structure (stem-loop 2:3), which results in transcription readthrough. In conditions of tryptophan excess, the ribosome reaches the leader peptide stop codon and dissociates. In this instance, formation of the anti-antiterminator blocks formation of the antiterminator, which thereby allows formation of the terminator hairpin (stem-loop 3:4), which causes transcription to terminate upstream from the trp operon structural genes. See text for details. [Adapted from the Encyclopedia of Molecular Biology (1999) with permission (1).] (b) Antitermination model of E. coli tnaCAB operon. RNAP pauses after formation of the pause hairpin. RNAP pausing provides time for a ribosome to initiate translation of the leader peptide. Under noninducing conditions (free tryptophan is not in excess), dissociation of the 70 S ribosome at the tnaC stop codon exposes a rut site allowing Rho binding. Rho translocates to paused RNAP leading to transcription termination. Under inducing conditions (excess tryptophan), the combination of tryptophan binding and nascent peptide interaction with the ribosome exit channel causes the ribosome to stall at the tnaC stop codon because the nascent peptide cannot be cleaved from the peptidyl-tRNA. The stalled ribosome prevents Rho from binding to the nascent transcript that leads to transcription readthrough. Tryptophan is represented by a star. See text for details. [Adapted from the Encyclopedia of Molecular Biology (1999) with permission (2).]
Transcription Antitermination Mediated by a Translating Ribosome

E. coli and several other microorganisms have the capacity to degrade tryptophan to serve as a carbon and nitrogen source. The degradative tryptophanase operon (tnaCAB) of E. coli is regulated by catabolite repression and by an antitermination mechanism (reviewed in Reference 4). This antitermination mechanism involves translation of a 24-amino-acid leader peptide (TnaC) containing a crucial Trp codon (7, 8). RNAP pause sites(s) between tnaC and tnaA (9), and Rho termination factor (10). A model that illustrates the tna antitermination mechanism is shown in Fig. 2b.

Under non-catabolite-repressing conditions, RNAP initiates transcription and continues until it reaches a pause site downstream of tnaC. Paused RNAP provides time for a ribosome to initiate translation of the leader peptide. Under tryptophan-limiting conditions, the translating ribosome reaches the UGA stop codon and is released, thereby exposing a rut (Rho utilization) site that immediately follows the stop codon. Rho binds to the rut site in the nascent tna transcript and translocates along the RNA in a 5′ to 3′ direction until it encounters paused RNAP, which ultimately leads to transcription termination upstream of tnaA. When cells are growing with inducing levels of tryptophan, the nascent TnaC peptide and a free molecule of tryptophan interact with the translating ribosome, which inhibits ribosomal peptidyl transferase activity such that peptidyl-tRNA remains bound to the ribosome (11–13). The stalled ribosome masks the rut site, which thereby prevents Rho interaction with the transcript. Eventually RNAP resumes transcription such that the structural genes encoding tryptophanase (tnaA) and a tryptophan permease (tnaB) are expressed (Fig. 2b).

The mechanism of ribosome stalling during translation of tnaC has been investigated in considerable detail. The nascent leader peptide (TnaC) interacts with ribosomal protein L22 and 23S rRNA in the narrow region of the ribosome exit channel (14). Ribosome recognition of the TnaC peptide results in specific binding of free tryptophan in the ribosome, which inhibits cleavage and the subsequent release of the nascent leader peptide (15, 16).

Transcription Attenuation Mediated by Proteins

Although the transcription attenuation mechanism of the E. coli trp operon served as a paradigm for several years, it is now apparent that several classes of effector molecules can be responsible for sensing and responding to environmental signals. Numerous examples exist in which RNA binding proteins perform this important task. In some instances, the presence of a specific metabolite leads to activation of the RNA binding protein such that it can bind to its RNA targets. In other examples, the metabolite leads to inactivation of the RNA binding protein. Despite this fundamental difference, several features of the corresponding mechanisms are remarkably similar to those described above. One clear distinction that is worthy of particular mention is that the mechanisms that rely on RNA binding proteins do not require coupling of transcription and translation. As a consequence, this general regulatory strategy has been identified in several eukaryotic organisms in which transcription and translation occur in separate cellular compartments (reviewed in Reference 3).

Transcription Termination Mediated by an RNA Binding Protein

The transcription attenuation mechanism for the trpEDCFBA operon in the Gram-positive organism Bacillus subtilis differs dramatically from that described for E. coli. Most notably, translation of a leader peptide is not involved in the B. subtilis mechanism. Instead, an RNA binding protein called TRAP (trp RNA-binding attenuation protein) is responsible for sensing the level of tryptophan in the cell and for the decision of whether to terminate transcription in the 5′ leader region, or to readthrough into the trp operon structural genes (reviewed in References 17 and 18). Transcription initiation of the trp operon seems to be constitutive, which suggests that the >1000-fold regulation in response to tryptophan occurs after transcription initiates. A transcription attenuation model for the B. subtilis trp operon is depicted in Fig. 3a.

The B. subtilis trp leader transcript contains inverted repeats that can form mutually exclusive antiterminator and intrinsic terminator structures. Soon after transcription initiates, RNAP can pause at the nucleotide just preceding the critical overlap between these two alternative structures, presumably to provide additional time for TRAP binding (19, 20). Under tryptophan excess conditions, tryptophan-activated TRAP binds to 11 closely spaced (G/U)AG repeats in the nascent transcript. As transcription proceeds, the RNA wraps around the protein ring with each triplet repeat interacting with several amino acids on adjacent subunits in the protein (21). Bound TRAP prevents formation of the antiterminator, which allows formation of the overlapping terminator structure. Thus, transcription halts in the leader region before RNAP reaches the trp operon structural genes. An RNA hairpin called the 5′ stem-loop (5′SL) also interacts with TRAP; however, the mechanism of TRAP binding to this structure is distinct from TRAP interaction with the (G/U)AG repeats (22–24). It seems that TRAP-5′SL RNA interaction increases the rate in which TRAP binds to the nascent leader transcript, which thereby increases the efficiency of termination (24). Under tryptophan-limiting conditions, TRAP is not activated and does not bind to the nascent trp operon transcript. As a consequence, formation of the antiterminator prevents formation of the terminator, which results in transcription readthrough into the trp operon structural genes (Fig. 3b).

The crystal structure of B. subtilis TRAP in complex with tryptophan shows that this protein consists of 11 identical subunits arranged in a ring (25). The TRAP oligomer is composed of 11 seven-stranded antiparallel β-sheets, with each β-sheet
Attenuation, Control of Gene Expression by Limiting β-glucosides

**Figure 3** (a) Transcription attenuation model for the *B. subtilis* trpEDCFBA operon. RNAP pauses after formation of the pause hairpin. RNAP pausing may provide additional time for tryptophan-activated TRAP to bind to the nascent transcript. In tryptophan-limiting conditions, TRAP does not bind to the 5′ SL and the (G/U)AG repeats (small black rectangles). Thus, formation of the antiterminator (stem-loop 2:3) results in transcription readthrough. In conditions of tryptophan excess, tryptophan-activated TRAP binds to the 5′ SL and the triplet repeats. Bound TRAP blocks or disrupts formation of the antiterminator structure, which thereby allows formation of the terminator hairpin (stem-loop 2:3), which causes transcription to terminate upstream from the trp operon structural genes. See text for details. (b) Antitermination model for the *E. coli* bglGFB operon. Under noninducing conditions (limiting β-glucosides), the phospho-BglG monomers cannot bind to the nascent transcript. Thus, formation of a terminator hairpin (stem-loop 2:3) causes transcription to terminate. In the presence of β-glucosides, the BglG monomers are dephosphorylated, which allows them to dimerize. BglG dimers bind and stabilize the antiterminator structure (stem-loop 1:2), which prevents formation of the overlapping terminator, and transcription continues into the structural genes. Note that overlapping antiterminator and terminator hairpins are found in the leader region and between the bglG and the bglF coding sequences. See text for details.

Four amino acid residues within each TRAP subunit (E36, K37, K56, and K58) are critical for RNA binding (29). Three of these residues form KKR motifs that are aligned on the perimeter of the *B. subtilis* and the homologous *B. stearothermophilus* TRAP ring. Several crystal structures of *B. stearothermophilus* TRAP–RNA complexes show that the RNA wraps around the TRAP protein ring with the bases pointing toward the protein, such that the majority of the hydrogen bond interactions are with the RNA bases and there are no contacts to phosphates (21, 30). The only direct hydrogen bond to the RNA backbone is to the 2′ OH of the ribose of the third G in each repeat. K37 forms a hydrogen bond with the A of each GAG repeat, whereas K56, R58, and E36 hydrogen bond with the G in the third position of each repeat. The side chain of K37 also forms hydrophobic interactions with the base in the first residue of each repeat. The spacer bases do not make specific contacts with TRAP, which is consistent with the lack of sequence conservation in spacers of natural TRAP binding sites and with findings that indicate that their composition in general is not critical for TRAP binding.
A model of TRAP-trp leader RNA interaction that encompasses several recent findings is as follows. Once a sufficient number of triplet repeats emerge from RNAP (e.g., three or four repeats), TRAP binds to the nascent trp leader transcript by contacting the 5′SL and the emerging triplet repeats. The nature of these contacts orients the protein so the RNA easily can wrap around TRAP’s periphery as transcription continues. TRAP sequencesthe remaining repeats in a 5′ to 3′ directionality as they emerge from RNAP (31, 32). As the tenth and eleventh repeats bind, the 5′SL is displaced as a consequence of the geometry of the TRAP-RNA complex in which TRAP completely is encircled by RNA (24). In conjunction with RNAP pausing, this RNA binding mechanism ensures that TRAP binds in time to block formation of the antiterminator structure.

Transcription Antitermination Mediated by an RNA Binding Protein

RNA binding proteins are responsible for mediating antitermination of several catabolic operons in bacteria. BglG, SacT, SacY, and LicT belong to a family of more than 50 proteins that regulate gene expression by a common antitermination mechanism; antitermination is mediated by the RNA binding proteins SacT and SacY, respectively (33). The structure of the RNA binding domain of SacT has been solved by both NMR (36) and X-ray crystallography (37). The domain exists as a dimer with each monomer consisting of a four-stranded antiparallel β-sheet. The structure of the RNA-binding domain from the LicT antiterminator protein in complex with a 29 base RNA shows that LicT binds mostly through hydrophobic and stacking interactions with the RNA and that binding “clamps” the RNA so as to stabilize the antitermination structure (38).

Transcription Antitermination

Mediated by an RNA Molecule

Antiserine RNA- mediated transcription attenuation controls the copy number of several Gram-positive plasmids (reviewed in References 39 and 40). In the case of the Streptococcus agalactiae plasmid, pIPS01, the antiserine RNA (RNA III) inhibits expression of repR, which is the gene encoding the essential RepR inhibitor protein, by binding to the nascent repR leader transcript (RNA II) (Fig. 4a). This interaction promotes formation of an intrinsic terminator upstream of the repR coding sequence. In the absence of RNA III-RNA II interaction, an alternative structure forms in the nascent RNA II transcript that blocks formation of the terminator. Although this alternative structure is more complex, it is functionally analogous to the antiterminator structures described above.

Interaction of RNA III with RNA II is initiated by the formation of a so-called kissing complex between single-stranded loops of both molecules, followed by propagation of the intermolecular RNA helix (41). The kissing interaction involves base pairing between loops III and IV from RNA III with loops subsequently to β-glucosides. It seems that BglF recruits BglG to the cell membrane in the absence of β-glucosides such that the system can respond rapidly to the presence of the stimulating sugar (35).

Expression of two sucrose utilization operons in B. subtilis, sacPA and sacB, is induced by sucrose by similar transcription antitermination mechanisms; antitermination is mediated by the RNA binding proteins SacT and SacY, respectively (33). The structure of the RNA binding domain of SacY has been solved by both NMR (36) and X-ray crystallography (37). The domain exists as a dimer with each monomer consisting of a four-stranded antiparallel β-sheet. The structure of the RNA-binding domain from the LicT antiterminator protein in complex with a 29 base RNA shows that LicT binds mostly through hydrophobic and stacking interactions with the RNA and that binding “clamps” the RNA so as to stabilize the antitermination structure (38).
Attenuation, Control of Gene Expression by RNAP

Figure 4. (a) Transcription attenuation model of S. agalactiae plasmid pIP501-encoded repR. When the concentration of RNA III is low, the absence of RNA III interaction allows formation of the antiterminator structure (stem-loop 1:2), which blocks formation of the terminator that leads to transcription readthrough. When RNA III is abundant, the kissing loop interaction between loops L III and L IV of RNA III with loops L I and L II within the nascent RNA II (repR) transcript blocks formation of the antiterminator, which promotes formation of the terminator hairpin (stem-loop 2:3) and transcription termination in the repR leader region. The positions of the residues that correspond to L II of RNA II (stem-loop A:B) are included in each structure for clarity. See text for details. [Adapted from the Encyclopedia of Molecular Biology (1999) with permission (1).] (b) General T-box model of tRNA-mediated antitermination. Under excess amino acid conditions, the amino acid on charged tRNA does not allow interaction of the tRNA with the discriminator sequence in the nascent transcript. As a consequence, formation of the terminator (stem-loop 2:3) halts transcription in the leader region. Under limiting amino acid conditions, uncharged tRNA binds to the nascent transcript via anticodon-specifying sequence base pairing as well as by base pairing between the 3' end of tRNA and the discriminator region. This second interaction stabilizes the antiterminator structure (stem-loop 1:2), which results in transcription readthrough. The tRNA is shown as a cloverleaf structure, and a boxed A.A. attached to the tRNA indicates it is aminoacylated. See text for details. (Adapted from Reference 45 with permission.)

II and I of RNA II, respectively (42). Moreover, loop L1 of RNA II contains a YUNR (Y = C or U; N = A, C, G or U; R = A or G) “U-turn” motif that is important for the kissing interaction; mutations that disrupt the U-turn motif result in higher plasmid copy number (43). Although RNA III and RNA II share extensive sequence complementarity, complete pairing is not necessary to promote transcription termination of RNA II. The finding that inhibition of transcription readthrough is considerably faster than stable complex formation suggests that inhibition precedes stable pairing (41). Furthermore, complete pairing only occurs at a low frequency because pairing of complementary folded RNAs often arrests at the stage of stable binding intermediates (41).

Transcription Antitermination Mediated by an RNA Molecule

In several bacterial species, uncharged tRNA serves as the effector molecule in controlling expression of several aminoacyl-tRNA synthetase genes and a few amino acid biosynthetic operons by a common mechanism termed T-box antitermination.
Most of these genes are from Gram-positive bacteria, although a few examples have been identified in Gram-negative organisms as well (reviewed in References 44 and 45). Expression of these genes is induced specifically by starvation for the cognate amino acid. A generalized model of the T-box antitermination mechanism is shown in Fig. 4b.

The untranslated leader region of each operon contains several conserved structural elements preceding an intrinsic transcription terminator. When the amino acid pool in the cell is sufficiently high, transcription terminates prematurely in the leader region upstream from the coding sequences. In addition to the conserved secondary structures, a critical 21-nt sequence called the T box is present in each leader region. A more structure of the leader region, which forms when the T-box base pairs with the 5' terminus of the terminator stem, functions as an antiterminator, which allows transcription to read through to the structural genes (46).

A critical feature is that the region of T-box control is a trinucleotide sequence that corresponds to a codon for the appropriate amino acid involved in regulating the structural genes. For example, in tyrS, which encodes tyrosyl-tRNA synthetase, the leader contains a UAC tyrosine codon (46). This coden-like sequence is always present within an internal loop of a conserved RNA structure. This triplet is designated the "specifier sequence" because it specifically interacts with the antiterminator, which allows transcription to read through to the structural genes (46).

A second base pairing interaction between the acceptor end of uncharged tRNA and the complementary sequence in the T box leads to stabilization of the antiterminator structure. The antiterminator side bulges contain a UGGN sequence, where N corresponds to a variable T-box position that covaries with the residue preceding the CCA acceptor end of the cognate tRNA (47, 48). The presence of the amino acid on charged tRNA prevents this second interaction from taking place (49). Although tRNA interaction with the specifier and UGGN sequences is firmly established, substantial genetic data suggest that additional RNA–mRNA contacts take place (44).

Reconstitution of the B. subtilis gleQ5 T-box mechanism in vitro with purified components suggests that additional protein factors are not required for RNA-mediated antitermination (50). Furthermore, the finding that charged tRNA can compete with uncharged tRNA for mRNAs binding indicates that the ratio of uncharged:charged tRNA is monitored by the nascent transcript, rather than simply the concentration of uncharged tRNA (49).

Although codon–anticodon interaction takes place at least 100 nt upstream from the terminator, nascent transcripts that extend just upstream of the terminator are still competent for RNA binding and antitermination (49). Thus, it is possible that RNA-P pausing participates in this regulatory mechanism by providing additional time for RNA binding, although a regulatory pausing event has not been firmly established (51).

Under starvation conditions for the corresponding amino acid, the ratio of uncharged-to-charged tRNA is relatively high. In this case, efficient interaction of uncharged tRNA with both the specifier and the discriminator sequence elements of the nascent leader transcript promotes formation of the antiterminator structure, which allows transcription to proceed into the coding region. The resulting increase in the level of RNA synthetase presumably allows more efficient charging of the scarce amino acid. As the concentration of the amino acid increases, an increasing proportion of nascent transcripts will interact with charged tRNA. In this case, the terminator hairpin will cause RNA-P to terminate transcription in the leader region.

Transcription Attenuation Mediated by Metabolites

The most recently identified class of transcription attenuation mechanism involves direct sensing of the effector molecule by the nascent transcript (52–54). These RNA sensors control metabolically diverse pathways. As for the other attenuation and antitermination mechanisms discussed thus far, recognition of the particular effector molecule occurs with the appropriate affinity and high specificity required for precise control of gene expression.

Transcription Termination Mediated by a Small Metabolite

In B. subtilis, 26 genes grouped into 11 operons constitute the S-box regulon. These genes are involved in sulfur metabolism as well as in the biosynthesis of methionine, cysteine, and S-adenosylmethionine (SAM) (55–57). SAM, the effector of the S-box regulon, binds to the nascent transcript with exquisite specificity; the closely related SAM analog S-adenosylhomocysteine binds with much lower affinity to the RNA targets and does not promote transcription termination in vitro (54, 58, 59). A simplified model of SAM-mediated attenuation is presented in Fig. 5a. Under limiting SAM concentrations, the nascent S-box transcript adopts an antiterminator structure that allows transcription to proceed into the structural gene(s). However, under conditions of SAM excess, SAM binds to the nascent transcript and stabilizes the antiterminator structure, which blocks formation of the antiterminator. As a consequence, the terminator hairpin forms and transcription halted in the leader region.

High-resolution structural information has been obtained for several RNA sensors (56, 60–63). A common feature of the leader RNAs is that they form tertiary structures in which single-stranded loops base pair with another (pseudoknot). For example, the tertiary structure of the SAM-sensing RNA contains a pseudoknot that occurs between the loop of an RNA hairpin and the single-stranded junction between two other helices (64). This pseudoknot seems to stabilize the global architecture of the structure and probably facilitates SAM recognition (56). A more common characteristic of these structures is engulfment of the ligand by the binding pocket of the folded RNA (63).
Transcription Antitermination Mediated by a Small Metabolite

Adenine sensing by the nascent *B. subtilis* pbuE (*ydhL*) transcript constitutes the most recently described antitermination mechanism (63–65). Although this is the only reported example of this type of antitermination mechanism, it is likely that additional examples will be identified. A model that depicts this antitermination mechanism is shown in Fig. 5a. Under limiting adenine conditions, an intrinsic terminator hairpin forms in the nascent pbuE transcript, which causes RNAP to terminate transcription in the leader region. In the presence of excess adenine, adenine binds to the nascent transcript and stabilizes formation of an alternative antiterminator structure, which thereby allows transcription to continue into the pbuE structural gene (65). Adenine-mediated activation of pbuE expression makes biologic sense as it encodes an apparent purine efflux pump (66). Deletions of the adenine-binding domain of the leader RNA caused constitutive expression of pbuE and conferred resistance to 2-fluoroadenine, which suggests that PbUE can pump this adenine analog out of the cell (65, 66). Analog binding studies demonstrated that the binding affinity of adenine and 2-aminopurine are similar, whereas the affinity of 2,6-diaminopurine is considerably higher. However, the association rate was faster for adenine compared with the two purine analogs. This finding supports a model in which regulation depends on the kinetics of ligand binding and the rate of transcription, rather than simple binding affinity (67).

The structure of the pbuE leader RNA is similar to the guanine-dependent RNA sensor in the xpt leader transcript (65). It was shown that a single C to U substitution in the loop of a triple helical junction swapped the xpt aptamer specificity from guanine to adenine. Importantly, the pbuE leader contains a U in the identical position. These results led to the hypothesis that this U residue in the pbuE transcript paired with adenine, whereas the C residue in the xpt transcript paired with guanine. This hypothesis was verified by NMR structural studies. Moreover, it was determined that adenine binding to pbuE leader RNA involved a base triple with two U residues, which includes the previously proposed uridine (68).

Concluding Remarks

The discovery of transcription attenuation over 30 years ago led to the realization that mRNAs serve a purpose beyond simply functioning as a conduit of information from DNA to protein. Indeed, the discovery of transcription attenuation established for the first time that RNA structure can modulate gene expression. It is now abundantly clear that expression of many genes is controlled by several different mechanisms after transcription.
initiates. Furthermore, it is apparent that transcription attenuation occurs by a variety of mechanisms that differ in the nature of the effector molecule (translating ribosome, RNA binding protein, RNA molecule, small metabolite), whether the effector promotes transcription termination (attenuation) or transcription readthrough (antitermination), as well as the structure of the RNA target in the nascent transcript. Although each mechanism described in this review contains key differences, it is also apparent that all but one of these mechanisms share an important feature: the presence of mutually exclusive terminator and terminator structures. One point of view is that these RNA-based regulatory mechanisms are relics of an RNA world in which both the storage of genetic information and the metabolic function were carried out by RNA molecules. Conversely, it is conceivable that these regulatory mechanisms have evolved more recently. In either case, it is apparent that what was once viewed as a biologic quirk is widespread and has been adapted to suit the physiologic needs of perhaps every organism.

Acknowledgments

The authors thank the members of their laboratories, both past and present, for stimulating discussion and discovery. This work was supported by grants GM 052840 to P. B. and GM 062750 to P.G. from the National Institutes of Health.

References

| Attenuation: Control of Gene Expression by Bacillus subtilis | }


Further Reading


This excellent review article summarizes the mechanisms responsible for processive antitermination in lambdoid bacteriophages, which includes the classic N-mediated antitermination of lambda and factor-independent antitermination of the related HK022 phage. This work was not included in this article because of space considerations.

See Also

- mRNA Untranslated Regions (UTRs)
- Protein-Nucleic Acid Interactions
- Small Molecule-Nucleic Acid Interactions
- Transcription, Activators and Repressors of Transcription, Initiation of
Epigenetic Modifications

Dori Huertas, Cancer Epigenetics and Biology Program, Catalan Institute of Oncology (ICO), Barcelona, Catalonia, Spain
Manel Esteller, Cancer Epigenetics Laboratory, Spanish National Cancer Centre (CNIO), Madrid, Spain
doi: 10.1002/9780470048672.wecb161

Epigenetic modifications play a fundamental role in chromatin structure and function. Histone modifications and DNA methylation regulate important biologic processes such as heterochromatin formation, genomic imprinting, X-chromosome inactivation, and transcriptional activation and silencing. Epigenetic mechanisms are responsible for a considerable part of the phenotype of complex organisms. The understanding of epigenetic modifications in chromatin may offer some clues to solve the mechanisms of cellular identity, tumorigenesis, stem cell plasticity, regeneration, and aging. These processes are of interest in the fields of human biology and human diseases. Here, we describe the principal components involved in epigenetic modifications, such as the enzymes that modify chromatin, the protein domains that recognize modified chromatin, some best-characterized downstream effectors, and the tools and techniques for studying epigenetics. However, the biological mechanism of the majority of components is still poorly understood, and the analysis of these components is fundamental to our understanding of epigenetics.

The identity of cells within an organism is determined by its heritable information, which is encoded by genetic and epigenetic information. Genetic information is the ordered sequence of nucleotides present in the genome of all cells of a given organism. Epigenetic information is based on epigenetic modifications, which are chromatin covalent modifications responsible for maintaining stable states of gene expression through mitotic divisions without alterations in DNA sequence. Although genetic information is homogenous throughout all cells within an organism, epigenetic modifications vary according to developmental programs that define the transcriptome of each cell type and ultimately determine the identity of cells (1). The epigenetic information of cells is stored as covalent modifications of DNA and histones. The most common epigenetic modification of DNA is the methylation of the 5' position of cytosines in the context of CpG dinucleotide (2). In addition, epigenetic modifications of histones include a variety of complex post-translational modifications, of which lysine acetylation and methylation and serine phosphorylation are the best characterized.

In the last decade, remarkable progress has been made to characterize epigenetic modifications and to associate specific patterns of modification with chromatin functionality. In this review, we will discuss the nature of DNA and histone epigenetic modifications. We will describe the modifying enzymes responsible for creating or eliminating each modification and the proteins that read such epigenetic modifications and recruit functional protein complexes. These are the downstream effectors that regulate chromatin structure and DNA accessibility.

The functionality of some chromatin modification patterns is well established; however, the vast majority remains poorly understood. Given that histones are subject to more than 100 post-translational modifications, considerable efforts to characterize the mammalian epigenome are ongoing.

Biological Background

Studies of a wide range of systems have contributed to our understanding of epigenetic processes in various model organisms. The majority of epigenetic processes are associated with gene silencing, for example, heterochromatin formation (3), genomic imprinting (4), X-chromosome inactivation (5), and Polycomb repression (6). In addition, the activation of transcription by the Trithorax group of genes is also an epigenetic mechanism (7).

Constitutive heterochromatin

The DNA within the cell exists as a nucleoprotein complex termed chromatin. Chromatin allows the packaging of DNA within the nucleus while at the same time is flexible...
epigenetic machinery (5). Gene silencing of the whole chromosome by recruitment of the Xist coats the entire inactive X chromosome, which leads to the X chromosome that will become silenced. This RNA transcript is likely that only a subset will have epigenetic inheritance. Whether the entire repertoire of histone modifications is heritable remains to be established. In fact, whether they are “epigenetic” in this sense or not is a matter of semantics, and the term may need to be redefined. Indeed Allis, Jenuwein, and Reinberg have proposed (10, 11) that epigenetics be defined as “the sum of the alterations to the chromatin template that collectively establish and propagate different patterns of gene expression (transcription) and silencing from the same genome.”

This definition includes transient chromatin modifications, which are associated with DNA repair or cell-cycle stages, and stable chromatin modifications that are maintained across multiple cell generations.

DNA methylation and the Polycomb/Trithorax system are the paradigms of epigenetic heritable systems. Cavalli and Paro (12) have shown that an activated state of the Fab-7 element, incorporated in transgenic flies, is inheritable mitotically through development and can be transmitted to the subsequent generations through female meiosis. This finding means that the Polycomb/Trithorax system can memorize gene expression patterns that have been set up by other cellular mechanisms.

The evidence that DNA methylation is heritable is self-sustaining. Maintenance DNA methylases recognize hemi-methylated DNA, the product of replication of the fully methylated DNA, and add methyl groups to the unmethylated DNA strand.

The study of epigenetic modifications is of great interest. Perturbations of chromatin structure can cause inappropriate gene expression and genomic instability that results in cellular transformation and malignant outgrowth. Proteins that control chromatin organization therefore constitute key players in cancer pathogenesis.

**Histone Modifications**

The DNA within the cell exists in the form of chromatin. The basic repeating unit of chromatin is the nucleosome, a structure consisting of 147 bp of DNA wrapped around an octamer of the core histones H2A, H2B, H3, and H4. The histone proteins are composed of a globular domain and unstructured and protruding N- and C-terminal tails. A striking feature of the tail histones is that they are subjected to several posttranslational modifications, which are associated with DNA repair or cell-cycle stages, and stable chromatin modifications that are maintained across multiple cell generations.

Histone modifications affect the organization of the chromatin to modulate its involvement in important nuclear functions such as DNA transcription, repair, and replication. The possibility that such combinations could modify chromatin behavior led to the hypothesis that they might constitute a molecular “histone code” (13) that is read by other proteins to mediate unique cellular responses.

A cetylation of the lysine residues of histones H3 and H4 was one of the first modifications of chromatin to be described (Fig. 1) (14). It is associated with the establishment of an open chromatin state that is active transcriptionally. This modification contrasts with hypoacetylation, which is associated with a compacted chromatin structure that is inaccessible to transcription machinery. The majority of all lysine residues of the N-terminal tails of histones H3, H4, H2A, and H2B have the ability to be...
## Table 1 A catalog of posttranslational chemical modifications of the histone proteins

<table>
<thead>
<tr>
<th>Histone</th>
<th>Residue</th>
<th>Modification</th>
<th>Biologic function</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2A</td>
<td>S1, K5, K9, K13, K119</td>
<td>Phosphorylation</td>
<td>Chromosome condensation</td>
</tr>
<tr>
<td>H2B</td>
<td>K5, K12, K15, K20, K123</td>
<td>Acetylation</td>
<td>Transcriptional activation</td>
</tr>
<tr>
<td>H3</td>
<td>T3, T11, S10, S28, R12, R17, R26, K4, K9, K12, K16, K27, K36, K79</td>
<td>Phosphorylation, Acetylation, Metylation</td>
<td>Mitosis, transcriptional activation, transcriptional repression</td>
</tr>
<tr>
<td>H4</td>
<td>S1, K5, K8, K12, K16, R3, K20</td>
<td>Phosphorylation, Metylation</td>
<td>DNA repair, transcriptional activation, transcriptional repression</td>
</tr>
</tbody>
</table>

![Histone modifications diagram](image_url)
Thus, SirT2 is a major contributor to the inheritance of epigenetic marks through mitosis (21). The methylation of histones can occur in various lysines of histone H3, such as K4, K9, K27, K36, and K79, and at lysine 20 of histone H4. Lysines can be mono-, di-, or trimethylated at all these sites. This modification could be associated with either the repression or activation of DNA transcription, depending on the site and methylation status (mono-, di-, or tri-) of the histone lysine residues. Chromatin immunoprecipitation (ChIP) experiments have shown that three methylation sites on histones are involved in transcriptional activation: methylation at lysine 4 (H3K4), at lysine 36 (H3K36), and at lysine 79 (H3K79) (22, 23). Three lysine methylation sites are connected with transcriptional repression: methylation of lysine 9 (H3K9) and lysine 27 (H3K27) of histone H3 and methylation of lysine 20 of histone H4 (H4K20). These modifications are epigenetic modifications of a repressed chromatin state (24).

Histone arginine methylation can be either mono- or dimethylated at residues R2, R8, R17, and R26 of histone H3 and at R3 of histone H4. It plays a role in transcriptional activation and repression. The biological mechanism of each histone arginine modification is mainly unknown; however some data implicates H4R3, H3R17, and H3R26 modifications with transcriptional activation. In contrast, H3R8 and H3R2 are associated with repression (25, 26). Furthermore, two different groups have recently provided the first mechanistic insight into the function of arginine methylation on chromatin (27, 28). They have shown that H3R2 is enriched throughout all silenced loci in budding yeast and that in all cases the pattern of H3R2me2 is mutually exclusive with the trimethylation of H3K4 (27), which marks active genes. The role of H3R2me2 in controlling H3K4me3 also is conserved in humans (28).

The phosphorylation of histones can occur at serine and threonine residues. The function of histone phosphorylation is related largely to either chromosome condensation and/or DNA repair of double-stranded breaks (DSBs). Phosphorylation at serine 10 of histone 3 (H3S10) plays a key role in chromosome condensation during mitosis (29). Phosphorylation at threonine 3 of histone 3 (H3T3) is required for normal metaphase chromosome alignment (30). Other lesser-known residues, such as phosphorylation at serine 1 of histone 4 (H4S1) and at serine 14 of histone H2B (H2BS14), also regulate chromosome condensation in budding yeast (31). Serine 139 of the histone variant H2A.X is phosphorylated in response to DNA DSBs and seems to be an early step in the response to DNA damage (32). Phosphorylation at other sites, such as serine 1 of histone H4 (H4S1), serine 129 of histone H2A (H2AS129), and serine 14 of histone H2B (H2BS14), also contribute to sense DSBs during DNA repair (33). In addition, the phosphorylation of histone H3 is involved in transcriptional activation through H3S10 or H3S28 (34).

**Histone-modifying enzymes**

The modification of histones is a dynamic process with contributing enzymes that can either direct or remove the modification. In the last 10 years, enzymes that play a role in both directions have been identified (Table 2).

**Epigenetic Modifications**

Histone acetyltransferases (HATs) form an extended family of enzymes that can be grouped into three families on the basis of sequence homology. The first of these families is the GNAT family (Gcn5-related N-acetyltransferase), which includes HAT1, the yeast Gcn5, and its human ortholog, PCAF. The second family is p300/CBP, which includes the two human paralogs p300 and CBP. The third family is the MYST family, which includes MOZ, Ybf2/Sas3, Sas2, and Tip60. Although HATs in general acetylate more than one lysine, some specificity has been detected. Structural comparison between the three families has shown that HATs contain a conserved core for Ac-CoA binding and that different domains fold around this core to facilitate substrate specificity (35). However, more analyses are necessary to elucidate the possible specificity of HATs for lysine residues.

Histone deacetylases (HDACs) are divided into four families on the basis of phylogenetic analysis. The class I family includes human HDAC1, HDAC3, and HDAC8, which are homologous to the yeast Rpd3. The class II family is homologous to the yeast Hda1 and is subdivided into the Iia (HDAC4, HDAC5, HDAC7, and HDAC9) and Iib (HDAC6 and HDAC10) subfamilies on the basis of their sequence homology. The class III family includes the NAD⁺-dependent enzymes of the Sir (or sirtein) family. Its components are yeast Hist proteins 1-4 and human Sirtuins 1-7. The class IV family contains the human HDAC 11. In general, HDACs are not specific for a particular acetyl group, but the sirtein family seems to prefer deacetylating the lysine 16 of histone H4 (36). The catalytic mechanism of the Sirtein family differs from those of other families in its use of NAD⁺ as a cofactor.

One important feature of HAT and HDAC enzymes is that they are integrated frequently within multiprotein complexes in which the noncatalytic subunits tend to regulate the substrate specificity that contributes to the molecular functionality in a cellular context.

Histone methyltransferases (HMTs) of lysines contain a catalytic SET domain of about 130 amino acids, with the exception of those of the Dot family. SET owes its name to the shared domain of the first three HMTs to be identified in Drosophila: Suppressor of position-effect variegation Su(var)3-9, Enhancer of Zeste E(z)2, and Trithorax (Trx). The HMTs can be classified into several families on the basis of the sequence homology within the SET domain and adjacent sequences (24, 37). The SET1 family includes the MLL family of proteins, which are homologous to the TRX proteins of Drosophila, and the Set1 protein of yeast. They specifically methylate lysine 4 of histone H3 linked to the activation of transcription. The SMYD family proteins also methylate H3K4 but include an MYND domain that binds to specific DNA sequences. The SUV39 family includes the SUV39H1 and SUV39H2 histone methyltransferases, SETDB1, and G9A, which specifically methylate lysine 9 of histone H3. The PRDM family contains a PR SET domain that is considered a subclass of the SET domain. The catalytic activity of the PRDM family is controversial. However, it includes crucial proteins such as RIZ1 and BLIMP1 that are associated with the methylation of lysine 9 of histone H3, which is important in regulating the gene transcription of many biological processes. The EZH family includes the two homologs of Enhancer of
### Table 2: Representative capture strategies for preparing microarrays

<table>
<thead>
<tr>
<th>Family</th>
<th>Protein</th>
<th>Residue specificity</th>
<th>Molecular function</th>
</tr>
</thead>
<tbody>
<tr>
<td>HATs</td>
<td>GNAT</td>
<td>HAT1, PCAF, Gcn5</td>
<td>K-acetylation</td>
</tr>
<tr>
<td></td>
<td>p300/CBP</td>
<td>p300, CBP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MYST</td>
<td>Tip60, MOF, HBO1, MORF</td>
<td></td>
</tr>
<tr>
<td>HDACs</td>
<td>Clas I</td>
<td>HADC1, HADC3, HADC8</td>
<td>K-acetylation</td>
</tr>
<tr>
<td></td>
<td>Clas Ia</td>
<td>HDAC1, HDAC3, HADC8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clas Ib</td>
<td>HADC4, HADC5, HADC7, HADC9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clas II</td>
<td>Sirtuins 1-7</td>
<td>H4K 16 deacetylation preference</td>
</tr>
<tr>
<td></td>
<td>Clas IV</td>
<td>HDAC 11</td>
<td>K-acetylation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMTs</td>
<td>SET</td>
<td>MLL proteins, SET proteins</td>
<td>H3K4me</td>
</tr>
<tr>
<td></td>
<td>SMYD</td>
<td>SMYD1, SMYD2, SMYD3, SMYD4, SMYD5, SMYD6, SMYD7</td>
<td>H3K4me</td>
</tr>
<tr>
<td></td>
<td>SUN 39</td>
<td>SUN 39H, SETDB12, SETDB14, SETDB15</td>
<td>H3K9me</td>
</tr>
<tr>
<td></td>
<td>EZH</td>
<td>EZH1, EZH2</td>
<td>H3K27me</td>
</tr>
<tr>
<td></td>
<td>SET2</td>
<td>NSD proteins</td>
<td>H3K20me</td>
</tr>
<tr>
<td></td>
<td>ASH1</td>
<td>ASH1</td>
<td>H3K4me, H3K9me</td>
</tr>
<tr>
<td></td>
<td>Other SET proteins</td>
<td>SUN-39H/H2, PR-SET7</td>
<td>H3K20me</td>
</tr>
<tr>
<td></td>
<td>DOT</td>
<td>DOT1L</td>
<td>H3K9</td>
</tr>
</tbody>
</table>

(Overleaf)
<table>
<thead>
<tr>
<th>Family</th>
<th>Protein</th>
<th>Residue specificity</th>
<th>Molecular function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jumonji C (JmC)</td>
<td>PHF</td>
<td>PHF2, PHF8</td>
<td>nd</td>
</tr>
<tr>
<td>JARID</td>
<td>JARID1A, JARID1B, JARID1C, JARID1D, JARID2</td>
<td>H3K4me3/2</td>
<td>X-linked mental retardation</td>
</tr>
<tr>
<td>JARID</td>
<td>JARID3</td>
<td>UTX, UTY, JMJD3</td>
<td>H3K27me3/2</td>
</tr>
<tr>
<td>HDMTs</td>
<td>LSD1</td>
<td>LSD1</td>
<td>H3K4me2/1, H3K9me2/1</td>
</tr>
<tr>
<td>type I PRMTs</td>
<td>PRMT1</td>
<td>H4R3</td>
<td>Transcriptional activation</td>
</tr>
<tr>
<td>type II PRMTs</td>
<td>PRMT5</td>
<td>H3R8, H3R26</td>
<td>Transcriptional repression</td>
</tr>
<tr>
<td>type II PRMTs</td>
<td>PRMT6</td>
<td>H3R2</td>
<td>Transcriptional repression</td>
</tr>
<tr>
<td>type III PRMTs</td>
<td>PRMT7</td>
<td>H3R2</td>
<td>Imprinting in male</td>
</tr>
<tr>
<td>HADs</td>
<td>JMJD6</td>
<td>H3R2, H4R3</td>
<td>Development</td>
</tr>
<tr>
<td>DNMTs</td>
<td>DNMTS</td>
<td>cytosine within CpG</td>
<td>Transcriptional repression</td>
</tr>
<tr>
<td>DNMT1a</td>
<td>DNMT3a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNMT3b</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Histone deacetylases (HDACs), histone acetyltransferases (HMTs), histone demethylases (HDMTs), protein arginine methyltransferases (PRMTs), histone arginine demethylases (HADs), and DNA methyltransferases (DNMTs).
Histone demethylases can remove all three histone lysine methylation states (mono-, di-, and trimethylation). The family is divided into several subfamilies, including JHDM1, JHDM2, JHDM3, JHDM4, JHDM5, JHDM6, and JHDM7. These demethylases have been associated with important roles in chromatin structure and development. For example, JHDM3 and JHDM4 demethylate H3K9 and H3K36, which antagonizes HP1 recruitment to chromatin and regulates gene expression. Moreover, very recent studies have confirmed the important role of histone demethylases in histone modification dynamics. It has been shown that four proteins of the JARID family possess the demethylase activity specific to H3K27 and contribute to cell-fate determination (39). Two more studies have established that members of the UTX/UTY family possess demethylase activity specific to H3K27 and are involved in HOX gene regulation and development (40, 41).

Histone arginine methylation is achieved by protein arginine methyltransferases (PRMTs), which are evolutionarily conserved from yeasts to humans. PRMTs are classified as Type I and II based on the nature of the methylation introduced. Type I catalyzes symmetric demethylation of arginine, and type II catalyzes the asymmetric one. Eleven PRMTs have been identified in humans based on sequence homology; however, their substrate specificity and their biologic functions still remain unknown. PRMTs catalyze many cellular proteins, and some of them also can methylate histones, such as PRMT1, PRMT4, PRMT5, and PRMT7 (26). PRMT1 methylates H4R3 contributing to transcriptional activation. PRMT4 or CARM1 can methylate H3R327 and H3R26, which are marks linked to gene activation. PRMT4 also can methylate H3R32 in vitro with low efficiency. PRMT5 methylates H3R38 and H4R3, and it has been copurified with chromatin remodeling complexes such as HSW1/SNF and NURD. It has been implicated also in transcriptional repression of cell cycle regulator and tumor suppressor genes. PRMT7 mediates H4R3 methylation at the imprinting loci in embryonic stem cells. Very recently, it has been shown that PRMT6 catalyzes asymmetric H3R2 dimethylation in human cells thereby preventing the recruitment of MLL histone acetyltransferase complexes and H3K4 methylation. Reciprocally, H3K4me3 prevents H3R2 methylation by PRMT6 (26).

Of special interest is the finding of the first histone arginine demethylase, the Jumonji domain-containing 9 protein (JMJD9). It shares homology with the jmjC domains found in the histone lysine demethylases. JMJD6 can demethylate H3R2 and H4R3 in biochemical and cell-based assays (42). The recent data that are emerging place histone arginine methylation as a key player of cell growth and cell proliferation processes.

Reading the histone modifications

Histone modifications generate two types of downstream effects on chromatin organization. The first type is produced mainly by the acetylation of lysine residues, which neutralize positive charges of lysine and allow the unfolding of the chromatin through an electrostatic mechanism. The DNA then is more accessible to the transcription machinery. The second effect arises from the generation of docking sites that are read by specific proteins with the purpose of tethering enzymatic activities to the chromatin.

The acetylation of lysines is recognized by the bromodomains, which are small domains included in an extensive family of proteins. The first bromodomain was identified in the Drosophila Brahma protein. Bromodomains were later found widely distributed among different enzymes that acetylate, methylate, or remodel chromatin (43). The bromodomain is present in the members of different families of histone acetyltransferases, such as the Gcn5/PCAF and p300/CBP families. They also are part of some histone methyltransferase enzymes such as Ash1, RIZ, and MLL. Bromodomains are important components of the remodeling enzymes that use ATP to remodel chromatin structure, such as BRM or members of the SWI/SNF complexes. Sequence-specific lysine acetylation recognition by a bromodomain is dependent on how the bromodomain recognizes residues flanking the acetylated lysine, but currently the exact specificity is poorly understood.

Different protein domains recognize the methylation of lysine: the chromodomains, the Tudor domain, the WD-40 repeat, and the PDZ domain. The chromodomain of HP1 within heterochromatin, HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9 and deacetylases. Methylation of lysine 27 of histone H3 is recognized specifically by the chromodomain of HP1 within heterochromatin. HP1 can recruit other heterochromatins as the histone methyltransferases Suvar3-9
chromatin unfolding. Secondly, histone modification can create a docking site that can be read by a protein domain. This protein may act as an independent effector or alternatively tether other complexes with different enzymatic activities that collectively modify the chromatin environment in response to histone modification. In this section, we will discuss some best-characterized biologic processes that involve specific histone modifications. Histone modification is of paramount importance to the establishment of constitutive heterochromatin at centromeres.

**Figure 2**  Constitutive heterochromatin at centromeres. a) Mammalian centromere model. SUV39H1/2 produces H3K9me3 residues, which are recognized by the chromodomain protein HP1α and HP1β. Both proteins interact with other HMTs, such as SUV420H1/2, that methylate adjacent histones that lead to the spreading of heterochromatin. Mammalian heterochromatin is enriched in H3K9me3 and H4K20me3 residues.
b) S. pombe centromere model. The assembly of centromeric heterochromatin in S. pombe is very well established. The components of the RNAi machinery play an essential role. Dicer produces siRNA from the centromeric dsRNA transcripts. siRNA are loaded within the RITS (RNA-induced transcriptional gene silencing) complex to target heterochromatin by recruiting Clr4, which methylates lysine 9 of histone H3. H3K9me2 will be recognized by Swi6 and Chp1. Chp1 is part of the RITS complex that also includes Ago1 and with the RDRC complex (which contains Rdp1, an RNA-dependent RNA polymerase) and dicer process nascent transcripts into siRNA. Clr4 will be recruited in a positive feedback loop that propagates the heterochromatin. Methylation of lysine 9 is accomplished after recruiting histone deacetylases (HDAC) complexes to heterochromatin to deacetylate lysine 9 residues.
Epigenetic Modifications

(Fig. 2) The histone methyltransferase SUV39H1/2 trimethylates lysine 9 of histone H3, which generates targets for the chromodomain-containing proteins HP1α and HP1β. HP1 molecules recruit other histone methyltransferases, like SUV39H1/2 and SUV420H1/2, that di- and trimethylate histones of adjacent sequences spreading the heterochromatin (44) (Fig. 2a). The initial targeting of SUV39H1/2 to heterochromatin involves components of the RNAi silencing machinery, as has been shown in the centromeres (Fig. 2b) of Schizosaccharomyces pombe. Centromeric transcripts can generate siRNA that are loaded within the RITS (RNA-induced initiation of transcriptional gene silencing) complex to target heterochromatin, probably by base-pairing homology with centromeric transcripts. The RITS recruit Clr4, the S. pombe homolog of Clr4, that will methylate lysine 9 of histone H3, acting as recognition sites of the S. pombe homolog of HP1, Swi6, and favoring heterochromatin spreading (45).

DNA methylation also is present within tandem-repetitive DNA sequences in the heterochromatin. They also are thought to inhibit recombination between homologous repeats, which could lead to genomic instability (46). In fact, mutations of the DNMT3b cause centromeric instability. In addition, HDACs lie at the heart of heterochromatin pathways and can be recruited by transcriptional repressor or by methyl–DNA binding proteins.

The epigenetic gene silencing by Polycomb protein complexes is one of the best examples of how epigenetics modulate gene transcription across generations. When the Polycomb repressive complex 2 (PRC2) is recruited to the Polycomb target genes, its histone methyltransferase component, EZH2, trimethylates lysine 27 residues of histone H3 (Fig. 3). The Polycomb repressive complex 1 (PRC1) can recognize trimethylated H3K27 through the chromodomains of the Polycomb protein, a component of the PRC1. This interaction might bring adjacent nucleosomes into the proximity of the PRC2 complex to facilitate widespread methylation over extended chromosomal regions. PRC1 mediates recruitment of DNA methyltransferases, chromatin compaction, and ubiquitylation of the lysine 119 of histone H2A, which is thought to contribute to transcriptional inhibition, but the precise mechanism of gene silencing is unknown (46). Very important contributions to this biologic process have been made recently. The proteins UTX and JMJD3 can demethylate H3K27me3 within the HOX genes, which leads to transcriptional activation. Moreover, UTX can associate with the H3K4me3 histone methyltransferase MLL2. This association is consistent with a model in which the coordinated removal of repressive marks, the polycomb group displacement, and the deposition of activating marks are important for the stringent regulation of the transcription during cellular differentiation (40, 41).

Figure 3  Epigenetic gene regulation of HOX genes by Polycomb complexes and H3K27 demethylases. The HMT EZH2 is a component of the PRC2 complex, and its binding to the HOX genes leads to H3K27me3, which is recognized by the PRC1 complex through its Polycomb member. The polycomb binding produces chromatin compaction and transcriptional repression. The mechanism of transcriptional repression is unknown but might be achieved either by the ubiquitylation of the H2AK119 by the ubiquitin E3 ligase activity present in the PRC1 complex or by the recruitment of DNA methyltransferases by the PRC2 complex. Recently, two H3K27 demethylases associated with MLL2 have been identified. Moreover, UTX associates with MLL2, an H3K4me3 HMT. These results suggest a model in which H3K27 demethylases associated with MLL2 can reverse the transcriptional silencing mediated by Polycomb. The complex would remove H3K27me3 residues and create H3K4me3 marks that lead to the activation of transcription. In this model, the hypothetical association of the PRC2 complex with H3K4me3 demethylases facilitates transcriptional silencing.
The best-characterized model that describes the activation of transcription at a molecular level has been developed in Saccharomyces cerevisiae (47). Activator factors recruit the Rad6–Bre1 complex, which is loaded into RNA polymerase II (RPNII) and catalyzes the ubiquitylation of lysine 123 of histone H2B (Fig. 4). This ubiquitylation favors the recruitment of the histone methyltransferases Set1, Set2, and probably Dot1. Ubiquitylation is necessary for methylation of H3K4 and H3K79 by Set1 and Dot1, respectively. During elongation, the C-terminal domain of RNP II is phosphorylated at serine 2 and Set1 dissociates from RNP II, whereas Set2 methylates H3K36. H3K4 serves as a binding site for the recruitment of histone acetyltransferase SAGA complex through a specific interaction between methyl H3K4 and the chromodomain of its component Chd1. Histone acetylation by the SAGA complex leads to transcription activation. An important contribution has been made recently regarding the implication of histone arginine methylation in transcription. The activity of the Set 1 complex is regulated by the asymmetric dimethylation of H3R2. The patterns of H3R2me2a and H3K4me3 are mutually exclusive because H3 R2 methylation can prevent Spp1, a component of the Set1 complex, from binding H3K4me3 (27).

DNA Methyltransferases are responsible for maintaining the methylation patterns following DNA replication. DNMT1 exhibits a 5- to 30-fold preference for hemimethylated substrates. This property led to the identification of DNMT1 as the enzyme responsible for maintaining the methylation patterns following DNA replication. DNMT3a and DNMT3b were identified by EST database searches and were proposed as the enzymes responsible for de novo methylation. DNMTs originally were classified as either maintenance or de novo DNA methyltransferases, but several lines of evidence indicate that all three DNMTs cooperate and possess both these functions in vivo (49).

The cell reads the DNA methylaction code by the methyl-CpG binding proteins (MBDs). This family of proteins consists of five members, MeCP2, MBD1, MBD2, MBD3, and MBD4. They target the transcriptional repressive chromatin machinery to hypermethylated CpG islands (50). The first finding connecting DNA methylation and chromatin modification involved MeCP2, which can repress the transcription of methylated DNA through the recruitment of a histone deactylase complex, HDAC. Other evidences that connect DNA methylation with histone modifications are that DNMTs and MBDs recruit repressive HM Ts and HDACs complexes (51).

CpG islands of tumor-suppressor genes are hypermethylated in cancer cells. Therefore, they recruit multiple repressors that lead to a characteristic histone modification pattern: the deacetylation of histones H3 and H4, the methylation of lysine 9 of histone H3, and the demethylation of lysine 4 of histone H3 (52). In addition, it has been shown that the polycomb protein EZH2 associates with DNMTs and that specific methylation of lysine 27 of histone H3 is required to establish DNA methylation in a subset of target genes (65).
DNA methylation in mammals is an epigenetic modification involved in a range of cellular functions and pathologies, including X chromosome inactivation, genomic imprinting, tissue-specific gene expression, cell differentiation, regulation of chromatin structure, carcinogenesis, and aging, and is indispensable for the survival of differentiated cells.

Chemical Tools and Techniques

Epigenetic modifications are key translators between genotypes and phenotypes. The goal of scientists is to understand the significance of each chromatin modification within a biologic process, genomic region, and given organism. The simplest scenario would be that every epigenetic mark corresponded to definable and predictable outcomes. However, many lines of investigation have shown that the function of these chromatin marks is more complex than previously thought and depends entirely on the biologic context.

We can study the function of epigenetic modifications from different points of view: this includes the investigation of the global epigenetic modifications of every cell type to generate its epigenome and the other components involved in epigenetic modifications, such as the histone-modifying enzymes, the proteins that read each modification, and the enzymatic activities or functional protein complexes that are recruited to this modification, as well as the downstream biologic effects. Below we will discuss the technology that may be used to analyze epigenetic modifications.

The epigenome is the map of epigenetic modifications of a given cell in terms of DNA methylation and histone modifications. Any techniques for studying epigenetic modifications at specific loci exist, and several of them have been adapted for large-scale analyses.

Histone modifications

The most powerful technique for studying histone modifications coupled with particular DNA sequences is chromatin immunoprecipitation, ChIP, with antibodies against specific modifications. The immunoprecipitated DNA is analyzed by PCR with specific primers to investigate the presence of a candidate DNA sequence. ChIP may be scaled up for global analyses with microarrays (ChIP-on-chip). Its use with genomic platforms has begun to yield extensive maps of histone modifications in model organisms such as Arabidopsis thaliana, yeast, Drosophila melanogaster, mice, and recently normal human cells, including stem cells (53). The ChIP-on-chip assay can be applied to any histone modification for which an effective antibody is available, although it has been used only with lysine acetylation and methylation. Some important histone patterns nevertheless have been established already. ChIP analyses are limited by their requirement for good-quality antibodies, and they need to be highly specific if they are to yield results that can be comparable across experiments. The majority of ChIP experiments still use polyclonal antibodies, but the generation and use of monoclonal antibodies, by which differences between antibody batches may be minimized, is an important goal.

Mass spectrometry is the most accurate technique for detecting global levels of histone modification. However, it requires great technical expertise and is difficult to apply across genomes. Acceptable data on global levels of histone modification currently can be obtained by combining other methods. For example, all histones (H3, H4, H2A, H2B, and H1) can be isolated by HPLC, and the corresponding eluted fractions can be analyzed by HPLC and liquid chromatography-electrospray mass spectrometry (LC-ES/MS). Specific modifications at each amino acid residue also can be characterized by using antibodies in western blots, immunostaining, and tandem mass spectrometry (MS/MS).

DNA methylation

The most accurate method for analyzing DNA methylation is the bisulfite treatment of DNA, which reproducibly changes unmethylated cytosines to uracil but leaves methylated cytosine unchanged. Any laboratory can study DNA methylation by using the bisulfite treatment combined with genomic sequencing or amplification by methylation-specific PCR. To detect minimal amounts of aberrant DNA methylation, quantitative PCR-based methods can be used, such as the bisulfite treatment in combination with MethylLight or pyrosequencing (54). Some other approaches are based on the uses of restriction enzymes that can distinguish between methylated and unmethylated recognition sites in genes of interest. This technology is less accurate because incomplete cutting of the restriction enzymes within the studied regions is a limitation. Both methods can be coupled with several genomic approaches for detecting DNA methylation patterns. Recently, a ChIP-on-chip-based method has been applied to DNA methylation, the methyl-DIP (54). The DNA that is immunoprecipitated with an antibody against 5-methylcytosine can be used as a probe for hybridization to genomic microarray platforms, which allows the rapid identification of multiple Cpg sites and simplifies the analysis of the DNA methylation.

References

Epigenetic Modifications


Further Reading
MethDB, Human DNA Methylation Database, http://www.methdb.de/

See Also
AdoMet-Dependent Methyltransferases, Chemistry of DNA. Covalent Modifications of Histone Acetyltransferases, Selective Inhibitors of Post-Translational Modifications to Regulate Protein Function
Pre-mRNA splicing

John Karijolich and Yi-Tao Yu, University of Rochester Medical Center, Rochester, New York
doi: 10.1002/9780470048672.wecb520

Pre-mRNA splicing is an RNA processing reaction by which introns are removed from an mRNA precursor and exons are precisely joined together to form a functional mature mRNA. Splicing involves two successive trans-esterification reactions that occur in the spliceosome, which is a multicomponent complex composed of a large number of protein factors and five small nuclear RNAs (snRNAs), each functioning as an RNA–protein complex (ribonucleoprotein or snRNP). A significant body of evidence indicates that both spliceosome assembly and catalytic core formation are orchestrated by an intricate network of RNA–RNA interactions. Most eukaryotic introns invariably start with the dinucleotide GT (GU in RNA) and end with the dinucleotide AG. Splicing of these introns occurs in the major spliceosome containing five abundant snRNPs (U1, U2, U4, U5, and U6). In high eukaryotes, a rare class of introns also exists, beginning with the dinucleotide AT (AU in RNA) and ending with the dinucleotide AC. Removal of this class of introns requires the participation of a different set of snRNPs, namely U11, U12, U5, U4atac, and U6atac. Recent advances in pre-mRNA splicing have provided strong evidence for an RNA enzyme catalyzing the two trans-esterification reactions in the spliceosome.

Introduction

It was discovered in 1977 that several adenovirus genes (in the form of DNA) contained intervening sequences that were not present in their mature mRNAs (1, 2). Shortly thereafter, other eukaryotic genes were similarly found to contain intervening sequences that disrupt the real hereditary message. Thus, the intervening sequences were named introns and the protein coding sequences (including the 5’ and 3’ untranslated regions) were termed exons (3). Since the discovery of introns, the phenomenon of intron-containing genes (often referred to as “split genes”) has been found to be widespread in eukaryotes. In fact, almost all genes in high eukaryotes contain introns (the exceptions are those encoding histones and interferons). Even in budding yeast, more than 250 genes contain introns. Early work on RNA metabolism in eukaryotes identified a large number of short-lived heterogeneous nuclear RNAs (hnRNAs) that are much longer than mature mRNAs in the cytoplasm. The hnRNA sequence can form a perfect double-stranded RNA–DNA duplex when hybridized with its own gene (4). Taken together, these observations produced the realization that transcription of a gene initially produces a perfect copy of RNA known as mRNA precursor (pre-mRNA or hnRNA) containing not only exons but also introns. Introns must be removed from the precursor before the resulting mature mRNA is transported to the cytoplasm where it directs the translation of protein. The removal of introns from mRNA precursor is catalyzed by an RNA processing reaction known as pre-mRNA splicing. Pre-mRNA splicing occurs in the spliceosome, a massive protein–RNA complex consisting of a large number of proteins and five small nuclear RNAs (U1, U2, U4, U5, and U6 snRNAs), each of which functions as a small nuclear ribonucleoprotein (snRNP) (5–8). Over the years, the spliceosomal snRNAs (or snRNPs) have been extensively studied and their roles in pre-mRNA splicing have been well established. A fairly detailed picture has emerged in which intricate networks of interactions among the components of the spliceosome, in particular the RNA–RNA interactions, orchestrate the assembly of the spliceosome as well as the formation of the catalytic center for catalysis (chemical reaction of splicing) (6, 9–11). Here, we discuss spliceosome assembly and pre-mRNA splicing, with an emphasis on the U snRNAs as well as some important spliceosomal proteins (both snRNP proteins and non-snRNP proteins).
Pre-mRNAs Contain Conserved Sequence Elements at the Splice Sites

In the late 1970s and early 1980s, the sequences of many eukaryotic genes became available for study, thus making it possible to conduct a comparative analysis to identify important sequence elements in pre-mRNA. Careful inspection of higher eukaryotic genes revealed several consensus sequences in introns at or near the 5′ and 3′ splice sites (12, 13). The 5′ splice site consensus sequence is G/GURAGU (ü represents the 5′ exon-intron junction; R depicts a purine; the underlined dinucleotide GU is invariant). The 3′ splice site is YAG/G (here, ü represents the 3′ intron-exon junction; Y is a pyrimidine; the underlined dinucleotide AG is invariant), which is frequently preceded by a GU-rich region. In addition to the 5′ and 3′ splice sites, there is another consensus sequence, namely the branch site YNYURAGC (N can be any nucleotide and the underlined adenine is invariant) located 20–40 nucleotides upstream of the 3′ splice site (Fig. 1). In budding yeast, the three consensus sequences at the 5′ and 3′ splice sites (including the branch site) have also been identified; however, there is no or a less-conserved GU-rich region preceding the 3′ splice site. In contrast to the loosely defined consensus sequences in high eukaryotic pre-mRNAs, the sequence elements in yeast pre-mRNA are almost absolutely conserved: G/GUAGU at the 5′ splice site, CAG/G at the 3′ splice site, and UACUAAC at the branch site. Mutation of any invariant nucleotides within the consensus or conserved elements results in a total inhibition or a decreased

Figure 1 Major spliceosome assembly and action. First, U1 (red) assembles onto the pre-mRNA (in the 5′ complex) to form a commitment complex (CC). Then U2 (green) joins, converting complex CC to complex A. Subsequent joining of the U4–U5–U6 (blue–brown–purple, respectively) tri-snRNP forms the major spliceosome (complex B1). The newly assembled spliceosome then undergoes a dynamic rearrangement of RNA–RNA interactions resulting in the release of U1 and U4 and the formation of complex B2. Formation of complex B2 triggers the first catalytic step, the formation of the cut-off 5′ exon and the lariat structure, thus leading to the formation of a new complex (complex C1). After additional conformational changes, step 2—the joining of 5′ and 3′ exons—occurs and complex C2 forms. mRNA is then released, resulting in the formation of complex I. Finally, the lariat intron is released and the spliceosome disassembles. The 5′ splice site (5′-SS), the 3′ splice site (3′-SS) and the branch point adenosine (BP) are indicated. The consensus sequence elements at the 5′ and 3′ splice sites and the branch site are shown (R, purine; Y, pyrimidine; Yn, the polypyrimidine tract; N, any nucleotide). The thick lines represent snRNA strands (in colors) or the intron (black), and the boxes are exons. The short thin lines between RNA strands represent Watson–Crick base-pairing interactions. The lightning bulbs depict non-Watson–Crick base-pairing interactions. The bulged-out branch point adenosine is pictured after the joining of U1. The blue arrows indicate the chemical reactions, trans-esterifications, or nucleophilic substitution reactions. The yellow areas in complexes B2 and C1 represent the catalytic center for catalysis. U2–U6 duplexes [H helices 1 (or A and B), II and III] in the spliceosome are shown. The U6 intramolecular stem structure (intra-stem) is also pictured. The extended U6 intramolecular stem (before the first step of splicing, complex B2) or the U6 intramolecular stem plus U2–U6 H helix B (before the second step of splicing, complex C1), constitutes a structure that resembles domain V (active site) of group II introns (self-splicing introns). Note: Different nomenclatures are used to describe splicing complexes in mammals and yeast. Here, the nomenclatures are adopted from the mammalian system (66).
level of pre-mRNA splicing in both yeast and higher eukaryotes, further confirming their importance in the splicing process. The question that then develops is what role do these consensus sequences (and the invariant nucleotides) play during splicing? Are they recognized by other trans-acting cellular components, and if so, what are they?

### Early Recognition of the 5′ Splice Site and the Branch Site

Although identified and characterized in the 1960s (14, 15), U1 and U2 snRNAs had not been assigned functions until the discovery of split genes (or pre-mRNA splicing). In an attempt to link U1 to pre-mRNA splicing, researchers first analyzed the sequence complementarity between the snRNA and the conserved sequence elements in the pre-mRNA. Excitingly, the sequence analyses revealed a striking complementarity between the evolutionarily invariant 5′ end of U1 and the pre-mRNA consensus sequence at the 5′ and 3′ splice sites (16, 17). Although the original proposals for the involvement of U1 snRNA in splicing suggested that the 5′ end of U1 may base-pair with consensus sequences surrounding the 5′ and 3′ splice sites, in 1981 Mount and Steitz modified the model after observations that certain sequences in U1 predicted in binding the 3′ splice site were not conserved (18). Driven by this hypothesis, researchers carried out a series of experiments to confirm the involvement of U1 in pre-mRNA splicing. For instance, in 1983 the Steitz lab provided the first evidence that U1 snRNA’s primary role was the recognition of the 5′ splice site (19). Using an ultrathin transcribed RNA fragment of the major mouse β-globin gene, which contains an intron, they localized the U1 snRNP binding site by T1 ribonuclease treatment of the snRNA-RNA complex, immunoprecipitation of the resulting trimmed complex with anti-U1 antibodies, and T1 ribonuclease fingerprinting of the immunoprecipitate. Fingerprinting analyses revealed three oligonucleotideides present: CAG, UUG, and UAUCCAG. All three of these fragments can be found together in a single region that encompasses the 5′ splice site (GGGCGAGUGUUGUCCAG) (19). In a separate experiment, the 5′ and 3′ end region of U1 was destroyed by oligodeoxynucleotide-directed RNase H digestion, resulting in a loss of splicing activity (20). In addition, an elegant genetic suppression assay was used to verify the base-pairing interaction between U1 and pre-mRNA (21). In that work, a point mutation was first introduced into the 5′ splice site region of a pre-mRNA. As expected, splicing of the pre-mRNA was inhibited. However, when a compensatory point mutation was introduced into U1 snRNA at the site (within the 5′ end region) that was predicted to restore the base-pairing interaction, splicing activity was indeed rescued (21). Moreover, posttranslation cross-linking also detected base-pairing interaction between U1 and pre-mRNA (22). Along the same line, U1 was physically detected in the splicing complex (see below). Such an enormous amount of data has clearly demonstrated that, indeed, U1 recognizes pre-mRNA during splicing and that the recognition involves a direct base-pairing interaction between the U1 5′ end region and the pre-mRNA 5′ splice site (see Fig. 1). Binding of U1 as well as other splicing factors to the 5′ splice site of a pre-mRNA leads to the formation of an early complex, termed a commitment complex, that commits pre-mRNA to the splicing pathway (see below).

### Splicing Factors Recognize the 3′ Splice Site

U2AF (U2- auxiliary factor) is one of several splicing factors that have been implicated in 3′ splice site recognition (36, 39). U2AF has two subunits, namely U2AF65 (65 KD) and U2AF35 (35 KD). It has long been known that U2AF65 binds to the polypyrimidine track upstream of the 3′ splice site of pre-mRNA in high eukaryotes (36, 39, 41). However, the role of U2AF35 had remained elusive until 1999 when a connection between U2AF35 and the dinucleoside AG at the pre-mRNA 3′ splice site was finally established (42–44). Using a variety of experimental approaches, including U2AF depletion and reconstitution, cross-linking and immunoprecipitation, Soléx, nucleoside protection, and mutational analysis, three research groups independently demonstrated that U2AF35 is actively recognizing the dinucleoside AG at the 3′ splice site during splicing (42–44). Other splicing factors may also play important roles in 3′ splice site selection. For example, using the yeast genetic approach and the mammalian in vitro splicing system, researchers demonstrated that the splicing factor SiU7 is important for the selection of the canonical AG at the 3′ splice site during splicing (45, 46).
Splicing Occurs Via a Two-Step Trans-esterification Reaction Pathway

In the early 1980s, a mammalian in vitro splicing system (HeLa nuclear extracts) was developed, leading to the discovery of two splicing intermediates, the cut-off 5'-exon intermediate and the 2/3 lariat intermediate (intron-3' exon in a lariat form) (47, 48). Importantly, the lariat (or branched) structure was also identified in celluly derived RNAs (49, 50). Soon after, the yeast in vitro system also became available, and it, too, allowed the detection of the cut-off 5’ exon and the 2/3 lariat intermediates (51). A two-step trans-esterification reaction pathway was thus established (Fig. 2). In the first step, the 2’-OH of the branch point nucleotide adenosine attacks the phosphate at the 5’-exon–intron junction (5’ splice site), resulting in the cleavage of the phosphodiester bond between the 5’-exon and intron and the concurrent formation of a new 5’-2’ phosphodiester bond between the 5’ end of the intron and the branch point adenosine. Thus, a lariat-structured intermediate (the 2/3 lariat intermediate) and the cut-off 5’-exon intermediate are produced. In the second step, the 3’-OH group of the cut-off 5’-exon attacks the phosphate at the intron-3’-exon junction (3’ splice site), releasing the lariat intron product and generating the spliced mature mRNA product. According to this pathway, the branch point adenosine and the guanosines at the 5’ and 3’ ends of the intron are the key nucleotides as they directly participate in the chemical reactions. In fact, these three nucleotides are among the most conserved nucleotides in pre-mRNA (see above).

Spliceosome Assembly Pathway

It was anticipated that splicing occurs in a large RNA-protein complex (or the spliceosome), given that multiple snRNAs and proteins can be identified as functional components that participate in splicing. However, how big the spliceosome is and, perhaps more importantly, how it is assembled had remained largely unclear. The availability of the in vitro splicing systems in vitro during splicing (51). A two-step trans-esterification reaction pathway was thus established (Fig. 2) (66, 67). Finally, the lariat intron product is released at the branch point adenosine and its 2’-OH group are pictured. The lines represent the intron and boxes depict exons.

The assembly starts with a nonspecific H complex that forms as soon as the pre-mRNA is mixed with the splicing extracts (62, 66, 68, 69). This complex is then turned into a commitment complex (CC complex or the E complex) after binding with U1 snRNP and various protein factors, including SR proteins, U2AF (Mud2p in yeast), and SF1 (BBP/Msl5p in yeast) (23-28). The commitment complex is a complex that is committed to splicing out the intron on which it assembles. U2 snRNP then joins the commitment complex, converting it into a presplicing complex, namely complex A (27, 35, 58, 60-62, 66, 67, 69). After binding with the preformed functional U4/U6/U5 tri-snRNP (70-74), in which U4 and U6 are extensively base-paired with each other, complex A is then converted to complex B1 (61, 66, 67). Complex B1 undergoes a conformational change, and it becomes complex B2, which is now competent for the first step of splicing (62, 66, 67). After the first step of splicing and additional conformational changes, complex B2 is turned into complex C1 in which the second step of splicing occurs (63, 66, 67). Complex C1 then changes to complex C2, which contains spliced products (61, 66, 67). The mature mRNA is then released, changing complex C2 to complex I (62, 66, 67). Finally, the lariat intron product is released.
Dynamic Rearrangements of RNA–RNA Interactions within the Spliceosome

Although kinetic analysis of spliceosome formation led to the discovery of a stepwise assembly pathway, the detailed network of interactions among the components within the massive spliceosome had remained unclear until the early 1990s when a series of cross-linking and yeast genetic suppression experiments were carried out to tackle this problem. Several cross-linking techniques, including psoralen cross-linking to detect Watson-Crick base-pairing interactions, 4-thioU cross-linking to detect non-Watson–Crick contacts between two nucleotides, and direct short wavelength UV cross-linking to detect various interactions, were used effectively to capture the interactions between U snRNAs and pre-mRNA and among U snRNAs themselves (22, 75–79). Cross-linking was performed at various time points in parallel with splicing and spliceosome assembly assays, such that any detected cross-linking species could be assigned to a particular stage (splicing complex) during spliceosome assembly and splicing. In the meantime, sensitive yeast and mammalian genetic suppression assays also identified several novel RNA–RNA interactions (80–88). Recently, some of the RNA–RNA interactions within the spliceosome were further refined through the use of NMR spectroscopy (89). The large amount of data accumulated over the years provided a detailed picture revealing a network of RNA–RNA interactions within the spliceosome.

In addition to the early base-pairing interactions between U1 and the 5′ splice site and between U2 and the branch site, a number of new RNA–RNA interactions were detected at late times during spliceosome assembly, especially after the spliceosome is fully assembled (complex B1) (Fig. 3). Interestingly, the formation of some new interactions is accompanied by the dissociation of old interactions, indicating that dynamic rearrangements occur within the spliceosome. Specifically, entry of the U4/U6–U5 tri-snRNP to form complex B1 may be mediated at least in part through the base-pairing interaction between the 5′ end of U2 and the 3′ end of U6 (U2–U6 Helix II) (78, 82, 86). As soon as the U4/U6–U5 tri-snRNP enters the splicing complex, U5 snRNA interacts, in a non-Waston-Crick pairing manner, with the exon sequence at the 5′ splice site, and possibly with the exon sequence at the 3′ splice site as well (75–77, 83, 84, 87). In the meantime, a conserved sequence of U6 (ACA-GAGA) displaces U1 in interacting with the 5′ splice site (79, 88). Concurrently, or immediately after, U6 dissociates from U4. The freed U6 sequence then forms new duplexes with U2 (Helices I and III, see Fig. 3) (22, 81, 85, 89), which are believed to be at least part of the catalytic center (9, 30). This dynamic rearrangement results in the release of both U1 and U4 from the spliceosome and the formation of an active catalytic center for the first step of splicing (complex B2). After the first chemical reaction, the spliceosome undergoes additional conformational changes, resulting in the disruption of the base-pairing interaction between U6 and the 5′ splice site, the formation of U2–U6 Helix 1b, and the repositioning of the reaction substrates within the catalytic center (complex C1) (80, 89, 90). These conformational changes directly trigger the second step of splicing that generates the mature mRNA and a lariat intron products (complex C2). After the second step of splicing, further conformational changes occur, leading to the release of mature mRNA (complex I). Final conformation changes then take place, resulting in the release of the lariat intron and disassembly of the spliceosome (or complex I).

Protein Factors Facilitate and Modulate Spliceosomal Conformational Changes

It has long been proposed that the rearrangements of RNA–RNA interactions and/or conformational changes are orchestrated by spliceosomal protein factors (7). Indeed, multiple such protein factors have been identified, and many of them are members of the family of DEAD box ATPases, which can either directly promote the unwinding of RNA–RNA duplexes (also referred to as unwindases) or disrupt RNA–protein interactions (or even protein–protein interactions) (91). These spliceosomal ATPases function in a substrate-specific way, such that they each promote a specific conformational change (RNA–RNA interaction) at various stages along the pathway of spliceosome assembly and catalysis. For instance, Prp5 is an RNA-dependent ATPase that bridges U1 and U2 snRNPs at early times during spliceosome assembly, promoting stable interaction of U2 with the intron branch site (40, 92) and thus promoting complex A formation. As is, there is evidence that the incorporation of U2 snRNP into the spliceosome is facilitated by Sub6/UPF56 (U2AF-associated protein), another RNA-dependent spliceosomal ATPase that acts through disruption of an interaction between U2AF and the branch site (93). Brm, a U5 snRNP-associated protein and also an RNA-dependent ATPase, promotes the dissociation of U6 from U4 (94) in the spliceosome (presumably soon after complex B1 formation). In vivo analysis in S. cerevisiae suggests that the spliceosomal ATPase Psp28 may be actively disrupting the interaction between U1-C and the 5′ splice site so as to allow for the exchange of U1 for U6 (95, 96), which is part of the rearrangement occurring during the transition from complex B1 to complex B2. The spliceosomal ATPase Psp22, on the other hand, promotes the first step of splicing (97–101), presumably also through rearranging the RNA–RNA interactions required for catalysis. The spliceosomal ATPase Psp15 promotes the conformational changes required for the second step of splicing (transition from complex B2 to complex C1) (90, 102–105). Psp22, a spliceosomal ATPase as well, facilitates the conformational changes required for the release of mRNA (106, 107), promoting the transition from complex C1 to complex I. Finally, another spliceosomal ATPase Psp43 orchestrates

Pre-mRNA splicing

![Diagram of spliceosome](image-url)
the conformational changes required for the release of the lariat intron (108).

Besides DEK/DOG box ATPases, Snu114, a component of US snRNP and a GT-Pase, is also required for splicing, presumably through promoting the rearrangement between PpD and Snu114 itself, thus leading to the release of U1 and U4 before the first step of splicing (109). In addition to the ATPases/GT-Pases, several other spliceosomal proteins facilitate the rearrangements as well. Included in these proteins are the US snRNP-associated proteins p116 and p20 (Prp3p in yeast), which are thought to play a regulatory role in the unwinding of U4/U6 (110).

Yeast genetic analyses indicate that PpD also plays a role in modulating the function of the spliceosomal ATPases, Prp16 (90, 100) and Prp22 (113). Likewise, Isy1, a component of Prp NineTen Complex (NTC), plays a role in facilitating the function of Prp16 (112). Similar roles can be assigned to U1C and Cus2, which are implicated in modulating the functions of PpD and Prp5, respectively (96, 113). Importantly, the spliceosomal ATPases and their modulators are not only the key factors in orchestrating the conformational changes during spliceosome assembly and catalysis, but also play important roles in maintaining the fidelity of splicing. In this regard, it has been proposed that incorrect conformations (e.g., those caused by mutations in pre-mRNA and/or other spliceosomal factors) are discarded through these conformational changes catalyzed by the spliceosomal ATPases and their modulators (90, 108, 114–116).

**Prefabricated Spliceosome?**

In contrast to the stepwise spliceosome assembly model, recent observations have indicated that the spliceosome exists as a pre-formed complex that engages the pre-mRNA as such (117–121). The strongest evidence for this model comes from the S. cerevisiae system from which a 45S snRNP complex was isolated that contained all five snRNPs (121). This penta-snRNP complex contains nearly all known U1, U2, and U4/U6-US snRNP proteins as well as non-snRNP splicing factors. The complex has also been shown to function in yeast cell extracts depleted of endogenous RNAs by nuclease digestion. Importantly, while assembling onto the pre-mRNA, the snRNA constituents of the penta-snRNP do not exchange with endogenous snRNPs. Consistent with the idea of a preassembled spliceosome, it has also been reported that both U2 and the U4/U6-US snRNP function before the formation of complex A (122). Furthermore, recent analysis of splicing complex formation in HeLa nuclear extracts argues that a preassembled 200S RNP complex containing all snRNP components can assemble onto a short RNA containing a 5′ splice site (123). However, using chromatin immunoprecipitation/more (ChIP) analysis, the Neugebauer group (124, 125) and the Rambach laboratory (126, 127) have more recently suggested that the recruitment of spliceosomal snRNPs to nascent pre-mRNA (in vivo spliceosome assembly) in yeast occurs via a stepwise assembly pathway, which is similar to that observed in vitro (see above, Spliceosome Assembly Pathway). Thus, it remains controversial as to whether spliceosome assembly is a one-step process (the prefabricated penta-snRNP complex assembles onto pre-mRNA) or it proceeds in a stepwise fashion. Further study is necessary to clarify this important issue.

**Is Splicing an RNA-Catalyzed Reaction?**

It has long been suspected that the two trans-esterification reactions to remove the introns from a pre-mRNA are catalyzed by the RNA constituents in the spliceosome (128). This idea is bolstered by the fact that U snRNAs in the spliceosome do form functional structures (the extended U6 intramolecular stem in complex B2 and the U6 intramolecular stem plus U2–U6 Helix I b in complex C1; see Fig. 1) resembling domain V of group II introns, some of which are self-spliced via the two-step trans-esterification pathway identical to that of spliceosome-catalyzed pre-mRNA splicing (see above). Domain V of group II introns constitutes the catalytic center for catalysis during self-splicing (129–131).

Over the past decade, a growing body of evidence has suggested that pre-mRNA splicing in the spliceosome may indeed be catalyzed by its RNA constituents. For instance, Sontheimer et al. have shown that the first catalytic step of splicing occurs through a metal-ion-dependent pathway (132), an observation consistent with the two metal-ion model proposed for the spliceosome active site(s) (133). Later, using sulfur substitution followed by manganese suppression, Lin and colleagues have demonstrated that, through metal ion coordination, U6 plays a critical role in the catalysis reaction (134). Most recently, Valadkhan and Manley carried out an even more direct experiment to address this issue (135). In that experiment, they used only three short RNA oligonucleotides, corresponding to U2, U6 (designed according to U2–U6 Helices I, II, III, and the U6 intramolecular stem, see Figs. 1 and 3) and the branch site, respectively, and no proteins were included. Remarkably, the U2 and U6 oligonucleotides do form the functional structure for the first step of splicing (U2–U6 Helix I plus the extended intramolecular stem of U6), leading to the production of X-RNA, a product generated by a splicing-like reaction by which the branch point adenosine in the branch site oligonucleotide attaches a phosphate in the U6 oligonucleotide (Fig. 3). Taken together, the evidence accumulated thus far strongly supports the RNA-catalysis model for the spliceosome.

**A Parallel Spliceosome**

In high eukaryotes, there also exists a minor population of introns containing distinct consensus sequences at their 5′ and 3′ ends (136). The removal of these introns takes place in an analogous spliceosome containing a different set of U snRNPs (U11, U12, U5, U4atac, and U6atac) (137). Although initially dubbed AT-A C introns based on their termini, extensive genomic database surveys have revealed that the standard GT-A G terminal sequences are more prevalent (138, 139). Instead, what appears to distinguish this class of introns is longer and more
constrained consensus sequences at the 5′ end of the intron and the branch site, as well as the absence of a polypyrimidine tract upstream of the 3′ splice site (138, 139). The splicing machinery responsible for the removal of these introns is of much lower abundance (∼10^6 copies per cell) relative to components of the major spliceosome, which is compatible with the low frequency in which these introns appear in the genome (∼1/300 human introns) (140).

Remarkably, U11, U12, U4atac, and U6atac form structures that are almost identical to their counterparts in the major spliceosome, namely U1, U2, U4, and U6, respectively, despite the fact that the two sets of snRNAs are quite different in primary sequences (136, 137). Equally strikingly, the network of RNA–RNA interactions detected in the major spliceosome also exists in the U12-dependent spliceosome, further validating the importance of these dynamic interactions during spliceosome assembly and splicing (136, 137).

The main mechanistic differences between the two spliceosomes occur at the stage of intron recognition, rather than catalysis. Indeed, recognition of the 5′ splice site and the branch site occurs simultaneously by the U11-U12 di-snRNP (141). Furthermore, there is a requirement among U12-dependent introns for 5′ exon sequences to form U6atac-3′ splice site interactions (142). Lastly, more constrained consensus sequences, as well as the lack of a polypyrimidine tract, suggest that the assembly of this spliceosome is more dependent on snRNA-based interactions than assembly of the major spliceosome (140, 143).

Acknowledgments

We thank our colleagues in the Yu lab for inspiration. Our work was supported by grant GM62937 (to Y.-T. Yu) from the National Institute of Health.

References

Pre-mRNA splicing


Pre-mRNA splicing


Pre-mRNA splicing


RNA Interference, Mechanisms and Proteins Involved in

Wang-Xia Wang and Peter T. Nelson, Department of Pathology and Sanders-Brown Center, University of Kentucky, Lexington, Kentucky
Guiliang Tang, Department of Plant and Soil Sciences and KTRDC, University of Kentucky, Lexington, Kentucky
doi: 10.1002/9780470048672.wecb517

RNA interference (RNAi) and microRNA (miRNA) pathways participate in posttranscriptional gene expression regulation. RNAi and miRNA biochemistry involve short (~22 nucleotide) regulatory RNAs: small interfering RNA (siRNA) and miRNA. Both siRNAs and miRNAs guide RNA–protein complexes to “target” complementary messenger RNAs (mRNAs) for cleavage or translational repression. Many proteins involved in these pathways have been elucidated. The biochemical mechanisms of miRNA/siRNA processing have been dissected in detail in plants and in animals using biochemical and genetic approaches. Here we review the current literature on RNAi and miRNA mechanisms with a special focus on the protein components.

Introduction

Much gene expression regulation occurs after mRNA transcription (1), and it is becoming increasingly clear that noncoding “regulatory” RNAs play central roles in posttranscriptional gene expression regulation. The breakthrough discovery of small regulatory RNAs and their multiple functions in gene regulation has heralded a fascinating new field in biology. Pioneering observations on the phenomenon of posttranscriptional gene silencing were reported in plants (2–5). However, the broader relevance of gene silencing mechanisms is confirmed in many other organisms, including the Nobel Prize winning discovery of RNA interference (RNAi) in worms (6). RNAi-related biological phenomena have now been elucidated in all known eukaryotic organisms. The similarities in RNAi mechanisms between plants and animals underscore the importance of this ancient gene regulatory mechanism. Moreover, in addition to mRNA “gene expression regulation,” RNAi-related mechanisms were shown to have diverse functions including transposon silencing (7), chromosome remodeling (8), and DNA methylation (9).

The molecular processes responsible for RNAi and its related pathways have been elucidated extensively over the last few years. Here we review the literature about the mechanisms of RNAi and miRNA pathways with a focus on the proteins involved.

Complementary Research Strands Come Together

In 1990, Jorgensen (10) attempted to enhance flower color by overexpressing an extra copy of chalcone synthase (chsA) gene in transgenic petunia plants (3). Unexpectedly, among many transgenic petunia plants both the endogene and the transgene of chalcone synthase were silenced. The transgenic flowers showed multiple colors or generated white segments. Loss of cytosolic chsA mRNA was not associated with a reduction of the gene transcription (11), which indicates that the phenotype resulted from a post-genetic transcriptional event. The phenomenon of silencing an endogenous plant gene, triggered by extra copies of a transgene with the same sequence, was termed “cosuppression” (10, 12).

Important advancements in the field that would eventually encompass RNAi and miRNAs came from the laboratory of Victor
A. Ambros et al. discovered that short noncoding RNAs (the product of the lin-4 gene) regulated the translation of another gene product (lin-14) via partially complementary sequences in the 3′UTR of lin-14 mRNA (13, 14). The significance of these data remained temporarily obscure—it was not until seven years later that a similar small regulatory RNA was discovered (15).

Thereafter, it was recognized that these molecules were representatives of a large, biologically important class of RNA (the designation “microRNAs” was first described in separate papers in 2001) (16-18). Although miRNAs work through pathways somewhat different from siRNAs (see below), the paradigm of antisense recognition by small regulatory RNAs, coupled with posttranscriptional regulation of mRNA, was thus established in animals.

In 1995, when worms were injected with antisense and sense RNAs to target a specific gene for down-regulation, Gao and Kemphues (19) found that both the control sense and the antisense RNAs induced similar gene silencing effects in worms. This phenomenon could not be explained fully at that time. In the seminal work of Fire et al. (6), dsRNA was finally demonstrated as a potent trigger in gene silencing by introducing tiny amounts of purified dsRNAs into C. elegans, and this phenomenon was termed “RNA interference”.

Although RNAi phenomena were described in different experimental contexts (labeled cosuppression in plants, RNAi in animals, and quelling in fungi), all these gene-silencing mechanisms have converged on a general paradigm. The common features of this paradigm share at least three characteristics: 1) Short (null) regulatory RNAs (siRNAs) of 21–23 nucleotides (nts) are the key players in mediating specific RNA degradation, 2) the degradation of target RNA depends on specific sequences, and 3) the machinery of the RNA-induced silencing complexes (RISCs) is similar in structure and function across most organisms.

The Mechanism of RNA Interference

Although RNAi and miRNAs were discovered in worms, major progress in revealing the RNAi pathway came from in vitro biochemical studies using a cell-free system from Drosophila syncytial blastoderm embryos, which was first carried out by Tuschi et al. (20). It was hypothesized that a population of 21–23 nt small RNAs could be generated in fly embryo lysates, which triggers the cleavage of radioactively labeled target mRNA at 21–23 nt intervals (21). Experimentally, these short RNA fragments were indeed detected in extracts of Drosophila Schneider 2 (S2) cells that were transfected with dsRNA (22). Moreover, RNA molecules of similar size were also found to accumulate in transgenic silenced plants (23). Based on these studies, Zamore et al. (21) proposed an early model of RNAi: Long dsRNAs are first cleaved into short dsRNA fragments of 21–23 nts. These short dsRNAs are then separated and integrated into a ribonucleoprotein complex, and the siRNA-ribonucleoprotein complex recognizes and binds the siRNA-specific complementary RNA that leads to destruction of the latter. This model laid a framework for the subsequent dissection of the RNAi pathway.

Generation of siRNAs and miRNAs

SiRNAs can be derived from several sources including long dsRNAs. dsRNAs can be produced many ways. They can be generated endogenously, for example, from transgene transcripts by RNA-directed RNA polymerase (RdRP) (24-27); by simultaneous transcription of sense and antisense DNA fragments from specific genomic loci (28); from viral replication intermediates (27, 29); or via introduction of genes that encode inverted repeats by genetic engineering (30, 31). Endogenously generated or exogenously introduced dsRNAs are then converted into 21-30 bp small RNAs by dicer (DCR), which is an RNase III family enzyme that initiates the RNAi (32).

miRNAs, which are derived from endogenous genes, are processed differently from siRNAs. A single-stranded RNA can form a “hairpin” or “stem-loop” structure where the RNA folds back and base pairs with itself. The “stem” of this structure thus serves as dsRNA for additional processing. Some “stem-loop” RNAs are processed to become miRNAs. MiRNA genes are transcribed by RNA polymerase II (pol II), mostly or in a few cases by pol III (33, 34), and the primary transcripts “pri-miRNAs” are processed by a microprocessor composed of a Drosha-like enzyme of RNase III family and a Pasha-like protein of dsRNA binding protein family (35-39). Or, in plants, pri-miRNAs are processed by the DCR family member DCL1 (40). Like all RNase III enzymes, Drosha leaves two nt 3′ overhangs and 5′ monophosphate groups. Processed pre-miRNAs are transported to the cytoplasm via the export-receptor Exportin-5 (41-43). MiRNAs not only are generated from miRNA genes located within intergenic regions, but also are derived from introns. A subset of intronic miRNA precursors from Drosophila can bypass the microprocessor and directly serve as DCR substrates (44, 45). These miRNAs are known as mirtrons. The spliced-intron lariats or mirtron precursors are debranched and refolded into stem-loop structures that have similar Drosha ends and are recognized by DCR1 in Drosophila (44, 45).

The final stages of miRNA maturation are deceptively complex. As indicated above, both dsRNA and stem-loop RNA structures such as miRNA precursors are recognized and processed by the cytoplasmic RNase III enzyme DCR (32). Humans and C. elegans encode only one DCR, which can process both dsRNA and mirtron precursors; Drosophila has two DCR genes, and the Arabidopsis genome encodes four DCR homologs with each responsible for distinct pathways (46). DCR binds and cleaves dsRNA endonucleolytically and generates a 21-25 bp long dsRNA (47-50). The short dsRNA fragments are separated, and one strand is incorporated into an effector complex, called RNA-induced silencing complex (RISC) (22), where it binds directly to a member of the Argonaute protein family. A Argonaute proteins— in plants as well as in animals—are at the heart of the function of RNAi and its related pathways.
The activity of Argonaute is affected dramatically by the sequence specificity of the loaded small regulatory RNA (i.e., the degree to which the small RNA is complementary to a "target" mRNA) (see below). In plants, miRNA-containing RISC guides predominantly target cleavage because plant miRNAs are usually extensively complementary to their target genes (51), although bona fide translational repression has been described in plants (52). In contrast, mammalian miRNAs are usually not fully complementary, and miRNA-guided posttranscriptional regulation leads primarily to the repression of gene expression via translational repression. To distinguish them from mRNA cleavage reactions mediated by RISC, effector complexes that contain mammalian miRNAs are often referred to as miRNA-RNP (53, 54).

The RISC Components and Function

The action of RNA interference is carried out by RISC. RISC is a ribonucleoprotein particle (RNP) with a "core" structure composed of a single-stranded siRNA or miRNA and an Argonaute protein (AGO). Depending on both the isoform of the AGO and the degree of complementarity between the siRNA/miRNA and the target sequence, interaction of RISC with mRNAs can result in multiple silencing events including endonucleolytic cleavage, translational repression, sequestration to P-body compartment, and mRNA degradation (Fig. 1).

Although AGO and small regulatory RNAs are the sine qua non, other proteins are also important to regulate the function of RISC. The complete protein repertoire of RISCs is not elucidated fully. The reported sizes of RISC vary from 160 kDa (minimal RISC) to 280 kDa (holo-RISC). The "holo" RISCs were found to be associated with many proteins in addition to the AGO protein and the small RNA. Both in vitro and in vivo studies showed that an AGO protein and a guide strand siRNA alone can form a functional "core" RISC to cleave the target mRNA. For example, either recombinant human AGO2, or biochemical-purified RISC from human or Drosophila cell lysate, are active complexes capable of substrate binding and cleavage (22, 55–61). Similarly, in Arabidopsis, the immunopurifiﬁcation of AGO1 proteins can interact with single-stranded siRNAs to form active RISCs (62). Both recombinant and endogenous RISC-directed target cleavages occur in a divalent cation (Mg$^{2+}$ or Mn$^{2+}$)-dependent manner. Other examples, however, indicate that activity of RISCs needs additional protein components or cofactors for its function. For example, in S. pombe, site-directed cleavage of the promoter region of the his3 gene requires the association of RITS complex with AGO1 protein (56–58).

Figure 1: Mechanisms of small RNA-protein complex mediated gene suppression. In cytoplasm, siRNAs and miRNAs are sorted and integrated into different types of RISC/RNP complexes. RISCs/RNPs that contain AGO proteins with or without RNase H activity (slicer) and the loaded siRNAs or miRNAs, execute gene suppression function via different mechanisms: 1) target mRNA cleavage by the slicer; 2) prevention of translation initiation by excluding the translation initiation factor eIF4E from binding to the mRNA-cap (m$^7$G) via an eIF4E-like domain (N2C) located on the AGO proteins; 3) ribosome drop-off from the target mRNA, which is induced by the RISC/RNP; 4) sequestration of target mRNAs to P-bodies for temporary storage or degradation. Under certain conditions (e.g., stresses), miRNPs help release the target mRNAs from P-bodies to resume the target mRNA translation. In the nucleus, siRNA-loaded RITS complex directs Chp1 to methylate of the ninth lysine of histone H3. RITS complex recruits the cellular proteins such as Chp1, Swi6, and TAS3, which lead to the chromatin remodeling and shutdown of gene expression in S. pombe. In humans, synthetic siRNAs that target promoter regions can interact with AGO2 protein and demethylate the H3K9, which leads to the activation of the target genes.
example, the RISC-directed multiple turnover target cleavage needs ATP, although a component that binds and hydrolyzes ATP has not yet been identified (58, 60, 61, 63). Several key components such as DCR, helicases, the GEMIN3 (a DEAD box helicase) (53) and its interacting protein GEMIN4, and dsRNA endonucleases have been identified in different organisms for their roles in RISC function.

### RLC and siRNA Loading onto RISC

After cleaving dsRNA to produce siRNAs, DCR may remain in association with siRNAs through interactions with its partner proteins, which change under different circumstance and different species (64, 65). These partner proteins include RDE-4 in C. elegans (66), the RPD-2 in flies (67), and the TRBP in humans (68). DCR interacts with these small RNA binding proteins and helps sorting small RNA duplex to form the RISC-loading complex (RLC). As its name suggests, RLC is responsible partly for loading the siRNA or miRNA into RISC. The transition from double-stranded to single-stranded RNAs during RISC assembly is achieved via RNA-protein and protein-protein interactions. The RLC is partially characterized in humans (69), Drosophila (see below) and C. elegans, but the exact composition is unknown.

siRNA-directed RISC assembly in Drosophila is one of the most experimentally dissected and understood systems in vitro. One of Drosophila’s two DCR homologs (DCR1s), DCR2, is involved in siRNA production and subsequently siRISC assembly, whereas DCR1 is involved in miRNA biogenesis (70), and its dsRNA binding partner is Loquacious (71, 72). Recent data indicate that the two pathways in Drosophila are interchangeable. A subset of miRNA duplexes can be sorted into the siRNA pathway after their biogenesis (64, 65). The sorting of miRNAs into RISC pathway requires the disassociation of miRNA duplex from DCR1 complex and a recruitment of the duplex to DCR2 complex. DCR2 is tightly associated with its partner, R2D2, to form a stable heterodimer (DCR2-R2D2), without which both are unstable in cells (67). The formation of RLC starts from the interaction between the siRNA duplex and the DCR2-R2D2 heterodimer (73, 74). DCR2 is a sensor protein that tends to form thermodynamically more stable siRNA duplex end and places the siRNA duplex onto DCR2-R2D2 in the appropriate orientation (73, 74). This directional positioning of the siRNAs on RLC during RISC assembly referred to as “asymmetric assembly of RISCs” or “asymmetric assembly of RLC” is a key function of RLC. The asymmetric complex assembly is defined mainly by R2D2 and the structures of the small RNA duplexes (73).

Whether siRNA/miRNA are loaded as double stranded, or are first unwound with only single-stranded guide siRNA being loaded into RISC, is still not fully understood. Initially, it was thought that a nonprocessive helicase that separated the two strands of the duplex was involved in RLC (74). In later studies, it has been suggested that siRNAs are loaded initially into RISC as duplexes, and AGO2 cleaves the passenger strand of the siRNA, which facilitates its displacement and leaves the siRNA guide strand bound stably to AGO2 (75–78). A recent report by Robb and Rana (79) demonstrated that human RNA helicase A (hRHA) is a siRNA-loading factor. hRHA interacts in human cells with siRNA, AGO2, TRBP, and DCR, and functions in the RNAi pathway. In RNAi-depleted cells, hRHA was reduced as a consequence of decreased intracellular concentration of active RISC assembled with the guide-strand RNA and AGO2. It is envisaged that other components of RISC are waiting to be identified.

### RISC-Mediated Target mRNA Interaction

The rules for the interactions between RISC and target mRNA are probably different between species and perhaps between tissues in a given species. Two recent papers from the Zamore laboratory (64, 65) provided their important findings about RISC-target interaction in Drosophila S2 cells or embryo lysate. Despite distinct pathways of siRNAs and miRNAs biogenesis, the authors showed that siRNAs and miRNAs are sorted by a common process that involves two types of functionally distinct Argonaute protein complexes: AGO1 and AGO2. The sorting of small RNAs between AGO1 and AGO2 is an active process that depends on the structure of the double-stranded siRNA and the double-stranded assembly intermediates. For example, DCR2-R2D2 acts as a gatekeeper for the assembly of AGO2-RISC, which promotes the incorporation of siRNAs and diverts the integration of miRNAs. An independent mechanism (yet to be identified) acts in parallel to favor assembly of miRNA duplexes into AGO1-RISC and to exclude siRNAs from incorporation into AGO1. Partitioned small RNAs may be loaded onto either AGO1 or AGO2, and this association then determines how the small RNA functions to repress gene expression either via transcript cleavage or via translation inhibition.

Both siRNA- and miRNA-guided RISCs can lead either to the destruction of the target mRNAs or to the suppression of the target mRNA translation. At least three conditions are essential for target mRNA to be cleaved site-specifically by RISC: 1) the sequence complementarity between the guide RNA and the mRNA target must be complete via the nucleotides numbered 2–11 complementary to the 5′ “seed” region of the siRNA, 2) RISC must comprise an AGO protein (e.g., AGO2 in humans) with RNase H-like endonuclease “slice” activity for forming a cleaving RISC, and 3) a scissile phosphodiester bond must exist on the mRNA target accessible to the endonuclease active motif located in the PIWI domain of the AGO protein. In general, target recognition, binding, and cleavage by RISC are determined mainly by the base pairing between the 5′ portion (~10 nt) of a siRNA or miRNA and their target mRNA. When recognizing complementary mRNA, activated RISC forms an effector complex with the target mRNA (80, 81). After cleavage, the target mRNA is degraded. Activated RISC, as a multturnover enzyme, is recycled to cleave additional mRNA targets (56). Only a few AGO proteins, such as AGO2 in humans (61, 82, 83); dAGO1 and dAGO2 in Drosophila (84); RDE1, PPD-1,
ALG1, and ALG2 in C. elegans (66); and Arabidopsis AGO1 and AGO4 (62, 85) have RNase H-like motifs at the PIWI domains that have been identified to have or potentially to have endonuclease (“slicer”) activity. The cleaving activity of the AGO “slicers” is divalent cation (e.g., Mg^{2+}) dependent (61, 86). It catalyzes the hydrolysis of the phosphodiester linkage of mRNA target at specific site and yields 3′ hydroxyl and 5′ phosphate termini on the cleavage products (59, 80). This specific cleavage site is predetermined on RISCs and is located precisely at the phosphodiester bond in the mRNA target between the nucleotides paired to base 10 and 11 of the guide RNA from the 5′ end (57, 61, 88, 87).

Crystal structure studies and biochemical analysis revealed that the phosphate group at the 5′ end of the guide RNA has an impact on the cleavage site fidelity by a RISC “slicer” active site (89–93). The guide RNA normally positions on the AGO protein in a RISC effector with the 5′-end anchored into the phosphate-binding basic pocket in M-id/Piwi-domain, the 3′-end bound with the PAZ-domain, and the middle sequence channelled with the “slicer” active site between the M-id/Piwi-domain and the PAZ-domain. To make the cleavage event occur, the scissile phosphodiester bond at the target must be positioned conducive by the conformational changes induced by the interaction among the AGO protein, the guide small RNA, and the target mRNA.

The structure of target mRNA also has a role in RNAi efficiency. Although the base pairing between a siRNA/miRNA and target site is important, in vitro approaches and bioinformatic prediction demonstrated that the sequence context that surrounds miRNA-binding sites influence the sensitivity to repression by a siRNA/miRNA (93–96). Sequence context could influence miRNA efficacy by mediating the binding of hypothetical cofactor proteins or by affecting the secondary structure of a target site and its accessibility to binding by the siRNA/miRNA. Multiple additional steps in the RNAi pathway including RISC assembly, siRNA strand selection, and target site accessibility might be affected by the target sequence composition. In a recent work by Ameres et al. (97), the authors demonstrate that RISC scans for its target sequences by transiently and nonspecifically binding single-stranded RNA, and it promotes siRNA-target RNA annealing when siRNA meets its target sequence. Their findings also show that cleavage competence of RISC is influenced by the secondary structure of the RNA target and that the 5′ portion of the siRNA determines the stable association of RISC with its target. This study provides insight into the cleavage event in RNAi and could improve the design of potent siRNAs.

RISC-Mediated Translation Inhibition

As stated, translation inhibition mediated by miRNPs/miRISC depends largely on the sequence complementarity of loaded miRNA to target mRNA and the type of AGO protein that is recruited in RISC. Partially complementary siRNAs/miRNAs loaded onto AGO protein that lacks “slicer” activity, is thought to produce translational inhibition. The following are at least four possible mechanisms to explain miRISC-mediated target mRNA translational inhibition: 1) RISC represses protein translation at the stage of translation initiation and postinitiation, 2) small RNA-guided RISC assists relocation and sequestration of the target mRNAs to cytoplasmic processing bodies (P-bodies, stress granules, etc.), 3) RISC guides mRNA decay triggered by rapid degradation of the mRNA, and 4) siRISC may cause immediate and/or fast protein degradation after translation (98).

Small RNA-guided RISC can inhibit translation at the initiation step of protein translation (99, 100) or at postinitiation stages (101–103). Pillai et al. (100) showed that in mammalian cells, let-7 miRNPs inhibits translation at the initiation step. The authors employed two independent approaches to investigate the inhibition process. One involved tethering the human AGO protein to the 3′-UTR of a reporter gene to mimic the mRNA-mediated translational repression in HeLa cells. The second approach was to assay the endogenous let-7 miRNPs for their inhibition of protein translation of a reporter mRNA that contains the let-7 targeting sites. The results demonstrated that the mRNA guided/associated miRNPs blocked the protein translation initiation in an m7G-cap-dependent manner, which suggests that miRNPs interfere with recognition of the cap.

The 5′ end of eukaryotic mRNAs is modified by the addition of a 7-methyl guanosine cap. Eukaryotic initiation factor 4E (eIF4E) binds the m7G cap directly, and this interaction is essential for the initiation of translation of most eukaryotic mRNAs. Regulation of translation initiation is the most common target of translational control, and preventing binding of eIF4E to the m7G cap is a commonly employed cellular strategy to inhibit translation (104). A recent study by Kirikidou et al. (105) indicates that the translation inhibition mediated by miRNP occurred at the initial translation stage. Kirikidou et al. (105) identified a cap-binding-like domain (MC) in the middle domain of AGO2. The MC domain demonstrates m7G-cap binding activity and is required for translational repression but not for assembly with miRNA or endonucleolytic activity. In addition to their finding, the authors propose that AGO represses the initiation of mRNA translation by binding to the m7G cap of mRNA targets, which is likely to preclude the recruitment of eIF4E, intrinsically. Thermann and Hentze (106) have shown evidence that DrosoPhila mir-2 function is mediated by inhibiting m7GppG cap-mediated translation initiation in association with the formation of large RNP complexes. miRISC may also mediate dropoff of translating ribosomes by causing ribosomes to exit prematurely from their associated mRNAs (103). Alternatively, protein translation might not be affected, but the nascent protein is degraded rapidly by the ability of miRNPs to recruit proteolytic enzymes that would degrade nascent polypeptides that emerge from the actively translating ribosomes. The hypothesis that miRISC/miRNPs repress protein translation posttranslationally is also supported by the fact that many mRNAs are associated with actively translating, endogenous mRNAs in polysomes (107). A recent study also found that in human and worm cells alike, depletion of the protein eff6 (an “anti-association” ribosome inhibitory protein known to
to prevent productive assembly of the 80 S ribosome abrogates miRNA-mediated regulation of target protein and mRNA levels (108).

The AGO proteins and the repressed mRNAs are enriched in the cytoplasmic processing bodies (P-bodies, also known as GW-bodies) (100, 109, 110). P-bodies represent discrete cytoplasmic foci, which are enriched in proteins involved in mRNA decay and translational repression. These proteins include deadenylases, decapping enzymes such as DCP1 and DCP2, and 5′-3′ exonucleases (111). On the other hand, P-bodies lack ribosomes and protein translation machinery. In P-bodies, mRNAs can undergo decapping and degradation (112), or they can be stored temporarily. As such, under certain conditions (e.g., stress) (113), the target mRNAs sequenced to and stored in P-bodies can be released and re-recruited by the ribosome to resume the translation.

Chu and Rana (114) showed that in some human cells, miRNA function requires RCK/p54, which is a DEAD-box helicase known to be essential for translational repression. RCK/p54 interacts with AGO1 and AGO2 in vitro and in vivo, facilitates formation of P-bodies, and is a general repressor of translation. Disrupting P-bodies did not affect RCK/p54 interactions with AGO proteins or its function in miRNA-mediated translation repression. Depletion of RCK/p54 disrupted P-bodies and dispersed AGO2 throughout the cytoplasm but did not affect RISC functions significantly. These studies also suggest that translation suppression by miRISC does not require P-body structural integrity and suggests that location of miRISC to P-bodies is the consequence of translation repression.

Downregulation of gene expression by miRNAs is also associated with a substantial degradation of some target mRNAs (115–118). Some mRNAs that are destined to undergo miRNA-mediated decay are first deadenylated (115, 119, 120). The CCR4:NOT1 was identified as deadenylase complex in Drosophila S2 cells (119). In addition, the GW182 protein was identified as a key factor required for miRNA deadenylation, miRNA degradation by miRNAs requires GW182, the CCR4:NOT deadenylase, and the DCPSLP decapping complexes. The following is a model proposed for this posttranscriptional gene suppression: AGO1-containing RISCs bind to mRNA targets by means of base-pairing interactions with miRNAs; AGO may then recruit GW182, which in turn marks the transcripts as targets for decay via a deadenylation and decapping mechanism (115, 119, 121).

Although miRNAs generally have been identified as negative regulators of expression of the target mRNAs in most cases, accumulating evidence now suggests that, in some circumstances, miRNAs are found to enhance protein translation from their target mRNAs. In some cases (e.g., under stress conditions), miRNA associated RISCs simply help the repressed target mRNAs be released from the P-body and recruited by the ribosome to resume protein translation as discussed above (113, 122). This derepression of the target miRNA translation by miRNA needs protein cofactors that are likely induced by stresses and can release translational repression by interacting with the 3′UTR of the target miRNA or by helping the target miRNA to reassociate with polyribosomes. The currently identified protein cofactors include AU-rich-element binding protein HuR (113 122) and apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G (APOBEC3G or A3G) (123). Moreover, synthetic siRNAs that target the promoter regions of the human genes E-cadherin, cyclin kinase inhibitor p21 (p21(CIP1)) (p21), and vascular endothelial growth factor induced the expression of these genes through chromatin remodeling related mechanism that requires the human AGO2 protein (124). These findings suggest more diverse role for miRNAs and siRNAs in the regulation of gene expression than previously appreciated. However, the degree to which these mechanisms are biologically widespread remains to be observed.

RITS Complex-Directed Chromatin Remodeling

Different from cytoplasmic RNA interference pathways is the RNA silencing pathway, which is named RNA induced transcriptional silencing (RITS) and occurs in the nucleus. RITS plays a general role in the construction of centromeres (125, 126). Extensive studies, most notably in the fission yeast Schizosaccharomyces pombe, have established RNA-dependent chromatin remodeling. H3K9 methylation is a central process for RNA-dependent formation of heterochromatin, by which genes are silenced in regions that contain repetitious DNA sequences (127–130).

Three “RNAi” genes are required for this process: RNA-dependent RNA polymerase (RDP1), DCR1, and AGO1. RDP1 is thought to amplify or to produce the initial dsRNA trigger for centromeric silencing in S. pombe. DCR1 is presumed to generate siRNAs from the trigger dsRNA. AGO1 is a component of the RITS complex, which is the effecter complex of siRNA-directed transcriptional silencing. In this pathway, siRNA homologous to the repeated sequences are generated by DCR and associate with an Argonaute via mechanisms not defined fully. These siRNAs/RISCs direct histone-modifying components, such as Ccr4, a H3K9-specific lysine methyltransferase (HKMT), to homologous loci. Ccr4 then methylated lysine 9 of histone H3 (H3K9). Methylated H3K9 recruits effectors like Swi6, Chp2, and other components (e.g., Ta53), which lead eventually to the formation of condensed heterochromatin structures (8, 129, 131–133). Interestingly, synthetic siRNAs that target the promoter regions of several human genes activated the gene expression that needs the human AGO2 and is associated with a loss of lysine-9 methylation on histone 3 at siRNA-target sites (124). It seems that methylation and demethylation of H3K9 to switch off and on genes are via related small RNA pathways in S. pombe and humans.

Summary and Conclusions

RNAi and miRNA pathways involve ancient evolutionary mechanisms for gene expression regulation. These pathways overlap in many aspects with regard to the processing of small regulatory RNAs and to the targeting of mRNAs for post-transcriptional gene expression regulation. Organisms have
evolved different ways of using these mechanisms, and researchers are only beginning to understand the biologic importance of these pathways. If the prior scientific literature is any indication, then many more surprises will develop as we seek to uncover the roles played by small regulatory RNAs and their protein partners.

Acknowledgments

G.T. is supported by the NIH K08 NS050110 and also NIH P30 49008400.


complementary and partially palindromic or partially comple-
mentary single-stranded siRNAs. RNA Biol. 2006;3:82–89.
78. Leschper P, Aamens SL, Kuoeng S, Martinos J. Cleavage of the
79. Robb GB, Rana TM. RNA helicase A interacts with RISC in
85. Baumberger N, Baulcombe DC. Arabidopsis ARGONAUTE1 is
86. Schwarz DS, Tomari Y, Zamore PD. The RNA-induced silenc-
ing complex is a Mg 2
87. Elbashir SM, Harborth J, Pelletier J, Tuschl T. Duplexes of 21-nucleotide RNAs mediate RNA inter-
95. Heale BS, Sofer HS, Bowers C, Rossj jj, siRNA target site
100. Pillar RS, Bhattacharyya SN, Aita CG, Zoller T, Cogott N, Bayuk E, Bertrand E, Filippowicz W. Inhibition of transla-
101. Olson PH, Ambros V. The lin-4 regulatory RNA controls develop-
104. Richter JD, Sonenberg N. Regulation of cap-dependent transla-
106. Thermann R, Hentze MW. Drosophila mirR2 induces pseudo-
107. Maconnay PA, Yu Y, Fisher J, Nilson TW. Evidence that microR-
110. Sen GU, Blau HM. Arbovirus NS5 RNA polymerase in sites of mam-
112. Sheth U, Parker R. Decapping and decay of messenger RNA oc-
113. Bhattacharyya SN, Habermacher M, Martine U, Closs E, Fil-
RNA Interference, Mechanisms and Proteins Involved in


Transcription Factors

KE Hauschild, CD Carlson, LJ Donato, R Moretti and AZ Ansari, Department of Biochemistry and the Genome Center, University of Wisconsin-Madison, Madison, Wisconsin
doi: 10.1002/9780470048672.wecb603

The genome of an organism is deciphered by transcriptional processes to generate RNA and protein molecules that determine cellular fate and perform all cellular functions. The transcriptional machinery itself displays limited specificity and is involved in transcribing all genes in the genome. The exquisite specificity with which gene networks are expressed is mediated by regulatory proteins called transcription factors. Transcription factors bind to gene-specific regulatory sites in the genome and help assemble the complex multi-subunit machinery that transcribes the target gene(s). Spatio-temporal regulation of gene expression permits selective expression of a subset of genes within the genome and thus governs the diversity of cell types and cellular function in response to physiologic signals. The central role of transcription factors in regulation of specific genes and networks is underscored by the fact that their malfunction is linked to the onset of a wide array of diseases, including developmental disorders, cancer, and diabetes. Thus, a major goal at the interface of chemistry, biology, and molecular medicine is the ability to generate synthetic molecules that function as transcription factors. Here we discuss the common properties of transcription factors, summarize the alluring value of targeting gene expression with small molecules, and summarize the current advances toward regulating the expression of desired genes and gene networks with artificial transcription factors.

Introduction

Multicellular organisms consist of a diverse array of cell types, yet this diversity belies the fact that they all contain an identical genomic complement. Specific gene regulatory programs are set into motion during cellular differentiation to confer a unique cellular identity (1, 2). Selective gene expression is directed by gene-specific transcription factors (TFs) that decipher various signals to regulate the expression of relevant target genes and gene regulatory networks. TFs target unique sites in the genome and nucleate the assembly of the transcriptional machinery (Fig. 1). TFs achieve this action by interacting with several protein complexes, including nucleosome-remodeling and -modifying enzymes, the proteasome, coactivators, corepressors, general transcription factors (GTFs), and RNA polymerase II (3-14). Generally, the process of gene transcription by RNA polymerase II can be described by three stages: 1) initiation, which involves the assembly of the transcriptional machinery at the gene promoters, 2) elongation of the nascent transcript, and finally, 3) termination, with release of the full-length transcript that is processed to yield mature mRNA (15, 16). Initially, TFs were thought to act only at the first stage of transcription, but recently they have been implicated in other steps of this intricately choreographed process (1) (Fig. 1a). In addition to TFs that control common cellular functions, different cell types have a unique cell type-specific complement of TFs that regulates the expression of genes and transcriptional networks that are unique to a given cell type.

Most TFs are composed minimally of two key functional protein domains: the regulatory domain (RD), which is used to activate or repress transcription, and the DNA binding domain (DBD), which targets the TFs to specific sites within the genome (1). Different TFs may contain additional domains that confer unique properties. For example, in the case of ligand-sensitive TFs, such as the nuclear receptor family, TFs also contain a ligand binding domain (LBD), which allows the TF to alter its functional state based on the presence or absence of small molecule signals (17-20) (Fig. 1b). The individual protein domains often are functionally self-contained, and these functional
Figure 1: Function of transcription factors. (a) TF binding. Highly compact chromatin, which correlates with hypoacetylation and reduced transcription, limits TF binding. To overcome this inhibition many TFs recruit chromatin-remodeling and -modifying enzymes to promoters. TFs typically recruit the transcriptional machinery to a promoter through protein–protein interactions with the mediator complex. (b) TF transport. Many nuclear receptors are kept in the cytoplasm until they bind their ligand. Ligand binding induces transport into the nucleus. (c) TF modification. Posttranslational modification of TFs before, during, or after transcription initiation frequently regulates TF activity or even DNA binding.

modules have been shown to be interchangeable between TFs (1). This realization led to the development of the widely used two-hybrid assay where DBD and RD are noncovalently assembled by the dimerization of attached proteins or small molecule ligands (21–25). In an additional application of modular design, ligand binding domains of nuclear receptors have been used to confer small molecule control on other transcription factors (19, 26, 27). More recently, engineered DBDs have been assembled in a modular fashion with RD and LBD to generate a class of artificial transcription factors (ATFs) (28–30). Modular assembly that uses synthetic counterparts of DBDs and RDs have also yielded synthetic ATFs that are not limited to peptidic RD and DBD, which allows the use of a wide variety of synthetic compounds as functional modules (29, 31–34). The assembly of ATFs with synthetic modules that harness the cooperative DNA binding properties of eukaryotic TFs has led to a class of molecules that function as protein–DNA dimerizers (34–36).

A significant amount of research has been dedicated to identifying the numerous factors involved in specific gene transcription. However, a greater focus has been on identifying the gene targets of TFs. Understanding the patterns of gene expression that are produced by TFs, individually and subsequently in complex combinatorial assemblies, will provide an increased understanding of how the genome is deciphered to give rise to a functional organism (37, 38). Nearly 6% of the human genome is thought to encode TFs (39, 40). Only a handful of these ~2000 TFs have been characterized to the point that we can comprehensively define their gene targets (41–48). As careful analysis of TFs from other organisms is performed, it is becoming increasingly clear that complex transcriptional circuits regulate a network of genes, which leads to the final cellular response to a given stimulus (49–51). The gene regulatory networks also have led to the realization that key nodes may regulate a defined cascade of gene transcription (2, 52–55). To aid in understanding this major problem of the postgenomic era, it will be essential to have an arsenal of engineered modulators of transcription, including ATFs, to help elucidate these mechanisms fully (29, 33, 34). These factors also will be useful for targeting transcription to alter the gene expression profiles of individual cells or even entire organisms to suit a required need. A striking example of this use has been seen in the emerging field of synthetic biology that unites several scientific disciplines toward the goal of manipulating genetic networks (56) and generating molecular precursors for industrial use (57). The ability to use ATFs to modulate transcription selectively, without the need for direct manipulation of endogenous factors, holds promise to expand their application into several emerging technologies such as metabolic engineering (58).
These finely tuned transcriptional networks can go awry with devastating consequence (59). Although cells possess active systems to prevent and repair transcriptional errors, the disruption of transcription through mutations in TFs has been implicated in a wide variety of diseases (Table and references therein). As such, this situation makes transcriptional regulators a logical target for drug therapies. The vast majority of TF protein assemblies are driven by cooperative interactions that involve several protein–protein interfaces, yet enzymes involved in transcription are a more common target for drug design. Because of the inherent difficulty in disrupting such interfaces, TF–protein–protein interactions are often an underexplored target (1, 60).

Below we discuss examples where chemical approaches have been used to alter transcriptional function by using small molecule inhibitors or activators. We also address strategies to overcome the limitation of targeting protein–protein interfaces. It is clear that the ability to manipulate transcription will be a powerful tool not only to elucidate gene networks or to generate molecule inhibitors or activators. We also address strategies to overcome the limitation of targeting protein–protein interfaces. A glimpse into the incredible potential of targeting TFs with small molecules is provided by the success achieved through regulating nuclear receptors and histone deacetylases. Below we touch on the insights gleaned from studying the nuclear receptor class of TFs and on the impact of targeting histone-modifying enzymes. We then discuss exciting new developments in the creation of ATFs and the importance of additionally targeting transcriptional processes for chemical intervention.

**Ligand-Dependent Transcription Factors: Nuclear Receptors**

Nuclear hormone receptors are ligand-inducible TFs that modulate transcriptional rates in response to their cognate ligands (18). Based on homology, the receptors contain a variable amino terminal region that harbors a ligand-independent activation activity, a highly conserved zinc finger DNA binding domain, and a carboxy-terminal ligand-dependent activation domain (LBD) (Fig. 2a) (61, 62) (19). Ligands for nuclear receptors include a diverse array of lipophilic compounds such as steroid hormones, fatty acids, eicosanoids, bile acids, and oxysterols (18). However, some receptors identified by sequence homology have no known ligands and therefore are referred to as orphan receptors; a subset of these receptors may have no natural ligand. When ligand binding occurs, the LBD undergoes a conformational change in which the carboxy-terminal helix 12 is positioned to form a hydrophobic groove that is necessary for coactivator binding (Fig. 2a) (63). Whereas some receptors are bound to heat shock proteins in the absence of a ligand, the structure of helix 12 in other unliganded and antagonist-bound receptors favors the binding of corepressors (proteins that maintain the surrounding chromatin in a transcriptionally inactive state) to a region of the receptor that overlaps the coactivator binding site. The change in helix 12 position and the resulting switch from co-repressor to coactivator interaction when agonist binding occurs translates into a transcriptional response (17).

**Synthetic ligands to modulate nuclear receptor activity**

The biologic systems that are modulated by nuclear receptors include many metabolic responses, which makes them attractive targets for drug discovery. In fact, currently marketed drugs exist for many receptors including the vitamin D receptor (VDR), glucocorticoid receptor (GR), estrogen receptor (ER), and peroxisome proliferator activated receptors (PPARs) (64–67). Transcriptional activation by nuclear receptors is accomplished through ligand binding to the receptor; therefore, the most common mode of repressing the activity of this TF class is through the use of antagonist ligands that function by binding to the ligand pocket of the receptor and retaining the inactive, corepressor-bound state (17). Some of these ligands can act as agonists or antagonists depending on the target tissue (68). One of the most well-studied examples of non-natural ligands used in the clinical setting is the selective estrogen receptor modulators (SERMs) in the treatment of breast cancer (69). Two examples of SERMs, Tamoxifen and Raloxifene, bind to and act as antagonists of the estrogen receptor in breast tissue (Fig. 2c). Both drugs also have been shown to protect against osteoporosis. However, Tamoxifen treatment has shown an increased risk of uterine cancer because it acts as an agonist in the uterus, whereas Raloxifene does not (68, 70). The tissue-selective function of SERMs seems to be determined by the coregulatory proteins that are expressed in the target tissues. SERMs induce a conformational change in the receptor such that it selectively interacts with a subset of coregulators and correspondingly activates or represses the transcription of the target genes (70).

Similarly, synthetic thyromimetics such as CGS 23245 that target the thyroid hormone receptor beta are designed to lower plasma concentrations of cholesterol in patients with hypothyroidism without incurring the negative cardiac side effects that administering the natural hormone would induce (Fig. 2a) (71). Recently, a new thyromimetic compound, CO23, has been designed specifically to target the TR alpha form in vitro and in vivo and could be used to study the effects of the thyroid hormone receptor modulated transcription in the heart (72). SERMs and thyromimetics are only a small selection of the plethora of synthetic agonists and antagonist ligands that exist for nuclear receptors. Many of these synthetic ligands can target specific isotypes/isofoms of their target receptors and target these receptors in a tissue-specific manner. This class of small molecule regulators of transcription contains the largest and most diverse compilation of molecules that regulate transcription and has yielded numerous therapeutic agents.

As an alternative to using synthetic analogs for disease treatment, the natural ligand for the retinoic acid receptor, all-trans-retinoic acid (ATRA), is used effectively in the treatment of acute promyelocytic leukemia (APL) (73). Chromosomal translocations in APL patients are responsible for the cellular transformation in this disease. The transformation results in a fusion of the retinoic acid receptor alpha (RARA) with another protein (73). Treatment of APL patients with pharmacologic doses of ATRA induces complete remission by restoring normal granulocytic differentiation (74, 75).
### Table 1: Transcription factors in disease

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>Event/Change</th>
<th>Disease Description</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mineralocorticoid receptor</td>
<td>Mutations causing abnormal interactions with the ligand, target DNA sequences, or aberrant nucleocytoplasmic trafficking</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Event</th>
<th>Disease</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STAT3</strong></td>
<td>Constitutively activated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Transcription Factor</th>
<th>Event</th>
<th>Disease</th>
<th>Refs</th>
</tr>
</thead>
</table>
Transcription Factors

Nuclear hormone receptor structure and ligands. (a) The functional domains of nuclear hormone receptors. They act as either homodimers or heterodimers with a ligand binding domain and a DNA binding domain that are separated by a linker sequence. (b) The conformational change in helix 12 when ligand binding occurs (61, 62). All-trans retinoic acid is shown behind helix 12. (c) Examples of synthetic ligands for estrogen receptor (SERMs) and thyroid hormone receptor (Thyromimetics).

Progress in chemically targeting the nuclear receptor class of TFs continues unabated. The development of assays that use time-resolved fluorescence energy transfer and fluorescence polarization techniques has enabled high-throughput screening (HTS) to detect equilibrium binding, nucleic acid hybridization, and enzymatic activity with such targets as kinases, phosphatases, proteases, G-protein coupled receptors, and nuclear receptors (70, 76, 77). These HTS assays also are being used to screen large libraries of natural or synthetic ligands in search of both endogenous ligands of orphan receptors and synthetic ligands for receptors of potential therapeutic use. As SERMs show unanticipated side effects, new criteria for small molecule regulators of nuclear receptors are being formulated. An ideal drug lead would be a ligand that retained the beneficial effects of the activated receptor in the target tissues without incurring negative side effects. With information from ERα and TRβ and the power of HTS assays, synthetic NR ligands can be pursued to maximize potential therapeutic efficacy.

Small Molecule Regulation of the TF Environment: HDAC Inhibitors

Another approach to regulating the activity of transcription factors is to modify the chromatin state and therefore the accessibility of the promoter and TF binding sites (78, 79). Chromatin states are modulated partially by posttranslational modifications of the histones in nucleosomes, which have significant effects on transcription levels (14, 80). These modifications include phosphorylation, methylation, acetylation, and ubiquitination (80, 81). An early indication of the importance of these covalent modifications on transcription emerged from the observation...
that many transcriptional coactivators contain histone acetylase activity, whereas corepressors contain deacetylase activity (Fig. 12) (82). Several clinically important transcription factors, such as ERα, recruit chromatin-modifying enzymes that constitute a critical step in the function of the TF (14, 83). In the active or agonist bound state, ERα interacts with histone acetyl transferases (HATs), and in the inactive or antagonist bound state the TF instead recruits histone deacetylases (HDACs) (84).

When a transcriptional repressor recruits a HDAC to a gene promoter, the HDAC deacetylates proximal histones, a process which is thought to enhance chromosomal condensation and thereby reduce the ability of transcriptional activators to bind the promoter (14). The observation that chromatin condensation correlates with histone deacetylation was possible only by using small molecule HDAC inhibitors (85). The HDAC inhibitor activity of the short chain fatty acid sodium butyrate was identified and led to the suggestion that histone acetylation increases DNA accessibility for TFs and the transcriptional machinery (85). The hydroxamic acid containing natural product Trichostatin A (TSA) was a known and fungal agent for 14 years before it was determined to be a potent inhibitor of HDACs (86). Like many HDAC inhibitors, TSA acts by using a bulky hydrophobic group to block the HDAC pocket and a polar region to interact with the HDAC active site zinc atom (Fig. 3) (87).

HDAC inhibitors that were purified from natural sources and/or synthesized chemically have shown exciting biologic properties such as the induction of cellular differentiation, growth arrest, and apoptosis in several different cancer cell types in culture (88). Additionally, currently three classes of synthetic HDAC inhibitors are undergoing clinical trials (Fig. 3c) (89). These inhibitors include hydroxamic acid containing molecules such as suberoylanilide hydroxamic acid (SAHA), PXD-101, and MS-275 (90-92). The drug MS-275 is a class of benzamide containing HDAC inhibitors that show efficacy in treating some tumors and lymphoma (93). Both the hydroxamic acid and benzamide-containing inhibitors are general HDAC inhibitors thought to affect most of the more than nine different classes of HDACs. In contrast, a sulfonamide anilide-containing inhibitor currently being tested in clinical trials, MGCD-0103, is thought to be isotype specific, which could reduce negative side effects (94). In contrast to all of these synthetically prepared inhibitors, depsipeptide is a natural product purified from Chromobacterium violaceum that functions as an HDAC inhibitor and is proving to be highly effective in clinical trials (95). The search for highly specific HDAC inhibitors with extremely low toxicity and other negative side effects continues.

ERα serves as an example of a TF that is dependent on an HDAC for its activity (84). The reduction in expression of this nuclear receptor is a key step in the carcinogenesis of breast cancer and correlates with poor prognosis. ERα has been shown to bind HDAC1 directly in vitro and in vivo, and the overexpression of HDAC1 in MCF-7 cells leads to a reduction in both ERα protein levels and ERα transcriptional activity. Overexpression of HDAC1 also causes increased cell proliferation of MCF-7 cells (96). Thus, it has been suggested that HDAC1 plays a critical role in breast cancer progression. Consistent with this hypothesis, HDAC inhibitors including TSA, SAHA, and others have been shown to induce a re-expression of ERα and a corresponding increase in the expression of ER target genes in breast cancer cells that lack ERα activity. Moreover, cells treated with HDAC inhibitors have reduced proliferation and, separately, because of the increase in ERα levels, these cells have an increased sensitivity to anticancer drugs (97). These studies have led to clinical trials that use HDAC inhibitors in the treatment of breast cancer and several other cancers (84).

HDAC inhibitors have revealed the complex relationship between chromatin and transcription. One limitation of using HDAC inhibitors as therapeutics is the effects on the deacetylation of other nonhistone cellular targets. HDAC inhibitors are known to have effects beyond chromatin remodeling, including microtubule and aggresome regulation (98). Despite these limitations, applying a chemical strategy has unveiled the importance of chromatin dynamics in the mechanism of transcriptional regulation. Next generation HDAC inhibitors that specifically inhibit targeted HDACs or HDAC-substrate interactions in desired cell types and at specific promoters would increase the use of HDAC inhibitors (99). However, engineering such specificity is nontrivial, and targeting protein–protein interactions presents a significant challenge for a small molecule. Some successes in addressing this challenge are described below and offer lessons in the design of molecules that use coupled equilibria to target protein interfaces.

**TF Interfaces: Targeting Molecular Interactions**

The rewards of targeting NRs and HDACs strongly imply that transcriptional regulation (TFs in particular) is an exciting yet underexplored target for chemical intervention. Despite their seemingly simple architecture, TFs have been fairly recalcitrant to chemical perturbation—this situation may well be because of the weak molecular interfaces between TFs, their myriad and perhaps redundant interactions with components of the transcriptional machinery, or the lack of our understanding of the functional properties of the targeted TF or its closely related isoforms that also may exist in the same cell (60). Even so, examples exist of TF inhibitors that block TFs and prevent their association with cellular partners. These examples include small molecules that bind DNA to prevent TF–DNA interaction and include several TF–protein binding inhibitors. Moreover, new approaches have been developed to block (or enhance) competitively the function of TFs.

**TF–DNA interactions**

Several current cancer therapeutic agents, such as cisplatin, can inhibit cancer by conjugating to DNA and thereby preventing transcription, replication, and repair in actively reproducing cancer cells (100). Other successful therapeutic agents trap topoisomerase II in a covalent complex with genomic DNA, and these nucleoprotein complexes are very physical barriers to RNA and DNA polymerases (101). Unfortunately, such approaches can lead to a wide range of side effects in several
Figure 3  Histone deacetylase inhibitors. (a) The mechanism by which many HDAC inhibitors function. A hydrophobic moiety blocks the active site of the HDAC. A flexible linker extends into the cleft, and a polar group interacts with the active zinc atom. (b) Crystal structure of TSA (left) and SAHA (right) in the active site of a HDAC-like protein HDLP (87). (c) Examples from four major classes of HDAC inhibitors.
normal cell types that exhibit rapid growth. To overcome this significant limitation, new strategies are being developed to inhibit sequence-specific interactions between TFs and their binding sites. In a recent study, the Hypoxia-inducible factor-1α (HIF-1α) was targeted because this TF has been implicated in tumor evolution and metastasis (102). An ELISA-based screen led to small molecules capable of blocking DNA binding by HIF-1α. One screened molecule, echinomycin (a quinoxaline antibiotic, Fig. 4a (103), could disrupt HIF-1α DNA binding by specifically targeting the hypoxia-responsive element (HRE) in vitro and in U251 cells (102).

Polyamides, a class of sequence-specific, minor-groove DNA binders, also have been used to target HIF-1α and many other TF DNA interactions successfully (104–109). This targeting is achieved by steric occlusion, as when a polyamide occupies the site of a minor-groove binding TF (or a TF that has minor groove contacts), by steric interference of a major-groove binding TF with a tripeptide-conjugated polyamide (104), or by allosterically modifying the DNA groove width that inhibits binding in cases in which a TF alters DNA structure when binding (106). In addition, a polyamide conjugated to acridine, a DNA intercalator, was used to prevent DNA binding by a bZIP TF (107). Conjugating the acridine to a sequence-specific polyamide conferred site specificity to the acridine moiety that intercalated between the base steps of the bZIP binding site and prevented TF binding to DNA (107). Such strategies offer a rational approach to disrupting TF-DNA interactions at targeted sites.

A novel approach to disrupting DNA binding by nuclear receptors relies on displacing the zinc atom from its zinc finger DBD. The cysteine thiols of zinc fingers are particularly labile, and recently disulfide benzamide (DIBA), an electrophile, was found to block both ligand-dependent and ligand-independent ERα-mediated cell growth. DIBA disrupts the ERα zinc finger, which releases the chelated Zn(II), and thereby inhibits ERα dimerization and DNA binding, which leads to the loss of both ERα-mediated transcription and estrogen-mediated breast cancer growth (Fig. 4a (110, 111)). The attractive feature of this alternative method for modulating nuclear receptor activity lies in the fact that it targets a completely different region of the TF. Breast cancer patients taking anti-estrogens or prostate cancer patients taking androgen receptor antagonists often develop resistance to the treatment after acquiring a mutation in the LBD (112). A small molecule inhibitor of the DNA binding domain would be a useful treatment to overcome such drug resistance.

DNA-binding small molecules have been isolated to enhance the DNA binding of a mutated C2H2 zinc finger protein (30). A structural variant of zif268, C7, was mutated at two critical residues involved in Zn(II) coordination and subsequently used in a screen to identify small molecules that would complement the mutation. A potent compound (2-(4’quinoline)benzimidazole) was identified that significantly restored the DNA binding and transcriptional capability of the mutated protein (Fig. 4a). This method for developing a small molecule-dependent zinc finger TF may be a valuable tool in the study of this class of TFs.
The transcription factor p53 is a tumor suppressor protein important in regulating genes that control the cell cycle and especially programmed cell death, or apoptosis. The transcription factor p53 is a tumor suppressor protein important in regulating genes that control the cell cycle and especially programmed cell death, or apoptosis. (113, 114). Therefore, it is not surprising that p53 protein mutations are involved in the genesis of several cancers (115-117). Other than the direct protein mutations in p53, several other routes to p53 inactivation in cancer exist that involve MDM2 or hDM2 in humans. In normal functioning cells, MDM2 binds to p53 and maintains the p53 in an inactive state until cellular stress or DNA damage activate cell signals to induce p53 phosphorylation (118). These signal cascades lead to the phosphorylation of p53 at several key residues important for the p53-MDM2 interaction, which disrupts the interface with MDM2 and releases p53 to perform its regulatory function. In several cancers, the expression of MDM2 is highly upregulated and prevents p53 dissociation despite cellular signals that should trigger the p53 release (115). The p53-MDM2 interface is a well-defined deep hydrophobic pocket and has become an important target for small molecule inhibitors of the p53-MDM2 complex (116, 121-124). These compounds fall into distinct categories. The first category consists of the compounds that mimic the naturally occurring interface by using peptidomimetic drugs. Several examples of this exist including those based on β-peptides, such as p53-1 and p53-3, which were identified as inhibitors of the p53-MDM2 interface (122). p53-1 not only disrupted the p53-MDM2 interaction but also, more importantly, did not disrupt other protein interactions tested, which implies strong specificity for the p53-MDM2 complex. Another class of compounds based on a terphenyl backbone was used to mimic the secondary structure of the p53 N-terminal peptide (121, 124). Here, several terphenyl compounds were identified that disrupted the p53-MDM2 interaction and, importantly, also were shown to be membrane-permeable and able to activate p53 when tested in vivo.

The second category comprises small molecule inhibitors, including chaotropes (123), sulfonamides (125), and a promising new class of inhibitors that were identified from a chemical screen, the cis-imidazoline analogs called nutlins (Fig. 4c) (116, 120). The nutlin compounds not only increased p53 activity in vivo but also showed a 90% inhibition of tumor growth when added to a cancer cell line. Other successful examples of using small molecules to disrupt the interactions between TFs and protein targets, such as the CBP/CREB and Myc/Max interfaces, are discussed in recent reviews (126-128). In many cases, small molecules are found for a TF of interest through assays developed with known protein partners. Structures of the interacting protein domains also have been invaluable for drug discovery. For several important TFs, however, the specific cellular partners with which they interact are unknown. In this case, using small molecule screens on solid supports is an extremely useful tool for drug discovery. In one such screen, diversity-oriented synthesis (DOS (129)) was used to identify a small molecule binder of Hap3p, a yeast protein of the Hap2/3/4/5p complex that is involved with aerobic respiration (130). For this assay, 12,396 DOS compounds were attached covalently to a solid surface and subsequently probed with Hap3p to yield a specific small molecule (Fig. 4c). HTS approaches have been limited by the fact that chemical libraries often are designed to interact in deep hydrophobic binding pockets. Future studies need to target the relatively shallow protein–protein interactions that are found at TF interfaces.

Transcription Factor Mimics: ATFs

Although in some cases small molecules can be effective at influencing protein–protein or protein–DNA associations, not all interactions can be so targeted, nor do these molecules allow for gain-of-function effects. To surmount these difficulties, researchers are working to create artificial molecules with functional properties akin to those of natural transcription factors.

Modular design

The construction of ATFs has been aided by the modular nature of natural transcription factors (Fig. 5a, 29). TFs tend to have their DNA recognition and regulatory functions located on separable domains, as discussed earlier (19, 131). ATFs have been constructed mimicking this modularity. Early work on ATFs focused on combining the DNA binding module from one protein with the regulatory domain of another, which results in a TF with the DNA recognition of the first and the regulatory activity of the second (29). In addition to R and DBD, the modularity of nuclear receptors also has been exploited in such a fashion for nearly 2 decades with fusion of the ligand binding domains of an NR conferring ligand dependence on the chimeric protein (26, 132-134).

DNA binding domains

Zinc fingers commonly are used as DBDs in the construction of ATFs (Fig. 5a). The number of sequences that can be recognized by zinc finger proteins has been extended through structure-based engineering, phage-display, and other selection techniques (135, 136). Engineered zinc finger proteins have been expressed as fusions with several regulatory domains, which yield engineered regulators with diverse DNA recognition properties (137-139). Although zinc finger-based ATFs have been invaluable in the creation of the engineered TFs, the use of proteins as exogenously applied ATFs is limited somewhat by the requirement for cell entry, nuclear localization, and potential immune response (140, 141). Although cell uptake peptides and nuclear localization signals can be attached to proteins (142), these proteins usually are introduced into cells with gene therapy techniques. Thus, synthetic “drug-like” molecules may be more useful for therapeutics and in vivo studies.

Most effort thus far on artificial transcription factors has focused on the DNA binding domain (Fig. 5c). Several synthetic DNA binding molecules have been used. One approach is to recognize the double-stranded DNA, either with triplex forming oligonucleotides (143, 144) or with peptide nucleic acids.
Artificial transcription factors. (a) Protein-based ATFs can consist of a DNA binding domain (DBD) that recognizes DNA and an activation domain (AD) that interacts with and recruits the transcriptional machinery. (b) Protein–DNA dimerizers do not interact directly with the transcriptional machinery but can bind and recruit other DNA binding molecules, which in turn interact with the transcriptional machinery. (c) Examples of molecules that act as functional domains used in ATFs. A DNA binding domain is connected, via a linker domain, to either an activation domain or a repression domain.

Although triplex-forming polymers can display activation domains and increase expression (142), they are limited to targeting purine-rich DNA strands. A more versatile DNA binding molecule is the pyrrole-imidazole polyamide. Polyamides attached to natural activation peptides can promote transcription over 30-fold (143–152).

Regulatory domains

Efforts in ATF design have used naturally occurring regulatory modules, which typically include multikilodalton protein domains. Many regulatory modules can be reduced to short peptides that contain most of the regulatory activity (Fig. 5c) (29, 32). In addition to these naturally occurring peptides, genetic screens have found new activating peptides, many that have similarities with natural activator peptides (33, 153). Peptoids with regulatory properties that can function in intact cells also have been identified (154) (Fig. 5c).

Comparatively few nonpeptidic activation domains exist. Some of the first were found in genetic screens, which identify RNA molecules that can promote transcription in yeast (155–157). This finding provided the first evidence that activation modules did not necessarily have to be peptides. More recently, grafting functional groups (similar to those in natural peptidic activation domains) onto an isoxazolidine skeleton resulted in a class of small molecules that are capable of activating...
transcription to the same extent as a much larger peptide (158). Library screening and structure-based design studies have identified larger synthetic compounds that can function as activators (32). For example, screening for inhibitors of the endothelial specific transcription factor (ESK) /Sur2 interaction revealed compounds that bound to Sur2, part of the mediator complex in general transcription (159). More structure-based engineering to improve Sur2 binding produced shrenochor, which, when conjugated to a polypeptide DNA binding molecule, produced a synthetic ATF that activated transcription 3.5-fold in vitro (160, 162). Unfortunately, this synthetic artificial transcription factor is not cell permeable (160).

In comparison with activation, much less has been done with transcriptional repression. The work done thus far has relied exclusively on naturally occurring repression domains and a few synthetic peptides with modest activity (29, 31, 162). The strategy employed to reduce gene expression with small molecules has focused on the inhibition of TF-DNA binding with a competing DNA binding molecule rather than on the active repression of transcription (29). One possible application toward direct repression effects is targeting histone-modifying complexes to gene promoters (29, 163). A clive repression has great potential, as demonstrated by recent reports of the reduction of HIV virus production by zinc finger-repression domain fusions (164, 165).

**Protein–DNA dimers**

Although direct interaction with the transcriptional machinery is a common property of TFs, certain TFs do not interact with the machinery directly but instead interact with other TFs, which leads to the assembly of regulatory complexes (36). Synthetic molecules based on this principle were successful in nucleating the cooperative assembly of an ATF-TF complex on a target DNA binding site. The first generation of such “protein-DNA dimers” uses polyamides to display dipeptides that can interact with a specific TF. This class of bifunctional molecules is sufficient to change the distribution of a TF on DNA (35, 36, 166). This interaction also has been shown to be modulated by the linker that is used to connect the two domains (36). Protein-DNA dimers represent another advance in mimicking the characteristics of natural TFs.

**Ligand-responsive artificial transcription factors**

Although in the past few years much success with creating artificial molecules with functions similar to TFs has been achieved, these synthetic molecules still do not exhibit the versatility of TFs. Most notably, ATFs lag in their ability to be regulated by external signals. The activity of TFs is controlled finely by signals from the external environment and by internal cellular conditions. Some success has been found in the use of nuclear receptor LBDs in conferring small molecule control of zinc-finger-based ATFs (28, 59, 167, 168) and found with small molecule binding RNA aptamers (169). Other techniques have used chemical inducers of dimerization (170) or small molecule-dependent DNA binding elements, such as engineered zinc fingers and the Tet regulatory system (30, 171, 172), for ligand-responsive gene control. In the Tet regulatory system, doxycycline is used to regulate transcription through repression (Tet-OFF) or activation (Tet-On). In the Tet-OFF system, tetracycline transactivator (tTA) binds to the tet promoter (Ptet) and activates a gene of interest. The addition of doxycycline (Dox, a tetracycline derivative) prevents tTA binding to Ptet and effectively inhibits gene expression. For the Tet-On system, a reverse tetracycline transactivator protein (rTA) requires Dox for binding to Ptet therefore, dosing with Dox will activate the transcription of a gene of interest (173). A few reports of ligand-responsive small molecule-based ATFs used for gene control have been made.

**Targeting TF-Modifying Enzymes: Indirect Regulators**

A parallel approach to modulating gene expression is to target enzymes that modify and thereby control the function of TFs (Fig. 3c). Targeting the active sites of enzymes or “druggable” pockets that permit the interaction between enzymes and their substrates has been highly successful in yielding several important chemical tools and even therapeutic agents (59, 96, 174, 175). As more cell signaling pathways are investigated, it is becoming increasingly clear that the signal transduction culminates in the posttranslational modification of a specific set of TFs (176, 177). These modifications define the function of the TF. In some cases, the modification leads to the ability of the TF to translocate to the nucleus and to regulate target genes; in other cases, modifications can lead to proteolysis and degradation of the TF. A wide array of posttranslational modifications of TFs have been reported, including phosphorylation, glycosylation, acetylation, methylation, ubiquitination, sumoylation, and ribosylation (178–184). This wide array provides numerous opportunities to target enzymes that act on TFs and thus indirectly regulate the expression of desired genes and networks. Although promising examples of this approach already exist, the rational targeting of TFs by chemically perturbing the modifying enzyme is not practiced widely yet. An interesting recent example is that of the Myc/yeast enzyme factor 2 (MEF2) that is specifically deacetylated by HDAC3 (185). Like many other HDACs, HDAC3 has biologic targets beyond histones and MEF2 is one of its known substrates (186). In cells, hyperacetylated MEF2 is a potent activator and deacetylation by HDAC3 greatly attenuates MEF2 activation potency. Chemical inhibition of HDAC3 prolongs the acetylated state of MEF2, which leads to increased expression of its target genes (185). As a key regulator of the gene networks that trigger cells to differentiate into skeletal muscle cells, prolonged MEF2 function because of HDAC3 inhibition increases myogenesesis of C2H10 T1/2 cells (185). MEF2 activity also is regulated by other modifications, for example a phosphorylation-dependent switch between an acetylation and a sumoylation on lysine 403 that changes MEF2 from a potent transcriptional activator to a transcriptional repressor (187). The complex interplay of posttranslational modifications on the transcription factor MEF2 shows the rich promise of exogenously regulating transcription by targeting indirect regulators.
Several other examples of indirect targeting of TF activity have been reported, and a prominent example is Nuclear Factor kappa B (NF-κB), a TF that plays a central role in rapid cellular response to stress signals (188). NF-κB regulates the immune response to infection and is involved, therefore, in several inflammation pathways that contribute to diverse ailments such as arthritis, asthma, and cancer (189-191). The activity of NF-κB is modulated by Inhibitor of kappa B (IκB), which binds and inhibits the function of NF-κB (192). The inhibitory protein is in turn regulated by the IκB Kinases (IκK). When cellular signaling occurs, the IκK enzyme is activated and it phosphorylates IκB. This modification marks IκB for degradation and frees NF-κB to translocate to the nucleus and to activate transcription of its target genes (193). The aberrant overactivity of NF-κB in several diseases has prompted a concerted effort to identify antagonists of IκK activity (112). Many IκK inhibitors have been identified and exploited to modulate the activity of NF-κB and ameliorate disease symptoms (111). As we develop a better understanding of the relationship between cellular signals and the enzymes that regulate downstream TF function, we can anticipate the use of indirect targeting approaches to decouple signaling from the undesirable expression of genes and transcriptional circuits.

**Future Directions**

The genomic revolution is dramatically changing our perspective on gene regulatory networks that govern cell fate and function. The rapid pace of discovery at this scientific frontier will provide unprecedented insight into the networks that define cellular differentiation, the response to various internal and external signals, and a broad array of devastating diseases. In addition, the potential for regulating gene networks in non-natural ways for metabolic engineering offers an unimaginable boon to the biotechnology industry (194). Similarly, the ability to regulate their functional role in cellular physiology. TFs provide the promise of positive regulation of any desired gene or gene regulatory network (33). As such, they serve as a complementary tool to negative regulation by RNAi. A recent series of reports suggests that in addition to ATFs, duplex RNA molecules in some contexts may stimulate the expression of target genes (196).

As Niels Bohr famously noted, “It is very difficult to make an accurate prediction, especially about the future.” Although we cannot anticipate the new and creative approaches that will be developed to regulate desired genes and networks, no doubt exists that transcriptional regulatory methods will be developed to explore how information encoded in genomes is deciphered in the genomes of organisms. Once that understanding is at hand, these same molecules, be they RNAi, ATFs, or tools that remain to be invented, will be used to correct the malfunctioning TFs that participate in the onset of disease or to program the cells to perform new metabolic tasks. This exciting challenge faces the chemical biologists, synthetic biologists, genomics, and future practitioners of personalized molecular medicine.

**References**


16

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

- 1870 -

Transcription Factors


Transcription Factors


Aminoacyl-tRNA synthetases (aaRSs) are critical components of the translation machinery for protein synthesis in every living cell. Each aaRS enzyme in this family links a single amino acid covalently to one or more tRNA isoacceptors to form charged tRNAs. Identity elements within the tRNAs serve as molecular determinants or antideterminants that aid in selection by cognate aaRSs. Some aaRSs also have an amino acid editing mechanism to clear their mistakes. The canonical aaRSs and aaRS-like proteins have functionally diverged to perform many other important roles in the cell. Their versatility and adaptability have provided unique opportunities to develop biotechnology tools and to advance medical research.

**Ubiquity of aaRS Structure and Function**

The aaRSs are an ancient family of enzymes that have a lengthy and diverse evolutionary history. For their central role in protein synthesis, the aaRSs generate aminoacylated tRNAs, which are transferred to an elongation factor such as EF-Tu in bacteria for delivery to the ribosome. Some aaRSs have diverged functionally to perform other secondary roles that impact critical cellular activities. In addition, paralogs that bear sequence homology to aaRS domains participate in a wide array of activities in the cell.

**Aminoacylation of tRNAs**

The aaRSs catalyze the covalent attachment of an amino acid to the universal 3′-adenosine at the terminus of tRNA. Aminoacylation consists of a two-step reaction mechanism. First, an amino acid is activated via ATP to form an aminoacyl adenylate intermediate. Second, the amino acid is transferred to the 3′-end of tRNA, releasing AMP (Fig. 1a). Generally, amino acid activation can occur in the absence of tRNAs. However, glutamyl- (GluRS), glutaminyl- (GlnRS), arginyl- (ArgRS), and lysyl- (LysRS-I) tRNA synthetases require tRNA as a cofactor for the activation step.

Representative X-ray crystal structures have been solved for each of the 20 canonical aaRSs (Table 1). This wealth of molecular information has provided a starting point to begin to unravel diverse paradigms that govern substrate recognition, the mechanism of aminoacylation, and alternative functions. In many cases, multiple cocrystal complexes with various substrates, analogs, and inhibitors are available. Fifteen aaRS X-ray crystal structures have been solved in complexes with their cognate tRNAs. Crystalization of individual aaRS domains has also provided insights into structure-function relationships of this diverse family of enzymes.
Figure 1 Aminoacylation reaction. (a) The overall aminoacylation reaction is performed in two steps by the aaRSs. Two modes of amino acid editing can hydrolyze the mischarged tRNA product (posttransfer editing) or misactivated aminoacyl adenylate intermediate (pretransfer editing). (b) The first step of the ATP-dependent aminoacylation reaction activates amino acid to generate an aminoacyl adenylate intermediate. (c) In the second step, the activated amino acid is transferred to the tRNA molecule and AMP is released.

The aaRSs can be divided evenly into two classes based on the architectures of their catalytic domains: the presence of specific consensus sequences and their chemical properties (6). The catalytic core of Class I aaRSs is composed of a Rossmann dinucleotide binding fold that is marked by two signature consensus sequences: KMSKS and HIGH (Fig. 2). Class II aaRSs are typically dimers or tetramers and possess a more unique catalytic core that is made up of seven antiparallel α-helices flanked by β-strands. These enzymes have three consensus motifs [Fig. 2. Motif 1 1GΦX3XΦXΦΦΦ is at the dimer interface, whereas motif 2 [FRXE-H(RXXX)FXXX(D/E)] and motif 3 [GΦGΦGΦ(D/E)(RΦΦΦΦ)] are part of the active site that represents a hydrophobic amino acid]. Each of the two distinct classes of aaRSs aminoacylates a set of amino acids with diverse chemical properties that would be important for protein function. Interestingly, LysRS is represented in both classes (7).

Recently, two new aaRSs that activate O-phosphoserine and pyrrolysine have also been added to the Class II group (8, 9). Both classes of enzymes catalyze the common aminoacylation reaction but via different mechanisms (1). Class I and Class II aaRSs bind ATP in an extended and bent conformation, respectively (Fig. 2). In addition, class I enzymes bind the tRNA acceptor stem from the minor groove side, which orients the 2′-hydroxyl group of the A76 ribose for attachment of the amino acid (Fig. 3). In contrast, Class II aaRSs aminoacylate the 3′-hydroxyl of the terminal adenosine, because the enzyme binds to tRNA via its major groove. Class II phenylalanyl-tRNA synthetase (PheRS), which charges amino acids onto the 2'-hydroxyl group of A76 of tRNAφM, is the only known exception to this rule.

The aaRSs possess diverse polypeptide domains and insertions, in addition to their catalytic core. Likely, these domains evolved to enhance specificity and fidelity and, in some cases, confer other functions (4, 5). One such domain is the C-terminal anticodon-binding domain that is widely varied (1). For example, GluRS and GinRS have highly conserved active sites within their canonical aminoacylation cores but have appended N-terminal anticodon-binding domains that are composed primarily of either α-helices or β-strands, respectively. In addition, common RNA-binding protein domains such as the OB-fold have been incorporated into aaRSs such as LysRS-II. In at least half of the aaRSs, an internal or appended domain confers amino acid editing activity (3).

Alternative chemical activities

By virtue of their long evolutionary history, as well as of their capacity to bind RNA, ATP, and other small molecules such as amino acids, the aaRSs have been recruited to carry out many diverse alternative functions in cells (4) (Fig. 4a). One such function includes capitalizing on its aminoacylation function to proofread RNA processing and maturation in the nucleus that occurs before tRNA export to the cytoplasm for protein synthesis. In addition, aminoacylation of tRNA-like structures, such as the 3′-end of viral genomes in plants and tmRNA in Escherichia coli, is important for viral replication and for ribosome recycling, respectively.

The RNA binding properties of aaRSs, such as leucyl- (LeuRS) and tyrosyl- (TyrRS) tRNA synthetases, have also been exploited to enable excision of self-splicing group I introns in some mitochondria. Others, including threonyl- (ThrRS) and alanyl- (AlaRS) tRNA synthetases from E. coli, are involved in transcriptional and translational regulation through interactions with their mRNA and DNA, respectively. PheRS also binds specifically to DNA, but the function of this property remains unclear. Some of the most diverse roles for aaRSs include cytokine and anti-angiogenic activities for TyrRS and tryptophanyl-tRNA synthetase (TrpRS), respectively.
### Table 1: Classification of aaRSs (6, 8) and their crystal structures

<table>
<thead>
<tr>
<th>Class I Enzyme only</th>
<th>With tRNA Substrate analog domains</th>
<th>With amino acid/Individual domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeuRS</td>
<td>1WKB 1I1E</td>
<td>1WZ2 1Q0, 1H3N 1J25 1J20 2AJG 2AJH 2AJI</td>
</tr>
<tr>
<td>IleRS</td>
<td>1WK8 1ILE</td>
<td>1QU2 1QY2 1FYF 1J25 1J20 1WNY 1WNO 1UE0</td>
</tr>
<tr>
<td>ValRS</td>
<td>1IYW 1Q4T 1ABH 1W0Y 1IQG</td>
<td>1V5 1GAX 1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>CysRS</td>
<td>1L1S 1B52 1IQB</td>
<td>1U0B 1JYW 1QU2 1QY2 1FYF 1J25 1J20 1WNY 1WNO 1UE0</td>
</tr>
<tr>
<td>MetRS</td>
<td>1QQT 1A8H, 1W0Y, 1IQQ</td>
<td>1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>CysRS</td>
<td>1IYW 1Q4T 1ABH 1W0Y 1IQG</td>
<td>1V5 1GAX 1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>GluRS</td>
<td>1G59, 1N77 1N78 2CV0 2CV1 2CV2</td>
<td>1N75 1I09 2CFO 2CZ2</td>
</tr>
<tr>
<td>SepRS</td>
<td>1YFR, 1YFS 1YFT 1YGB 1IQQ</td>
<td>1EVK 1EVL 1FYF 2H0L 2HL1 2HL2 2HK2 2YQ0 2YWT 1JE 1KE 1TKG 1TKY 1SAH 1SAL</td>
</tr>
<tr>
<td>GlyRS</td>
<td>1AT1 1HCT 1N8 24L</td>
<td>1H76 1GGM 1H76 1GGM 1H76 1GGM</td>
</tr>
<tr>
<td>HisRS</td>
<td>1H4V 1H4T 1QEO 1WU7</td>
<td>1X59 1X59 1X59 1X59 1X59 1X59</td>
</tr>
<tr>
<td>LeuRS</td>
<td>1WK8 1I1E</td>
<td>1WZ2 1Q0, 1H3N 1J25 1J20 2AJG 2AJH 2AJI</td>
</tr>
<tr>
<td>IleRS</td>
<td>1WK8 1ILE</td>
<td>1QU2 1QY2 1FYF 1J25 1J20 1WNY 1WNO 1UE0</td>
</tr>
<tr>
<td>ValRS</td>
<td>1IYW 1Q4T 1ABH 1W0Y 1IQG</td>
<td>1V5 1GAX 1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>CysRS</td>
<td>1L1S 1B52 1IQB</td>
<td>1U0B 1JYW 1QU2 1QY2 1FYF 1J25 1J20 1WNY 1WNO 1UE0</td>
</tr>
<tr>
<td>MetRS</td>
<td>1QQT 1A8H, 1W0Y, 1IQQ</td>
<td>1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>CysRS</td>
<td>1IYW 1Q4T 1ABH 1W0Y 1IQG</td>
<td>1V5 1GAX 1PFA 1PFW 1FGS 1GPG 1GPG 1GPG 1GPG</td>
</tr>
<tr>
<td>GluRS</td>
<td>1G59, 1N77 1N78 2CV0 2CV1 2CV2</td>
<td>1N75 1I09 2CFO 2CZ2</td>
</tr>
<tr>
<td>SepRS</td>
<td>1YFR, 1YFS 1YFT 1YGB 1IQQ</td>
<td>1EVK 1EVL 1FYF 2H0L 2HL1 2HL2 2HK2 2YQ0 2YWT 1JE 1KE 1TKG 1TKY 1SAH 1SAL</td>
</tr>
<tr>
<td>GlyRS</td>
<td>1AT1 1HCT 1N8 24L</td>
<td>1H76 1GGM 1H76 1GGM 1H76 1GGM</td>
</tr>
<tr>
<td>HisRS</td>
<td>1H4V 1H4T 1QEO 1WU7</td>
<td>1X59 1X59 1X59 1X59 1X59 1X59</td>
</tr>
</tbody>
</table>
### Chemistry of Aminoacyl-tRNA Synthetases

<table>
<thead>
<tr>
<th>CLASS II</th>
<th>Enzyme only</th>
<th>With tRNA</th>
<th>With amino acid/ substrate analog domains</th>
<th>Individual domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIb</td>
<td>AspRS</td>
<td>1EOV, 1WYD, 1ASZ, 1GW9, 1G51, 1L0W, 1BBA, 1EFQ</td>
<td>1EFW, 1ASY, 1ASZ, 1C0A, 1IL2</td>
<td></td>
</tr>
<tr>
<td>AsnRS</td>
<td>1X56</td>
<td>1B70, 1B7Y, 1J/C, 2ALY, 2AMC, 2FY5, 2AKW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LysRS-II*</td>
<td>1BBW</td>
<td>1BBW, 1E1O, 1E1T, 1E22, 1E24, 1LYL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIc</td>
<td>PheRS</td>
<td>1PYS, 1EIY, 1B70, 1B7Y, 1J/C, 2ALY, 2AMC, 2FY5, 2AKW</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*These aaRSs have a hydrolitic editing pathway to clear misactivated and mischarged amino acids.

### aAR5-like proteins

Proteins with sequence homology to aAR5s perform diverse cellular functions (5) (Fig. 4b). Some of these proteins resulted from aAR5 gene duplications and are called paralogs. Other paralogs are composed of just one domain of an aAR5 that has multiple domains.

Paralog examples include E. coli YadB, which resembles GluRS. YadB attaches glutamate to a queuosine base yielding an essential tRNAAsp anticodon modification. In addition, two methionyl-tRNA synthetase (MetRS)-like proteins called Arc1p and Trib11 bind to tRNA. Although the role of Trib11 is not understood, Arc1p aids in delivery of tRNAs to their cognate aAR5s subsequent to nuclear export.

Several aAR5s possess amino acid editing domains, which hydrolytically clear mistakes in cis (3) (Table 1). Interestingly however, some archaeal and bacterial synthetases rely on paralogs of editing domains that hydrolyze mischarged products in trans (10, 11). These products include the YbaK and ProX (or PrdX), which edit Ala-tRNAPro, and AlaX that hydrolyzes mischarged Ser-tRNAAla and Gly-tRNAAla.

### Chemistry of Aminoacylation

#### Recognition of tRNAs by aAR5s

One of the most significant features of the aminoacylation reaction is the specific recognition of a set of tRNA isoacceptors by their corresponding aAR5 (2). A correct tRNA-aAR5 pairing relies on molecular “identity elements” that are composed of individual nucleotides, modifications, base pairs, or structural motifs. Positive identity elements (or determinants) enhance the selection of a tRNA by its cognate aAR5, whereas negative identity elements (or antideterminants) prevent the formation of incorrect tRNA-aAR5 pairs. These identity elements can be classified as either major or minor elements based on the level of their impact on recognition.

Most identity elements are present at the two distal ends of the molecule: the acceptor stem and the triplet anticodon (Fig. 5). In the acceptor stem, the unpaired discriminator base N73 is a crucial recognition factor for most aAR5s. Also, the first few acceptor stem base pairs are unique to many tRNA-aAR5 pairs. These identity elements can be classified as either major or minor elements based on the level of their impact on recognition. Binding of uncharged tRNA to a HisRS-like domain of the GCN2 kinase protein in Saccharomyces cerevisiae under starvation and stress conditions elicits a cascade of molecular events to upregulate genes for amino acid and nucleotide biosynthesis. The accessory subunit Polγ of mtDNA polymerase γ, which shares structural homology to both the catalytic and the anticodon binding domains of glycyl-tRNA synthetase (GlyRS), increases the processivity of the catalytic subunit of DNA-Pol and was proposed to serve as a primer recognition factor.

#### Recognition of tRNAs by aAR5s

One of the most significant features of the aminoacylation reaction is the specific recognition of a set of tRNA isoacceptors by their corresponding aAR5 (2). A correct tRNA-aAR5 pairing relies on molecular “identity elements” that are composed of individual nucleotides, modifications, base pairs, or structural motifs. Positive identity elements (or determinants) enhance the selection of a tRNA by its cognate aAR5, whereas negative identity elements (or antideterminants) prevent the formation of incorrect tRNA-aAR5 pairs. These identity elements can be classified as either major or minor elements based on the level of their impact on recognition.

Most identity elements are present at the two distal ends of the molecule: the acceptor stem and the triplet anticodon (Fig. 5). In the acceptor stem, the unpaired discriminator base N73 is a crucial recognition factor for most aAR5s. Also, the first few acceptor stem base pairs often serve as determinants for many tRNAs. The structural domains of aAR5s play specific roles in the recognition of tRNAs (12). In general, the more conserved catalytic domain binds the acceptor stem during aminoacylation. A is indicated, diverged domains in most aAR5s interact directly
Chemistry of Aminoacyl-tRNA Synthetases

Figure 2: Mode of ATP binding to aaRSs. (a) Active site of GluRS enzyme bound to ATP (PDB 1N75). The consensus sequences KISKR (KMSKS) and HVGT (HIGH) are highlighted on the protein. (b) Active site of GlyRS enzyme bound to ATP (PDB 1B76). The consensus motifs 1, 2, and 3 are highlighted on the protein. Class I and Class II aaRSs bind ATP in an extended and bent conformation, respectively. ATP is shown as dark spheres.

with the tRNA anticodons to enable discrimination. The only exceptions are LeuRS, SerRS, and AlaRS, which do not require anticodon interactions for aminoacylation. SerRS and LeuRS interact, respectively, with up to four and six tRNA isoacceptors that have highly varied anticodons. In some aaRSs, accessory domains enhance recognition.

Unique architectural features of tRNAs also can serve as identity elements (2). For example, the long variable loop of tRNA<sub>Ser</sub> interacts specifically with SerRS. SerRS and LeuRS interact, respectively, with up to four and six tRNA isoacceptors that have highly varied anticodons. In some aaRSs, accessory domains enhance recognition.

Antideterminants are defined as nucleotides that block mis-aminoacylation of the tRNA by a noncognate aaRS (2). A few examples of unmodified nucleotides or base pairs that act as antideterminants include A73 in human tRNA<sub>Leu</sub> that hinders aminoacylation by SerRS, U34 in yeast tRNA<sub>Glu</sub> that interferes with MetRS binding, and a G3:U70 base pair in yeast tRNA<sub>Ala</sub> that blocks ThrRS. In addition, lysidine 34 and m1G37 are located in the anticodon structures of E. coli tRNA<sub>Ile</sub> and yeast tRNA<sub>Asp</sub> and serve as antideterminants against MetRS and ArgRS enzymes, respectively. Interestingly, tRNAs that are charged by Class I aaRSs possess antideterminants against Class II aaRSs and vice versa (2).

Reaction mechanism

The overall two-step aminoacylation reaction relies on mechanistically distinct features of the Class I and Class II enzymes (1). In the Rossmann fold of Class I aaRSs, ATP binding is stabilized by interactions with the conserved KMSKS and HIGH consensus sequences. The β- and γ-phosphates interact with Mg<sup>2+</sup>. The α-NH<sub>3</sub> group of the bound amino acid

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

- 1877 -
Chemistry of Aminoacyl-tRNA Synthetases

Functions of aaRSs

- DNA Binding
- Aminoacylation of tRNA
- mRNA Aminoacylation
- Translation
- Proofreading
- Export
- Cytokine-like Activity
- Anti-Angiogenesis
- Mitochondrial RNA Splicing
- Transcriptional Regulation
- Aminoacylation of tRNA-like 3’ End of Viral Genomes

Cyclophilin

AaRSs-like proteins

- Arc1p, Trbp111 (RNA-binding)
- BirA (Biotin Synthetase/Repressor)
- PoxA/GenX, Pyruvate oxidase
- AlaX, ProX, YbaK (Trans-editing of mischarged tRNA)
- YadB (tRNA modification)
- GCN2 (Protein Kinase)
- DNA Pol γ (DNA Replication)
- AlaRS, ProRS
- SerRS
- MetRS
- AspRS
- AsnA (Asparagine Biosynthesis)
- HisZ (Histidine Biosynthesis)
- Arc1p, Trbp111
- AlaRS, PheRS, TyrRS, TrpRS, LeuRS, AspRS

Figure 4  Cellular roles of aaRSs and aaRS-like proteins. (a) Alternative functions of aaRSs. The canonical aaRSs perform diverse functions in the cell that are distinct from their primary role of aminoacylation in protein synthesis. (b) AaRS-like proteins. Proteins that either resemble aaRSs or are paralogs of aaRSs are widespread in nature and carry out a variety of activities.

is stabilized via hydrogen bonds with a conserved aspartate. Amino acid activation occurs by in-line nucleophilic displacement, where the ω-carboxylate oxygen of the amino acid attacks the ω-phosphorous atom of ATP (Fig. 1b). The flexible KMSKS peptide loop forms hydrogen bonds between the conserved lysine and serine and the pyrophosphate moiety of ATP during the transition state. Cleavage of the phosphoanhydride linkage between the ω- and β-phosphates of ATP releases PPi. In the second step, the 2’ ribose hydroxyl of the tRNA A76 attacks nucleophilically the carbonyl carbon of the aminoacyl adenylate intermediate to cleave the mixed anhydride. When amino acid is transferred to the tRNA molecule, AMP is released.

In Class II aaRSs, the bound ATP molecule interacts with motifs 2 and 3 (Fig. 2). The ω- and γ-phosphates bend toward the adenine ring and are stabilized by three Mg2+ ions. The ribose of the ATP molecule adopts a 3’-endo conformation, as opposed to the 2’-endo conformation in Class I aaRSs. The amino acid side chain tends to bind by a lock-and-key mechanism, whereas its backbone interacts via an induced fit. Similar to Class I aaRSs, activation proceeds by an in-line nucleophilic displacement mechanism. The transition state is stabilized by interactions between the ω- and γ-phosphates and arginine residues in motifs 2 and 3, respectively. The 3’-hydroxyl of the A76 ribose nucleophilically attacks the carbonyl carbon of the aminoacyl adenylate intermediate, forming aminoacyl-tRNA and AMP.

Amino acid editing

Nine aaRSs that include representatives of both classes possess a hydrolytic editing mechanism to clear misactivated amino acids (3) (Table 1). Hydrolysis of misactivated amino acids can occur either before or after transfer to tRNA (Fig. 3a). Pretransfer editing hydrolyzes adenylate intermediates but can be RNA-dependent. Posttransfer editing cleaves mischarged tRNA. In general, aaRSs that edit seem to employ both pathways, although one may predominate.
Figure 5  Protein–RNA interactions of aaRSs. The cloverleaf secondary structure of tRNA$^{\text{Leu}}$ folds into an L-shaped tertiary molecule. The tRNA can bind in an aminoaacylation complex, where the 3'-end is located in the canonical Class I or Class II core as shown in the upper right for the *P. horikoshii* LeuRS-tRNALeu aminoacylation complex. In aaRSs that edit, a second complex can be formed, where the 3'-end interacts with a separate domain such as the connective polypeptide insertion (CP1) that contains a hydrolytic active site as shown in the lower right for the *T. thermophilus* LeuRS-tRNALeu editing complex. (Table 1; PDB files 1WZ2 and 2BYT).
O-phosphoseryl-tRNA synthetase (SepRS) acylates tRNA synthetase (CysRS). Interestingly, a Class II enzyme called SepSecS (9) or Sec synthase (SecS) (13), converts O-phosphoseryl-tRNACys, which is then converted to Sec-tRNACys by Sec synthase (SelA). A pyridoxal phosphate-dependent enzyme called Sep-tRNA:Sec-tRNA synthetase (SepSecS) (9) or Sec synthase (SecS) (13), converts Sep-tRNACys to Sec-tRNACys. Finally, a pyridoxal phosphate-dependent enzyme called Sep-tRNACys synthetase (SepSecS) (9) or Sec synthase (SecS) (13), converts Sep-tRNACys to Sec-tRNACys.

In vivo, SepSecS has been achieved using E. coli strains that are auxotrophic for certain amino acids. Initially, cells are grown in the presence of a formyl group is added to the charged methionine via its amino moiety by the methionyl-tRNA formyltransferase enzyme, which uses N10-formyl tetrahydrofolate as the formyl donor. This formylated tRNA will then be expected to be modified to Pyr-tRNA by a mechanism that remains undefined (1). Formylation of methionylated tRNA allows differentiation of the AUG start codon from internal AUG codons (14). MetRS is a Class II enzyme that recognizes specific sec-tRNA sequences called selenocysteine insertion elements (SECISs), which are located 3' to a UGA codon on a stalled ribosome-bound mRNA. It is also responsible for delivering Sec-tRNA to the A site of the ribosome (9).

In archaea and eukaryotes, incorporation of Sec into polypeptide synthesis occurs by several mechanisms. In E. coli, SecRS produces the mischarged Ser-tRNA, which is then converted to Sec-tRNA by Sec synthase (SelA). A pyridoxal phosphate-dependent enzyme called Sep-tRNACys synthetase (SepSecS) (9) or Sec synthase (SecS) (13), converts Sep-tRNACys to Sec-tRNACys.

The appropriate aaRS can also be overexpressed to boost charged tRNA production of analogs that are poorly activated. Editing-defective aaRSs can also be capitalized on for global incorporation of amino acid analogs into proteins. Nonstandard amino acids can be either incorporated globally at multiple sites within a protein or inserted at specific locations (1, 12). Global misincorporation of nonstandard amino acids can produce protein analogs with altered physical properties that confer, for example, varied tensile strengths and elasticities (16). These unique biomaterials can be used in many medical applications, such as altering properties associated with cell adhesion. In other applications, routine replacement of methionine by selenomethionine aids in X-ray crystal structure determination. Global incorporation of amino acid analogs into proteins has been achieved in vivo using E. coli strains that are auxotrophic for certain amino acids. Initially, cells are grown in the presence of the analog to reach sufficient biomass. The standard amino acid is then replaced by an analog in the media, followed by induction of a target protein for expression. The appropriate aaRS can also be overexpressed to boost charged tRNA production of analogs that are poorly activated.
### Table 2: Indirect pathways of aminoacylation (9, 12)

<table>
<thead>
<tr>
<th>Product</th>
<th>Mischarged intermediate(s)</th>
<th>Organism</th>
<th>Enzyme steps</th>
<th>Chemical reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gln-tRNA(^{Gln})</td>
<td>Glu-tRNA(^{Gln})</td>
<td>Bacteria</td>
<td>1. ND-GluRS(^{†}) 2. Asp/GluAdT (GatCAB) or GluAdT (GatDE)</td>
<td><img src="image1.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Archaea/Eukaryotic organelles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn-tRNA(^{Asn})</td>
<td>Asp-tRNA(^{Asn})</td>
<td>Bacteria</td>
<td>1. ND-AspRS(^{†}) 2. Asp/GluAdT (GatCAB)</td>
<td><img src="image2.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Archaea/Bacteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sec-tRNA(^{Sec})</td>
<td>Ser-tRNA(^{Sec})</td>
<td>Bacteria</td>
<td>1. SerRS 2. SerA</td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Archaea/Eukaryotes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sec-tRNA(^{Sec})</td>
<td>Ser-tRNA(^{Sec})</td>
<td>Archaea/Eukaryotes</td>
<td>1. SerRS 2. PSTK 3. SepSecS</td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td>Cys-tRNA(^{Cys})</td>
<td>Sep-tRNA(^{Cys})</td>
<td>Archaea</td>
<td>1. SepRS 2. SepCysS 3. SepSecS</td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>Met-tRNA(^{Met})</td>
<td>Met-tRNA(^{Met})</td>
<td>Bacteria</td>
<td>1. MetRS 2. Met-tRNA transfor-mylase</td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>&quot;Pyl-tRNA(^{Pyl})</td>
<td>Lys-tRNA(^{Pyl})</td>
<td>Archaea</td>
<td>1. LysRS-I-LysRS-II complex 2. unknown</td>
<td><img src="image7.png" alt="Image" /></td>
</tr>
</tbody>
</table>

\(^{†}\) ND = nondiscriminating.

*An alternative direct pathway wherein pyrrolysine is charged to tRNAPyl by pyrrolysyl-tRNA synthetase (PylRS) has also been described (9).
proteins, can be cross-linked readily to introduce diverse chemical properties to the protein. Other novel amino acids that have been incorporated successfully into proteins via editing defective aaRSs include α-aminoarbutic acid by ValRS, norfloxacin and norvaline by IleRS and LeuRS, and unsaturated amino acids such as allylglycine, homoallylglycine, homopropargylglycine, and 2-butylnorvaline by LeuRS.

Site-specific incorporation of novel amino acids in proteins (orthogonal aaRSs)

Site-specific insertion of novel amino acids in proteins originally was accomplished in vitro by generating mischarged suppressor tRNAs, which have anticodons that pair with amber “stop” codons. The tRNAs were aminoacylated chemically with natural and novel amino acids and were then incorporated at specific stop codons using in vitro translation systems.

Orthogonal aaRS:tRNA pairs were developed to incorporate nonstandard amino acids at specific protein sites in vivo (15). Within an organism, the orthogonal aaRS cannot recognize any endogenous tRNA molecules. Likewise, the orthogonal tRNA cannot be aminoacylated by any genome-encoded aaRSs.

Two general strategies have been employed to engineer successfully orthogonal pairs. First, aaRS and tRNA pairs from distant organisms; for example, S. cerevisiae were used in an E. coli system. Second, these aaRSs and tRNA molecules were altered via rational and/or random mutagenesis, typically using in vivo selection systems, until they evolved to orthogonality.

Orthogonal pairs have been selected that can incorporate genetically more than 30 unnatural amino acids in E. coli, yeast, or mammalian cells. These amino acids include photosensitized amino acids, glycosylated amino acids, and amino acids with chemically reactive groups that are used to modify selectively native proteins. This group also includes amino acids with novel properties that could be used as probes in protein structure-function analysis in several techniques, such as FRET, X-ray crystallography, NMR, and IR spectroscopies. Perhaps the best characterized is a Methanococcus janaschii TyrRS:tRNATyrCUA orthogonal pair that is expressed in E. coli and incorporates the synthetic amino acid D-methyl-L-tyrosine. Other examples include the yeast AaspRS:tRNAλmerC, and yeast TyrRS/E. coli RNA84C orthogonal pair that function in E. coli, the LeuRS:tRNATyrCUA pair that functions in yeast, E. coli TyrRS:tRNATyrCUA and GlnRS:tRNAfMetCUA pairs, the Bacillus subtilis TrpRS:tRNATrpCUA, and the E. coli TyrRS/Bacillus steaetheromophilus RNA5pCUA pairs, all of which function in eukaryotic cells.

AARSs and Medicinal Therapeutics

Each member of the aaRS family is central to the viability of every living cell. Not only are they essential to protein synthesis, but also they are critical to many other important and diverse cellular activities (4). Species-specific inhibition of a single aaRS has been a proven drug target for antimicrobial development. Human aaRSs have long been associated with autoimmunity. More recent reports have linked aaRS defects to neurologic problems. Finally, the discovery of an impressive and growing list of aaRS alternative activities, which include roles in apoptosis, angiogenesis, and immunity provide new pathways to novel medicinal therapeutics.

aaRSs as antibiotic targets

Enzymes in the aaRS family are a promising target for the development of novel antibiotics (17). Selective inhibition of just one essential aaRS would be lethal to the pathogen. The best example is mupirocin, a commercially marketed IleRS inhibitor. Mupirocin, also known as pseudomonic acid, originally was isolated from Pseudomonas fluorescens and is used as a topical antibiotic against gram-positive bacteria, particularly Staphylococcus aureus. It binds directly to the first lysine of the conserved KMSKS sequence in the aminoacylation active site (18).

aaRS-linked autoimmune diseases

Antisynthetase syndrome is a condition that involves the production of antibodies that bind to and inhibit aaRSs (1). Several human aaRS antibodies have been identified as markers for this autoimmune disease. These antibodies are generally found in one third of all patients with polymyositis and dermatomyositis, which encompasses chronic inflammatory muscle and skin disorders. These patients exhibit muscle inflammation, interstitial lung disease and arthritis, and mortality related to cardiopulmonary complications. To date, six human cytoplasmic aaRSs, including HisRS, ThrRS, AlaRS, IleRS, GlyRS, and AaspRS, have been identified as autoantigens and have been paired, respectively, with Jo-1, PL-7, PL-12, OJ, EJ, and KS autoantibodies. In any given condition, the patient’s serum has only one of the above autoantibodies. Although the origin of antisynthetase antibodies is not known, it is clear that elucidating their functions in autoimmune diseases may provide avenues to new therapeutics (1).

aaRS genetically linked neurologic diseases

Mutations in aaRS genes have been linked to progressive neurodegeneration and neuropathies. In one example, a single amino acid substitution (A734E) in the amino acid editing domain of mouse AlaRS enzyme results in the accumulation of misfolded proteins in Purkinje cells within the adult cerebellum (19). This modest editing defect triggers the unfolded protein response and cell death in terminally differentiated neurons. Charcot-Marie-Tooth neuropathies can be caused by single mutations in genes encoding either TyrRS (20) or GlyRS (21, 22). The TyrRS-based mutation primarily affects aminoacylation and has been proposed to impair protein synthesis. In contrast, the molecular basis of pathogenesis remains unclear for GlyRS mutants, because they seem to retain aminoacylation activity (22).
aaRS-linked immunity and apoptosis

Under apoptotic conditions in human cells, TyrRS is secreted and then cleaved by the extracellular enzyme leukocyte elastase into two fragments. The two fragments have distinct cytotoxic activities (22). The C-terminal fragment is homologous to the endothelial monocyte-activating polypeptide-II cytochrome. It stimulates the production of myeloperoxidase, TNF-α, a tissue factor, and migration of polymorphonuclear and human umbilical vein endothelial cells. It also has leukocyte and monocyte chemotactic activities, which are dependent on a heptad loop located in the first β-strand of the β-barrel structure.

The N-terminal fragment (mini-TyrRS) is a proangiogenic factor. An ELR tripeptide found in CXC-chemokines with proangiogenic activity is located in the Rossmann fold of mini-TyrRS and is responsible for its angiogenic function. mini-TyrRS also has interleukin-8-like activity, which includes stimulation of polymorphonuclear cell migration and binding to the IL-8 receptor.

aaRS anti-angiogenesis activity

Human TrpRS is involved in the inhibition of angiogenesis; the formation of capillaries from blood vessels (23). It is activated either by alternative splicing or by proteolysis. An alternative splicing results in six different mRNA species, one of which codes for a protein called mini-TrpRS. Mini-TrpRS is regulated by INF-γ and blocks endothelial growth factor (EGF)-induced angiogenesis by inhibiting apoptosis of endothelial cells. The proangiogenic activity of mini-TyrRS and the antiangiogenic activity of mini-TrpRS suggest that the two aaRSs might play cooperative, but opposing, roles in the regulation of angiogenesis.

Proteolysis by leukocyte elastase generates two N-terminally truncated fragments called T1-TrpRS and T2-TrpRS. The larger T1-TrpRS is similar in size to mini-TrpRS and has comparable angiostatic activity. T2-TrpRS, the smaller fragment, is an inhibitor of retinal and vascular EGF-induced angiogenesis and thus has tremendous potential as a novel anticancer agent. T2-TrpRS is homologous to the IL-8 receptor.

References

Further Reading


Champagne KS, Sirlar M, Lambare Y, Double S, Franklyn C.S., Activation of the hetero-octameric ATP phosphoribosyl transferase through subunit interface rearrangement by a RNA synthetase paral-


O’Donoghue P, Sethi A, Whose CR, Luthey-Schulten ZA. The evolu-


Sauerwald A, Zhu W, Major TA, Roy H, Palourea S, John D, Whit-


U. F. 1004
Translational Recoding

Edith M. Osborne and Christopher J. Noren, New England Biolabs, Ipswich, Massachusetts
doi: 10.1002/9780470048672.wecb612

The limitation of a triplet code with only four bases from which to choose and the size limitation of viral genomes has placed evolutionary pressure on translational coding. Organisms and viruses have developed a variety of methods of translational recoding, including codon reassignment and frameshifting, temporarily altering the canonical reading of mRNA to meet their needs. A translational recoding event can be caused by alterations in the mRNA and/or the tRNA and may even necessitate specific ribosomal elongation factors. Scientists also desire to alter the genetic code by manipulation of natural recoding events and by exploitation of artificial genetic code expansion.

Introduction

With the availability of 64 three-letter combinations of the genetic code, 61 of which code for the 20 canonical amino acids and 3 of which are stop codons, both nature and humans have used a variety of mechanisms to expand the genetic code to incorporate additional amino acids. The genetic code is not entirely universal. A small subset of the genetic code has been reassigned in certain organelles and a limited number of species. In certain cases, the genetic code is expanded only when pertaining to a specific mRNA. These recoding events often compete with standard decoding. Thus, recoding events can play a role in regulation or increase of genetic diversity (1). Programmed frameshifting, translational bypassing, and codon redefinition are three types of translational recoding. Both frameshifting and bypassing occur by slippage of the mRNA in the ribosome during translation. Bypassing involves skipping a block of nucleotides during decoding and is frame independent, for example, it may or may not cause a change of frame. Codon redefinition temporarily assigns a new meaning for a codon. Natural mechanisms of codon redefinition to incorporate additional amino acids include N-formylmethionine, selenocysteine, and pyrrolysine (Table 1 and Fig. 1a). Scientists have incorporated a vast array of noncanonical amino acids through the means of artificial genetic code expansion.

Frameshifting and Bypassing

Several mRNAs have been shown to cause the ribosome to shift frames, either +1 (3') or -1 (5'), with frequencies of shift ranging from a few percent up to 50% in cases of programmed frameshifting (2). Although frameshifting occurs in a wide variety of prokaryotes and eukaryotes (1), many frameshifting events have been found in viruses, transposons, and insertional elements (3, 4). This may occur because recoding provides a means for viruses and mobile genetic elements to condense protein coding and/or regulatory information in their compact genomes (5).

Several factors enhance ribosome slippage. Slippery sequences, such as the heptameric sequence A–AAA–AAU, favor RNA movement or misalignment and thus increase the probability of frameshifting (6). The presence of rare sense codons, codons under aminoacyl–tRNA limitation (amino acid starvation), or stop codons can increase frameshifting, which suggests that ribosomal pausing contributes to initiation of peptidyl–tRNA slippage (7). The combined stimulatory signals encoded in mRNAs that undergo programmed ribosomal frameshifting can greatly increase the probability of tRNA slippage at a specific shift site, with up to 50% of the ribosomes changing frame in some cases. Programmed ribosomal frameshifting sites include a slippery sequence in the mRNA and an mRNA feature to enhance the slippage. In addition to pauses caused by a low-abundance tRNA or a stop codon, secondary structures, including stem-loops and pseudoknots, may also enhance programmed frameshifting events (4, 8).

Crystallographic, molecular, biochemical, and genetic studies have provided insight into the mechanics of the highly efficient levels of -1 ribosomal frameshifting caused by mRNAs pseudoknots (4). During the RNA accommodation step on the ribosome, the anticodon loop of the aminocyl–tRNA moves 9 Å and normally pulls the downstream mRNA a similar distance. A downstream mRNA pseudoknot wedged into the entrance of the ribosomal mRNA tunnel resists this translocation. The tension in the mRNA between the A-site codon and the mRNA pseudoknot can be relieved by unwinding the pseudoknot, which allows the downstream mRNA to move forward and be read in frame, or by slippage of the mRNA backward by one base, which causes +1 frameshift.
Frameshift into the residues at the 3' end of T7 gene particularly shift-prone codon (12). A short Shine–Delgarno-like programmed frameshifting site. The presence of a C 3' of the low RF2 concentration increases frameshifting efficiency and (10). High concentration of RF2 favors termination, whereas and production of full-length RF2 is an autoregulatory process concentration of RF2 in the cell, the frameshifting efficiency of termination at the UGA stop codon directly depends on the base is the first base of the stop codon. Only one leucine is in the 3' direction toward the overlapping UUU where the last sequence located three nucleotides 5' of the shift site, and which can pair with the 16S rRNA of the translating ribosome, may realign by creating tension in the mRNA.

Several features of the RF2 mRNA aid in creating this programmed frameshifting site. The presence of a C 3' of the UGA makes a poor termination context (11), and CUU is a particularly shift-prone codon (12). A short Shine-Delgarno-like sequence located three nucleotides 5' of the shift site, and which can pair with the 16S rRNA of the translating ribosome, may create a pause to allow more time for slippage, or may promote reorientation by creating tension in the mRNA.

Some mRNA sequences have been found that cause low-to-moderate frameshifts without the aid of nearby stimulator sequences. These sequences include the four consecutive U residues at the 3' end of T7 gene 10, which permit a -10% frameshift into the -1 frame to yield protein 10B (14) and the sequence CCC UGA, which permits a -3% +1 frameshift (15). Strings of four or more G or U residues that precede a stop codon have been shown to facilitate a low level of frameshifts (13). The UUU phenylalanine codon followed by a pyrimidine also causes low-to-moderate +1 frameshifts (16, 17). Sequences prone to frameshifting have been identified from phage-display isolates that were expressed, although they lacked an identifiable open reading frame (18). More studies of these mRNA sequences revealed that the sequence CCC CGA is a weak (1–2% efficiency) translational frameshift site in E. coli (19). Because the arginine-coding CGA is one of the rarest in E. coli (20), it likely creates a prolonged vacancy in the ribosomal A site. This vacancy would allow the peptidyl–tRNA* in the P site to slide rightward one base and still pair with a CCC triplet; thus, it would cause shifting to the reporter frame.

Frameshift-suppressing RNAs with an extra nucleotide in the anticodon loop were originally thought to cause a +1 frameshift by recognizing a four-base codon (21). Experimental contradictions to the four-base translocation model, including RNA suppressors with normal anticodon loop size but altered bodies (22), led to the development of a peptidyl–tRNA slippage model for several frameshift RNA suppressors of the CCCN (proline) and GGGN (glycine) suppression sites (23). In this model, slippage of the mRNA occurs when the suppressing RNA has moved to the ribosomal P site, and this slippage is in competition with the decoding of the next codon, where the presence of a stop codon or a low-abundance codon can increase slippage.

The presence of rare sense codons or codons under aminocacyl–tRNA limitation (amino acid starvation) contributes to translational bypassing or hopping (24, 25). In bypassing, ribosomes suspend translation at a certain site and then resume translation downstream without decoding a block of intermediate nucleotides, and thus a single protein is synthesized from two separated coding sequences. Depending on the number of nucleotides skipped, bypassing may cause a change of frame.

An extreme example of translational bypassing occurs in the translation of the bacteriophage T4 gene 60 (13). The ribosome reads the first 46 codons of the mRNA, pauses at a UAG stop codon, hops over 47 nucleotides (a 50-nucleotide coding gap), and resumes translation. This bypass requires matched codons. The peptidyl–tRNA pairs first with one codon, slips, and then pairs with the matching codon, which results in only one
amino acid being inserted for the two matched codons and any intervening nucleotides. The stop codon after codon 46 pauses the ribosome and allows time for slippage. The sequence context of the UAG codon creates a very poor termination site (11). In addition to the stop codon, a stem-loop structure that contains the stop codon and take-off site GGA within its stem and a cis-acting signal in the nascent peptide that consists of a stretch of charged and hydrophobic amino acids specified by codons preceding the gap enhance bypass efficiency (13).

**Codon Redefinition**

*N*-formylmethionine

All organisms initiate protein synthesis with either methionine or its derivative *N*-formylmethionine (fMet). In all organisms, two classes of methionine-tRNAs are present with one class used only for initiation and the other for peptide elongation (26). Archaeal and cytoplasmic eukaryal protein synthesis initiates with methionine. Bacteria and eukaryotic organelles (mitochondria and chloroplasts) require *N*-formylmethionyl–tRNAfmet as a translation initiator. Typically, the formyl group and the methionine are removed from the protein even while the protein is still being synthesized on the ribosome (26).

Synthesis of fMet occurs on its tRNA. The tRNAfmet is charged by MetRS with methionine, which then is formylated by methionyl–tRNA formyltransferase (MTF) in the presence of the formyl donor N10-formyltetrahydrofolate (27). Initiation factor IF2 sequesters the fMet–tRNAfMet and excludes it from the ribosomal A site. Instead, IF2-bound fMet–tRNAfMet is transported to the ribosomal P site. EF-Tu--GTP can bind methionyl–tRNAfmet, although at a lower affinity than

Unique structural properties, including the absence of a Watson–Crick base pair between nucleotides 1 and 72 at the end of the acceptor stem and the presence of three consecutive base pairs at the bottom of the anticodon stem, distinguish initiator tRNAs from elongator tRNAs (30). Observations from the E. coli crystal structure suggest the high specificity recognition between Met–tRNAMet and MTF mainly involves recognition of the acceptor arm (32). The lack of a Watson–Crick base pair at position 1-72 is important not only for the formation of this tRNA by methionyl–tRNA transformylase (32) but also for the prevention of the initiator tRNA from acting as an elongator tRNA. The three consecutive G–C base pairs on the acceptor stem of initiator tRNAs are required for targeting the tRNA to the ribosomal P site (33).

Selenocysteine

The trace element selenium plays an essential role in the activity of some bacterial and eukaryotic antioxidant enzymes (34). Selenium is incorporated into proteins in the form of the so-called twenty-first amino acid, selenocysteine, which is encoded by a UGA stop codon. Although the chemical structure of selenocysteine differs from cysteine only by the replacement of the sulfur atom with selenium, the lower pKa of selenocysteine results from the coaxial Watson–Crick base pair between nucleotides 1 and 72 at the end of the acceptor arm (31). The lack of a Watson–Crick base pair at position 1-72 is important not only for the formylation of this tRNA by methionyl–tRNA transylase but also for the prevention of the initiator tRNA from acting as an elongator tRNA. The three consecutive G–C base pairs on the acceptor stem of initiator tRNAs are required for targeting the tRNA to the ribosomal P site (33).

Selenocysteine synthesis occurs on tRNA^Sec (Fig. 1B). The secondary structure of tRNA^Sec differs from that of canonical tRNAs. Selenocysteine tRNAs are considerably larger than other tRNAs (36). Although tRNAs normally undergo processing of a 5' leader sequence, tRNA^Sec does not have a 5' triphosphate on its terminal guanosine (37). In eubacteria, archaea, and eukarya, a conservation of extensions of the acceptor stem and the D stem in tRNA^Sec exists. An extended 6 bp D stem, in eukarya, a conservation of extensions of the acceptor stem and phosphate on its terminal guanosine (37). In eubacteria, archaea, and eukarya (9), position between bacteria (8) and eukarya (37), a 13-bp A–T arm is necessary instead of the canonical 3–4 bp, was shown to be a major identity determinant for serine phosphorylation in eukarya (38). A 13-bp A–T arm is necessary for binding to SelB in bacteria and for serine to selenocysteine conversion in eukarya (41, 42). Investigation of posttranscriptional modification of the vertebrate tRNA^Sec revealed only four modified bases, which is fewer than canonical tRNAs (43).

In bacteria, seryl–tRNA synthetase initiates selenocysteine biosynthesis by charging tRNA^Sec with serine. The aminoacylation efficiency of this reaction is only 1-10% that of aminoacylation of tRNA^Met (44). In E. coli, selenocysteine synthase, encoded by selD, catalyzes the conversion to seryl–tRNA^Sec to selenocysteinyll–tRNA^Sec (45). Seryl–tRNA^Sec covalently binds to the pyridoxal phosphate (PLP) of selenocysteine synthase. From in vitro studies, after the elimination of a water molecule from the seryl moiety, formal addition of hydrogen selenide to the double bond of the aminocarboxylic intermediate occurs (46). Selenophosphate, which is synthesized by selenophosphate synthetase encoded by selD, acts as a selenium donor in vivo (47). In archaea (42) and eukarya (48), selenocysteine also starts with the serylation of tRNA^Sec by seryl–tRNA synthetase. In the presence of Mg^2+ and ATP, phosphoseryl–tRNA^Sec kinase specifically phosphorylates the seryl moiety of Ser–tRNA^Sec to produce O-phosphoseryl–tRNA^Sec (49). The conversion of O-phosphoserine (Sep) to selenocysteine proceeds by the action of the PLP-dependent enzyme Sep–tRNA:Sec–tRNA synthase (SepSecS) using selenophosphate as a selenium donor to produce Sec–tRNA^Sec (50). As the phosphate of Sep provides a better leaving group than the water of serine, Sep to Sec conversion is more chemically favorable than Ser to Sec conversion. This favorableness and the greater stability of the Sep–tRNA^Sec as compared with Ser–tRNA^Sec (49) would enhance the production of Sec–tRNA^Sec in archaea and eukarya, which have more extended selenoproteomes than bacteria (50).

To ensure that the opal codon codes selenocysteine only when needed, a cis-acting stem-loop structure designated as the Sec Insertion Sequence (SECIS) element must be present in the mRNA (34). In E. coli, a 38-nt SECIS is positioned immediately downstream from the opal codon, whereas the archaea and eukaryotic SECIS can be located several hundred nucleotides downstream in the 3' untranslated region of the gene. The presence of the SECIS in the 3' untranslated region yields greater sequence flexibility of proteins in the expanded selenoproteomes of archaea and eukarya. The ability of higher eukaryotes to translate selenoproteins with SECIS elements in the coding region suggests that the location of the SECIS element in the 3' UTR is not a requirement in higher eukaryotes but rather an evolutionary adaptation (51).

For delivery to the ribosome, special elongation factors are necessary. In E. coli, EF-Tu does not bind tRNA^Sec. Instead, a selenocysteine-specific elongation factor SelB, with an N-terminal domain that has sequence similarity to EF-Tu, recognizes the aminocarboxylic moiety and exclusively binds selenocysteinyll–tRNA^Sec (52). Like EF-Tu, SelB binds GTP. The C-terminal extension of SelB consists of four winged-helix domains arranged in tandem that bind the SECIS element and undergo a structural change during SECIS binding that is proposed to allow for communication between the tRNA and mRNA binding sites (53). A SelB conformational change during SECIS binding would be another mechanism to ensure that selenocysteine is not inserted at random codons (53). Thus, a quaternary complex of SelB, selenocysteyll–tRNA^Sec, GTP, and the SECIS element is necessary for selenocysteine insertion in E. coli. When in excess, this quaternary complex interacts with a SECIS-like element in the 5' nontranslated region of the selAB operon that encodes SelA and SelB to repress synthesis (54).

In archaea and eukaryotes, additional SECIS-binding proteins bridge a SelB-like protein and the SECIS element (55). M use EF–Sec and archaeal M. jannaschii Mseg (SelB) contain N-terminal domains functionally homologous to elongation factor EF–Tu (56, 57). The C-terminal extension does not interact with the SECIS. Instead, it interacts with bridging proteins. In eukarya, SECIS binding protein 2 contains both ribosomal and SECIS binding activity.
RNA binding domains, which are important for Sec incorporation (58). Ribosomal protein L30 found in eukaryal and archaeal kingdoms also has the ability to bind SECIS RNA (59). The exact roles and players in eukaryal and archaeal secenocysteine insertion remain unknown.

The location of the SECIS element in the 3′ untranslated region in arachaea and eukarya allows complete flexibility in the amino acid sequence surrounding the secenocysteine residue and allows for multiple secenocysteines in one protein (60). Expressing secenocysteines from archaea and eukarya or selective secenocysteine insertion in E. coli can be difficult because the E. coli SECIS element immediately follows the opal stop codon. In a case where the secenocysteine occurred at the N-terminus of an essential phage coat protein (M13 pII) to couple phage production effectively to nonsense suppression (66, 68, 69). The cloned sequences also suggested that the lower stem of the SECIS can be fully randomized and unpaired, and conservation of the loop and upper stem bulged U was observed. Additional phage library studies suggested much more flexibility in the sequence of the upper SECIS stem than previously thought, with one functional clone having five substitutions. This sequence flexibility would make it easier to express eukaryotic proteins in bacteria.

Secenocysteine displayed on the surface of the phage provides a uniquely reactive functional group, which permits the use of small electrophilic compounds for regiospecific covalent phage modification (68). This method has been used in setting up a system for enzyme evolution (76). Secenocysteine tags have also been used for protein purification and labeling using 4-phenylarsine oxide-Sepharose (PAO–sepharose) columns, which have also been used for protein purification and labeling using 4-phenylarsine oxide-Sepharose (PAO–sepharose) columns. The standard elongation factor EF-Tu directly recognizes RNA in vitro, which suggests that a specialized elongation factor may not be necessary for pyrrolysine incorporation (84). Studies of the activation of tRNAPyl by the PylRS from Desulfotomaculum hafniense showed that the discrimination between the acceptor and D stems is shorter by one nucleotide than the typical two-nucleotide junction. Mass spec analysis of the tRNAPyl revealed a relative mass of the RNA in vitro, which revealed a relatively low mass of nucleoside modifications, especially in the anticodon stem-loop (83). The standard elongation factor EF-Tu directly recognizes RNA in vitro, which suggests that a specialized elongation factor may not be necessary for pyrrolysine incorporation (84). Studies of the activation of tRNAPyl by the PylRS from Desulfotomaculum hafniense showed that the discrimination between the acceptor and D stems is shorter by one nucleotide than the typical two-nucleotide junction. Mass spec analysis of the tRNAPyl revealed a relative low mass of nucleoside modifications, especially in the anticodon stem-loop (83). The standard elongation factor EF-Tu directly recognizes RNA in vitro, which suggests that a specialized elongation factor may not be necessary for pyrrolysine incorporation (84). Studies of the activation of tRNAPyl by the PylRS from Desulfotomaculum hafniense showed that the discrimination between the acceptor and D stems is shorter by one nucleotide than the typical two-nucleotide junction. Mass spec analysis of the tRNAPyl revealed a relative low mass of nucleoside modifications, especially in the anticodon stem-loop (83). The standard elongation factor EF-Tu directly recognizes RNA in vitro, which suggests that a specialized elongation factor may not be necessary for pyrrolysine incorporation (84).

In contrast to Sec–tRNA Sec and fMet–tRNA fMet formation, Pyrrolysine (Pyl), the twenty-second amino acid, has been discovered incorporated in methyamine methyltransferases from Methanosarcinae, a branch of methanogenic archaea that has the ability to reduce a wide variety of compounds to methane including carbon dioxide, acetate, methanol, methylated thiol, and methylated amines (72). Methanogenesis from methyamines as (mono-, di-, or trimethylamine) uses non-homologous, substrate-specific methyltransferases: monomethylamine methyltransferase (mtMB), dimethyamine methyltransferase (mtMB), and trimethylamine methyltransferase (mtMB) (72). Sequencing of the Methanosarcinaceae (73) and Methanococcoides genomes shows all methanogen methylamine methyltransferase genes contain in-frame amber codons that must be suppressed to produce full-length protein (72, 74). The crystal structure of Methanosarcina bacteri MtmB revealed a lysine with its Nε in an amidic linkage with (4R,5S)-4-substituted-pyrrolidine-2-carboxylate at the coding position corresponding to the stop codon (75). The C-4 substituent of pyrrolysine was identified as a methyl group from tandem mass spectrometry of Methanosarcina bacteri MtmB, MtMB, and MtMB (76). In Mtb B, Pyl is located at the enzyme active site and likely serves as a strong electrophile (77). Each methylamine transferase specifically methylates its cognate corrinoid protein (MtMC, MtMB, MtMC, or MtMB) using its specific methylamine (74). Methylcobamide:coenzyme M methyltransferase (MtMB) demethylates the corrinoid using zinc to form the nucleophilic coenzyme M thiolate (78). Reduction of methyl-coenzyme M to methane is the major energy-conserving step of methanogenesis (79).

Near the mtnB1 gene in Methanosarcinae are the pytTSBCD genes, which represents a "genetic code expansion cassette" that allows for the production and insertion of pyrrolysine when transferred to E. coli (77, 80). PylS is an archael class II pyrrolysyl–tRNA synthetase (81). PylB, PylC, and PylD may play a role in pyrrolysine biosynthesis (82). The pyt gene encodes RNA (83), which has a CUA anticodon (82). Structural differences exist between RNA and other RNA molecules (82). In RNA, the anticodon stem can form an additional base pair creating a six-base pair stem and three-base variable loop. With only a five-base pair anticodon stem, the link between the D stem and anticodon stem is two nucleotides instead of the typical one nucleotide. The junction between the acceptor and D stems is shorter by one nucleotide than the typical two-nucleotide junction. Mass spec analysis of the RNA revealed a relatively low mass of nucleoside modifications, especially in the anticodon stem-loop (83). The standard elongation factor EF-Tu directly recognizes RNA in vitro, which suggests that a specialized elongation factor may not be necessary for pyrrolysine incorporation (84). Studies of the activation of RNA by the PylRS from Desulfutobacterium hafniense showed that the discrimina-
from gene \textit{mtmB1}, which contains an internal UAG stop codon (86, 87). These studies suggest that \textit{M} \textit{ethanarciacnaecae} may use both direct and indirect charging of tRNA\textsubscript{Pyl} to ensure efficient readthrough of the UAG stop codon (83).

Much debate surrounded the existence and necessity of a pyrrolysine insertion element (PYLIS) analogous to the SE\textsubscript{CIS} element until the examination of in vivo contextual requirements for pyrrolysine in \textit{M} \textit{ethanarciacnaecae} revealed that although either termination of translation or pyrrolysine insertion elements are present in \textit{M}. \textit{barkeri} with or without a PYLIS structure present, the presence of a PYLIS element increases the efficiency of pyrrolysine insertion (88). This study concluded that the need for high concentrations of \textit{m} \textit{ethylamine methyltransferases when grown on \textit{m} \textit{ethylamine would make read-through enhancement provided by the PYLIS at least greatly advantageous if not essential. The presence of pyrrolysine seems to be isolated to a small subset of prokaryotes. Searches of completely and incompletely sequenced prokaryotes genomes thus far have revealed only seven organisms that could use Pyl, including four members of archaea \textit{M} \textit{ethanarciacnaecae} genera, the archaea \textit{M} \textit{ethanarciacnaecae} \textit{burtonii}, the Gram-positive bacteria \textit{Deinococcus} \textit{thermoacidophilus} \textit{halnense}, and the symbiotic \textit{delatproebacterium} of the gutless worm \textit{O} \textit{livus} \textit{alagavensii} (82, 89). Analysis of Pyl-containing archaea revealed that less than 5% are predicted to terminate at UAG, and thus in these organisms, as complex a system as controls se-lenocysteine insertion may not be needed to regulate the read- ing of UAG as pyrrolysine (90).

Artificial Genetic Code Expansion by Directed Translational Recoding

The complement of known cotranslationally incorporated amino acids is currently limited to the “standard” twenty, plus selenocysteine, pyrrolysine, and N-formylmethionine. The ability of the ribosome and elongation factors to accept a variety of tRNA structures and appended amino acids has allowed the development of several methods that recruit the translational machinery for incorporation of noncanonical (“unnatural”) amino acids at defined positions in expressed proteins. Such expansion of the genetic code has allowed specific incorporation of amino acid-like structures with unique chemical, steric and biological properties into any protein of interest (91, 92). All of these methods rely on the following components: 1) a tRNA that will insert the desired residue efficiently at defined positions in a single expressed protein, while not being recognized by any of the endogenous aminoacyl-tRNA synthetases in the expression system being used, and 2) a method for coupling the desired amino acid to the acceptor stem of the tRNA.

Choice of tRNA

The first criterion for a candidate tRNA to be used as a delivery vehicle for noncanonical amino acids is the ability to insert its appended amino acid site specifically into a target protein. An obvious route is redefining existing codons to encode the noncanonical residue. Rather than using a cis sequence element to recode an existing coding triplet (analogous to the selenocysteine insertion pathway), a more straightforward approach is nonsense suppression. Because amber (UAG) suppressors are ubiquitous in nature and genomes have evolved to use the UGA and UAA as the predominant termination codons, most methods for noncanonical amino acid insertion rely on amber suppressor tRNAs. Frameshift suppressors have also been used (93). A third- tionally, synthetic orthogonal base pairs, in which one or more unnatural nucleotides in the tRNA anticodon, have also been used (94), although this method requires the use of in vitro protein expression systems.

The second criterion is that the tRNA not be recognized by the endogenous aminoacyl-tRNA synthetases in the host organism. This requirement is crucial, as the endogenous synthetases not only aminoacylate the tRNA with a canonical amino acid after it gives up its noncanonical residue on the ribosome but also may remove a noncanonical residue from the tRNA via natural proofreading mechanisms. Early work (95) in which translation was carried out in an \textit{E}. \textit{coli}-based in vitro system used an amber suppressor based on tRNA\textsubscript{TM} from Saccharomyces cerevisiae, which had been demonstrated to be unrecognized by the endogenous aminoacyl-tRNA synthetases in \textit{E}. \textit{coli}. Later work in vivo in expression systems employed, for example, archaeal tRNAs (\textit{i.e.}, \textit{M} \textit{ethanarciacnaecae} \textit{annachii} tRNA\textsubscript{TM} in an \textit{E}. \textit{coli} system (96), \textit{E}. \textit{coli} tRNAs in a yeast expression system (97), and bacillus tRNAs in a mammalian expression system (98), although the natural tRNAs are often measurably aminoacylated in these heterologous expression systems. With appropriate genetic selection systems (vide infra), a fully orthogonal mutant tRNA could be generated that is not recognized by the entire synthetase complement in the chosen expression system.

Methods of aminoacylation

The second half of the problem is how to get the amino acid onto the tRNA before translation. Early work by Peter Schultz and coworkers (95) employed a variation of a chemical misacylation strategy originally developed by Sidney Hecht (99). This method avoids the intractable problem of regiochemically coupling an amino acid onto a 76-nucleotide forest of competing functional groups by focusing the derivatization on the 3’ acceptor stem of the tRNA only. The chimeric DNA-tRNA \textit{dimer pDCpA} is synthesized and specifically aminoacylated on its 3’ end using the cyanomethyl ester of the desired amino acid. The use of deoxycytidine greatly simplifies the synthesis of the \textit{dimer} (100), and the use of the cyanomethyl ester obviates the need for internal protection of the \textit{dimer} before aminoacylation (101). \textit{T4 RNA ligase} then is used to ligate the resulting aminoacyl-\textit{pDCpA} onto an amber suppressor tRNA lacking its 3’ terminal CA generated by in vitro transcription (102), which generates the full-length aminoacyl-tRNA. Despite the presence of a deoxynucleotide at position 75 and a lack of any of the ubiquitous tRNA modifications, the tRNA can participate in translation and give up its loaded amino acid in response to the corresponding amber codon. The use of chemically misacylated tRNA requires the use of an in vitro protein expression system, but the strength of this approach is its flexibility. In theory, any
amino acid-like structure can be coupled to tRNA using identical chemistry, which gives the user the ability to incorporate many amino acid variants into a protein in a relatively short amount of time.

The principal drawback of the chemical aminoacylation approach is low yield of expressed protein. This low yield is a result not only of the inherent low yields of in vitro protein synthesis systems, but also of the tRNA not being recycled after giving up its amino acid on the ribosome. An obvious solution is to employ an aminoacyl-tRNA synthetase to aminoacylate the tRNA, which permits both in vivo expression and recycling of the tRNA. It has long been known that under auxotrophic conditions a naturally occurring synthetase can be tricked into incorporating a media-supplemented noncanonical amino acid that is structurally related to its cognate substrate, for example, incorporation of selenomethionine in place of methionine for X-ray crystallography (103). As this strategy uses the existing in vivo pool of tRNAs, incorporation is global rather than site specific. Nevertheless, this approach has become a powerful tool for generating novel biomaterials (104).

Homogenous, site-specific incorporation of a noncanonical residue in vivo requires a fully orthogonal synthetase-tRNA pair, in which the synthetase recognizes no other tRNA or amino acid, and the tRNA is not recognized by any other host synthetases. Such a pair can be generated by rational mutagenesis (105), but a more general approach uses a “double-sieve” genetic selection system (106). Briefly, amber suppressor tRNAs that are not recognized by host synthetases are selected from a library of tRNA variants coexpressed with a lethal gene (bar), which results in in-frame amber stop codons. Surviving tRNA variants are then coexpressed with the desired cognate synthetase in the presence of an amber suppressor tRNA with an in-frame amber codon, which results in selection of tRNAs that are recognized only by the desired synthetase. Synthetases that can aminoacylate only the now-fully-orthogonal suppressor tRNA with the desired amino acid, but not the canonical amino acids present in vivo, could be obtained via an analogous two step genetic selection (96): Coexpression of the synthetase library with the RNA and an amber gene with an in-frame amber codon in the presence of the desired amino acid results in selection of synthetases capable of aminoacylating the RNA with the desired amino acid or any of the canonical amino acids. Coexpression of these selected synthetases with a lethal gene with in-frame amber codons in the absence of the desired amino acid then yields the desired fully orthogonal synthetase-tRNA pair that will direct insertion of only the desired noncanonical amino acid in vivo. Similar strategies have been used to create orthogonal synthetase-tRNA pairs in yeast (97) and mammalian cells (98). In all cases, the noncanonical amino acid must be supplied exogenously in the media, which theoretically limits residues to those that are compatible with the cellular transport machinery and that are not substrates for intracellular metabolic pathways. These issues could be addressed by engineering a biosynthetic pathway for the noncanonical amino acid directly in the organism, which results in a fully self-contained organism with an expanded genetic code (107).

Applications

A detailed discussion of the hundreds of noncanonical amino acids incorporated via these methods is outside the scope of this article. Most applications fall into four main areas: protein labeling, protein modification, investigation of protein structure–function relationships, and engineering activity “switches” into enzymes (reviewed in References 91 and 92). Labeled residues can be incorporated directly via this method, such as isotopically enriched residues for NMR, spin labels, and fluorescent tags. Additionally, many residues have been incorporated that carry orthogonally reactive side chains (e.g., keto groups, acids, alkenes, and thiocysteates) that permit specific modification of a single residue with a variety of appended labels under conditions in which the rest of the proteome is unreactive. Other protein modifications (glycosylation, PEGylation, metal-binding sites, cross-linking agents, and redox-active groups) can be incorporated either directly or by posttranslational modification. A variety of backbone and side-chain modifications have been used to study protein folding and hydrophobic packing, and altered nucleophiles and hydrogen-bond donors/acceptors have been used to explore enzymatic catalysis. Finally, photochemical switches that allow precise control of proteases, transcription factors, and protein splicing have been incorporated.

Implications

A variety of amino acid-like structures has been incorporated via directed artificial translational recoding, including α-hydroxy acids, N-substituted amino acids, and severely constrained amino acids (e.g., 1-amino-1-carboxycyclopropane), along with myriads side chains that far exceed the length and shape parameters of the canonical 20 residues. The ability to accommodate such a broad range of structures suggests that the translation machinery may have evolved to accept many other amino acids beyond the canonical 20 and that other naturally occurring co-translational amino acids remain to be discovered, as evidenced by the recent discovery of the “twenty-second amino acid" pyrrolysine.

References


10. A dansi MS, Donly BC, Tate WP. Competition between frame shifting, termination and suppression at the frameshift site in the Escherichia coli release factor-2 mRNA. Nucleic Acids Res. 1993;21:5074-5078.


40. Baron C, Krol A. The length of the aminoacyl-acceptor stem of the selenocysteine-specific tRNA (Sec) of Escherichia coli is the determinant for binding to elongation factor SELB or Tu. J. Biol. Chem. 1991;266:20375-20379.


Translational Recoding

See Also

Aminoacyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Sitespecific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-

103. Hendrickson WA, Horton JR, LeBaron DM. Selenomethionyl
proteins produced for analysis by multidimensional anomalous
diffraction (MAD): a vehicle for direct determination of three-

104. Link AJ, Tirrell DA. Realignment of sense codons in vivo.

105. Kwon I, Tirrell DA. Site-specific incorporation of tryptophan
analogues into recombinant proteins in bacterial cells. J. Am.

106. Wang L, Schultz PG. A general approach for the generation
and function with an expanded genetic code. Angew Chem. Int.

107. Mehl RA, Anderson JC, Santoro SW, Wang L, Martin AB, King
DS, Hom DM, Schultz PG. Generation of a bacterium with a 21

Further Reading

Aminocyl tRNA Synthetases, Chemistry of
Translation: Topics in Chemical Biology
Phage Display
Translation Machinery: Modifications to
chemically aminocylation transfer RNA prepared by runoff trans-
The screening of chemical libraries is one of the major sources for new leads in drug discovery. The large size of chemistry space compared with the library sizes that are feasible to screen requires careful selection of the compounds for the screening library to maximize screening success. Besides issues around technology compatibility and chemical tractability of the compounds, the main objective is to increase the probability of obtaining hits for the screened targets. Diversity selection approaches have often shown only limited success. In the absence of any knowledge, it is proposed to screen smaller “lead-like” ligands with preference. When knowledge about the target is available, it can be used for target-focused compound selection or for library design. In the screening process, physical high-throughput screening (HTS) can be combined with virtual screening either to avoid the high-throughput primary screen of the whole library or to limit false negatives by combining primary HTS and virtual screening results. Screening an initial subset then using the results obtained to predict likely hits for subsequent screening rounds in sequential screening can lessen the number of compounds to be screened, but it causes a greater logistics effort and has the risk of missing compounds that are not well represented structurally by the initial set. Data analysis and visualization of the screening results are a necessary final step of a screening campaign to ensure that the prioritization of compounds followed up is based on all available relevant information.

When in vitro biologic assays replaced in vivo animal models as the first tool to assess biologic activity of molecules in drug discovery, the possibility existed to test many more compounds than was possible before. This triggered the hope that the slow process of lead discovery, which relies to a large extent on medical chemists’ intuition and serendipity, could be accelerated by a systematic brute-force screening of large collections of chemical compounds, for which the term “chemical libraries” has been introduced. Consequently, the pharmaceutical industry has built up high-throughput screening (HTS) facilities (see the article “High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology”), in which in vitro assays could be performed in a highly parallel, miniaturized, and automated way. With HTS available, it became not only possible to screen the historically accumulated compound collections of pharmaceutical companies, but also a much greater number of compounds exist. This finding triggered the demand for a highly parallelized and automated synthesis of compounds to feed the HTS machinery (see the articles “Combinatorial Libraries: Overview of Applications in Chemical Biology” and “Small Molecule Combinatorial Libraries”). Although large pharmaceutical companies screen compound libraries in the magnitude of 10⁶ molecules, this approach is far away from a systematic brute-force approach because the chemistry space is estimated to contain 10¹³–10⁶⁰ small molecules (1). The conservative estimate of 10¹³ molecules is based on well-established chemical reactions and commercially available reagents (2). Extrapolation from a systematic enumeration of all theoretically viable organic molecules up to 11 non-H atoms toward 25 non-H atoms (the average size of drug-like molecules) suggests the existence of 10²⁷ unique structures (3). Because only a small subset of the chemistry space can be screened, the compounds must be chosen appropriately to maximize the success of the screen. Three groups of criteria for this exist.
The enzyme in biochemical assays in an unspecified way can lead to the detection of false-positive hits. The exact cause for the aggregate formation and the mechanism and conditions of the enzyme sequestration are not understood completely; however, experimental procedures have been suggested to detect false positives caused by aggregation (7).

The second property of importance for bioavailability is the polar surface area (PSA) that is associated with intestinal absorption and cell membrane penetration by passive transport. Compounds with a high polar surface are less likely to penetrate the lipophilic environment of the cell membranes by passive transport. Like the logP, PSA can be computed by summing up fragment contributions (8) with H-bonding fragments as the main contributor.

The role of the physical chemical properties discussed so far is the rationale behind two popular rules of the thumb to estimate “drug-likeness”: Lipinski’s rule-of-five (9), in which counts of hydrogen bond donors and acceptors take the place of the PSA, and the “Egan Egg” (10) (see Table 1).

**Chemical purity and stability**

To rationalize the results of a screen and to derive structure activity relationships (SAR) guiding the additional optimization of the compounds, it is prerequisite that the activity of a compound sample results from the structure being attributed to it. This can only be ensured if the compounds included in the screening collection are reasonably pure. Typical purity requirements are in the range of 85% to 95%. Impurities that interfere with the assay technology especially must be avoided.

To remain in that state of purity, the compounds must be chemically stable under the conditions of storage, and because the fresh production of screening solutions from powder sample is not feasible for each individual HTS, the compound must be stable in DMSO solution over a period of time if the stock solution is intended to be used.

A chemically unstable compound is also not well suited to be marketed as a drug: therefore, insufficient chemical stability is also an issue for optimizability. If the compound contains chemical groups that are reactive toward DNA, the compound can become mutagenic, which constitutes another liability for optimization.

Practically, the stability issue is addressed mostly by the application of substructural filters to remove compounds with known labile and reactive functionality. Several sets of substructure filter sets published share a large degree of overlap (13, 14).

**Other reasons for technology incompatibility**

In addition to the general criteria discussed above, each assay technology can also suffer from technology-specific interference of chemical compounds. A sub-fluorescent compounds can interfere with any fluorescence-based readout. In assays that use the biotin-streptavidin interaction, biotin analogs are potential false positives.

First, the compounds must be compatible with the compound handling and screening technology used and should not cause assay artifacts. Second, the library must contain molecules with the desired activity. Last, once a hit is identified, the molecule must be optimizable into a drug candidate with suitable efficacy, bioavailability, therapeutic window, and, in the case of industrial drug discovery, patentability (see the article “Lead Optimization in Drug Discovery”).

**Technology Compatibility and Optimizability**

These two objectives are discussed together because many of the selection criteria for screening compounds are important to fulfill both objectives, namely physical chemical properties and chemical purity and stability. Especially with respect to optimizability, a violation of the selection criteria discussed here is not a definite reason for exclusion, but it is a liability of the compound, which needs to be addressed during the optimization of the compound that follows the discovery of the hit. The more such liabilities a compound exhibits, the more difficult the optimization of a lead compound will be. Also, not every violation of a technology compatibility criterion is a hard reason for exclusion, but rather it increases the potential of a compound to cause artifacts under a certain assay technology; appropriate experimental procedures are required to detect these artifacts.

**Physical-chemical properties**

Generally, biologic assays are performed in aqueous solution, typically in a concentration of up to 50–100 \( \mu \)mol. These solutions are produced by diluting a stock solution of the compound in dimethylsulfoxide (DMSO) in the millimolar concentration range with buffer. Therefore the compounds must be soluble in water and in DMSO under the respective conditions, or a potential activity of the compound remains undetected or is underestimated largely. Water solubility is equally important for bioavailability of the drug, in which sufficiently high blood plasma levels must be achieved for efficacy. Unfortunately, neither the experimental determination of water solubility nor its prediction by computational methods is straightforward, because both depend not only on the hydrophilicity of compound, but also on the lattice energy of the crystal (4). Thus, based on Yalkowski’s equation (5) as a guideline to estimate water solubility, the logarithm of the octanol–water partition coefficient (logP) has been used frequently. It can be predicted by summing up fragment contributions that have been fitted on experimental data (6) as the ratio between hydrophobic and hydrophilic fragments. From a high lipophilicity as indicated by a high computed logP (ClogP), it can be concluded that the water solubility of the neutral compound is low; however, a low ClogP does not guarantee high water solubility. Protonation of basic groups or deprotonation of acidic groups lead to ionic species that frequently have higher solubility than neutral compounds.

In this context, it is noteworthy that lipophilicity is not only related to low aqueous solubility, but also to the tendency of compounds to form aggregates. Such aggregates can sequester the enzyme in biochemical assays in an unspecified way and can lead to the detection of false-positive hits. The exact cause for the aggregate formation and the mechanism and conditions of the enzyme sequestration are not understood completely; however, experimental procedures have been suggested to detect false positives caused by aggregation (7).

The second property of importance for bioavailability is the polar surface area (PSA) that is associated with intestinal absorption and cell membrane penetration by passive transport. Compounds with a high polar surface are less likely to penetrate the lipophilic environment of the cell membranes by passive transport. Like the logP, PSA can be computed by summing up fragment contributions (8) with H-bonding fragments as the main contributor.

The role of the physical chemical properties discussed so far is the rationale behind two popular rules of the thumb to estimate “drug-likeness”: Lipinski’s rule-of-five (9), in which counts of hydrogen bond donors and acceptors take the place of the PSA, and the “Egan Egg” (10) (see Table 1).
Selection of Compounds with Desired Biologic Activity

In this section, it will be discussed how likely the screening library is to contain compounds that are active on the targets of interest. In the absence of any knowledge about the structure of the target or its ligands, only diversity-based sampling methods can be used. Knowledge about the target or its target family can be used for the design of target (family)-focused library.

Diversity-based strategies

The central hypothesis for all diversity-based strategies of compound selection is the similarity property principle, which states that molecules with similar structures can be expected to have similar properties and to bind to the same target proteins (15). Following this principle, it is only necessary to screen one representative molecule out of a group of molecules with similar structures because the other molecules of the group should have the same behavior as the representative. Consequently, many algorithms exist to select diverse subsets of molecules from a database that represent the groups of unselected molecules. These algorithms have been reviewed elsewhere (16, 17) and only a short overview is provided here. Most methods encode the molecular structures as a descriptor vector, from which similarity coefficients for pairs of molecules can be calculated without aligning the molecules. Then, these similarity coefficients are used in diversity selection or clustering algorithms (18). From a clustering solution, a diversity selection is obtained by choosing one or more representative molecules from each cluster. Each molecule in a diverse subset is expected to represent the nonselected molecules, and it can be interpreted as the center of a cluster formed by its similar neighbors. For the sake of a more descriptive discussion, the clustering viewpoint is assumed in the following paragraphs; however, the arguments made can generalize to other diversity-selecting procedures. A literature to clustering, rule-based methods, typically based on the molecular scaffold, can be used to create partitions of molecules from which the representatives are selected (19–21).

Despite initially high expectations, diversity-based strategies for compound selection have shown only limited success. Diversity selection from the MDL Drug Data Report (MDDR), a database that contains only molecules with documented pharmacological properties, led to an enrichment of covered activity classes (22). However, diversity selections from a compilation of screening data that includes inactive molecules did not lead to an enrichment of targets covered by selected compounds (23). In a clustering experiment, the intraclass similarity of the IC50 vectors of the compounds measured in a uniform panel of assays was not much greater than the intra-group IC50 similarities of compounds grouped randomly (24).

How can these results be understood? At first, the similarity property principle is only valid on rather short similarity ranges. According to a popular rule of thumb, molecules that have a Tanimoto similarity coefficient of 0.85 calculated over the Daylight fingerprints are supposed to be very similar; however, often they differ significantly in their protein binding properties (25). Also, the inversion of the similarity property principle that dissimilar molecules should also have dissimilar protein binding properties is not generally true (26).

Second, the theory that the screening of only one representative per structural cluster is sufficient to determine the activity of the cluster assumes that the screening procedure is error free.

### Table 1: Empirical “rules-of-thumb” to estimate the suitability of compounds at different stages of drug discovery based on structural properties

<table>
<thead>
<tr>
<th>Name</th>
<th>Rule</th>
<th>Purpose</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule of fives</td>
<td>Two or more of the following conditions violated:</td>
<td>Estimate whether a compound’s</td>
<td>(9)</td>
</tr>
<tr>
<td></td>
<td>MW ≤ 500 Da</td>
<td>absorption and membrane permeation are good</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ClogP ≤ 5</td>
<td>enough to be orally bioavailable</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HBD ≤ 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HBA ≤ 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egan Egg</td>
<td>Ellipse defined in the ClogP and PSA space.</td>
<td>Estimate whether a compound’s</td>
<td>(10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>absorption and membrane permeation are good</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>enough to be orally bioavailable</td>
<td></td>
</tr>
<tr>
<td>Lead-likeness</td>
<td>MW ≤ 460 Da and ClogP ≤ 4.2 and LogS &lt; −5 and HBD ≤ 5 and</td>
<td>Identify compounds that have the potential</td>
<td>(11)</td>
</tr>
<tr>
<td></td>
<td>RNG ≤ 4 and HBA ≤ 9</td>
<td>to be successful leads</td>
<td></td>
</tr>
<tr>
<td>Rule of the three</td>
<td>MW ≤ 300 Da and ClogP ≤ 3 and HBD ≤ 3 and HBA ≤ 3 PSA ≤ 60 Å2 and</td>
<td>Identify compounds that have the potential</td>
<td>(12)</td>
</tr>
<tr>
<td></td>
<td>RTB ≤ n3</td>
<td>to be successful fragment screening hits.</td>
<td></td>
</tr>
</tbody>
</table>
Any error in the screening for the representative molecule is extrapolated to the whole cluster and leads to its misclassification. To compensate screening errors and to determine its activity correctly, it is necessary to screen several representative molecules from a cluster of molecules with assumed common biologic activity. A statistical model to determine the number of representatives that need to be screened per cluster based on empirically estimated false positive and negative rates has been published by Harper et al. (27). They described that the probability that a compound is active is determined by the product of the probability $\pi_i$ of the cluster that contains the active compound (variable from cluster to cluster) and a probability $\alpha$ that describes the probability of an individual active compound, provided the cluster is active. In this model, $\alpha$ accounts for the average error of the screening process that leads to an erroneous determination of an individual compound’s activity and errors of the clustering procedures that lead to the erroneous grouping of a compound to a cluster with different activity. The activity related to the common “chemotype” or pharmacophore of the cluster $i$ is described by $\pi_i$, and even an active cluster with a high $\pi_i$ is likely to be missed if $\alpha$ is low. If $n_i$ compounds per series are screened, the probability to find at least one hit if the cluster is active equals $(1-1-\alpha)^{n_i}$.

The third reason for the limited success of diversity-based strategies is the low baseline probability for bioactivity, with hit rates of 0.1% as the typical order of magnitude. In the case where a clustering was highly predictive of biologic activity, with active clusters that show a 100-fold enrichment of active compounds, this would still indicate that, on average, only 10% of compounds in any given active cluster are active. If the cluster was sampled with only one compound, the probability that the cluster is identified correctly as active would only be 10%. This theory is illustrated qualitatively in Fig. 1.

**How large should a screening library be?**

Because of the inefficiency of an untargeted, diversity-oriented sampling, it is not possible to determine a meaningful library size as sufficient to sample the chemistry space. However, one can only get value out of a screening library that is matched by adequate resources for screening, compound logistics, and data handling. The costs for this and for the compounds practically limit the library size. The statistical model by Harper et al. (27) predicts the growth of the expected rate of successful screens. Successful means here that at least one hits series is found. According to this model the average number of hits series identified per assay increases linearly with the number of compounds in the library, assuming the average number of compounds per series is kept unchanged. However, the probability of success does not grow linearly with the size of the library. Each incremental addition of compounds leads to a smaller increase of the probability of success than the previous addition of the same number of compounds.

**How large should the molecules in the library be?**

The probability of a molecule to match the binding site of its target is assumed to be dependent on the size of the molecule. This relationship is described by the qualitative model of Hann et al. (28). After an initial increase, the probability that a ligand matches the binding site in exactly one orientation decreases with the ligand size. In addition, the number of potential molecules increases exponentially with the number of atoms; therefore, with increasing cutoff for the molecular size, it becomes more and more difficult to sample the chemistry space (3, 11). On the other hand, the larger the molecule, the more binding contacts it makes if it fits the binding site perfectly, which leads to a higher binding affinity (29). The binding affinity that is required minimally for detection depends on the sensitivity of the assay. An increase of assay sensitivity leads to a decrease of the required minimal size for ligands that have the potential to bind with a detectable affinity. For the reasons stated above, it makes sense to screen molecules in the size range that is large enough to allow for a detectable affinity, but not larger. Lead-likeness criteria have been formulated based on this finding (see Table 1). In fragment-based screening (FBS), highly sensitive biophysical assay technologies are used to detect the binding events of small molecular fragments to proteins (12). In the molecular size range used in FBS, the hit rate is, in accordance with the Hann model, much higher than in conventional biologic assays. The observed affinity is much lower (30), which requires the fragments to be amenable to chemical transformation to evolve them to molecules for which activity can be validated and be optimized using biochemical assays.
series, the ligand efficiency for the best compound after each optimization step is in most cases constant, what indicates that an increase of affinity coincides with an increase of molecular weight (MW). To achieve a final drug candidate with a potency of less than or equal to 10 nM and a MW less than or equal to 500 Da to comply with Lipinski’s rule, a LE of 0.016 pK_{DA}^{-1}

is the minimum requirement. Conventional HTS are typically sensitive enough to detect compounds K_{i} in the range of 1 μM.

Assuming that the ligand efficiency is indeed constant for a chemical series, and only ligands with LE greater than or equal to 0.016 pK_{DA}^{-1} have the potential to optimized into a suitable drug candidate, then it would be sufficient to screen compounds with a maximum MW of 375 Da. Likewise, for a biophysical fragment screen that can detect K_{i} in the range of 1 mmol/L, a maximum MW of 188 Da would be sufficient if the binding constant translates into an inhibition constant of the same order of magnitude. However, the criterion of LE greater than or equal to 0.016 pK_{DA}^{-1} may not be achievable for every target. In Reference 32, chemical series exist with lower LE values that nevertheless went into preclinical development. It must also be taken into account that a screening with a low LE can still be a suitable tool compound and may serve as a starting point to design a scaffold with a higher LE, which has not been present in the screening library. The definition of optimization of a chemical series used by Hajak is very narrow, and it allows an initial hit fragment to grow but not to have parts of it removed. In the exploration of HTS hits, pruning operations are frequent, although an ideal library also would have contained the compound that resulted from the pruning in the first place. Therefore, such LE-derived cut-off criteria for screening hits rarely can be applied stringently.

**Target family-focused screening libraries**

The limitations of the brute-force sampling of the chemistry space lead to attempts to identify structures that would have a greater chance of success. Virtual screening methods are now well-established technologies (33) (see the article “Computational Approaches in Drug Discovery and Development”). Moreover, these methods work on individual targets, of which some knowledge is needed either about protein structure or about ligands. When assembling a screening library for lead discovery, the main objective is to design a broadly usable collection that also contains ligands for targets about which very little is known. Fortunately, drug targets are not isolated in the pharmacology space, and many pairs of targets share some common ligands, especially if they belong to the same protein family (34). Within a protein family, ligands are often similar enough that searching for chemical similarity with the ligands of one representative member also identifies ligands of the other members without reported activity on the reference protein (35). According to the SAR homology concept (36), targets of one family often share similar ligands, and similar SAR behavior. This allows the library design to focus on a target family rather than an individual target and to leverage the whole set of knowledge for a target class (see the article “Target Family-Biased Compound Library: Optimization, Target Selection, and Validation”). Typically, chemical similarity searching is based on the comparison of molecular descriptor vectors with similarity coefficients (37). Although similarity searching with the individual ligands and combination of the results by data fusion can be highly successful (38), the numerical or binary nature of the descriptor vectors allows a whole range of machine learning techniques to be applied from other areas of multivariate statistics. Examples of these techniques are binary kernel discriminators (38), support vector machines (39), emerging patterns (40), naive Bayesian classifiers (41), and self-organizing maps (42). Self-organizing maps have become especially popular because of the intuitive visualization of their results (43, 44). Often, the results depend strongly on the target class chosen and the available data. For this reason, one key success factor is to compile as comprehensive and accurate a reference set as possible, which requires bioactivity databases that are well integrated into both bioinformatics databases that describe protein family membership and chemical databases that characterize the ligands. Although considerable room for improvement still exists in this sector, a wide range of databases has become available, and they were reviewed recently (45) (see the article “Small Molecules, Drug-Target Databases”).

Structure-based virtual screening technologies use the complementarity between the structural features of ligand and its target protein-binding site. Docking, which tries to predict explicitly the orientation of a ligand within the binding site (pose) and to estimate its binding energy (score), is the most frequently used method and has now become well established (46). However, a new method uses the target-ligand complementarity for the generation of predictive models without generating binding poses (47).

To make docking suitable to identify the ligands of a whole target family, it is necessary to address the issue of how to deal with family members without an available protein structure and how to overcome the inaccuracy of the scoring functions for the analysis of the docking results. A substantial improvement of the results can be obtained by using not only the docking score as a decision criterion to retain or reject a pose, but also the key interactions between ligand and protein. Based on binding affinity of known ligand, the scoring weights of different interactions can be adjusted to reflect the SAR of the ligands (48). Ligand-protein interactions can be described by discrete bit vectors comparable with chemical fingerprint descriptors, which allow the efficient and fast filtering of poses as described for the design of a kinase-focused library (49). Recently, a method has been described to train the weights of interactions in the scoring functions automatically based on a set of known ligands. This method allowed the authors to predict not only the activity for the kinase to which the ligand has actually docked into, but also, by using a training set of activity data for a second kinase that is different from the one used for docking, the activity for this second kinase. This method allows predicting activity for kinases without a known structure (50). It is an example of a method in which classical docking is combined with statistical learning. To apply these methods, both protein structures and activity information over large set of ligands is required. Even further generalizations of the binding sites are schematic descriptions of kinase binding sites summarizing the features of several kinase inhibitor complexes and the variations in their binding pockets between different kinases.
Such qualitative models have been used successfully to design kinase-focused libraries. In the case of G-protein coupled receptors (GPCRs), the structural information is rather sparse; bovine rhodopsin is the only GPCR for which a crystal structure is currently available. However, this structure could be used to determine which residues are positioned within the binding cavity. A qualitative model has been derived to visualize the molecular recognition of ligands in different GPCRs that depend on the amino acids side chains exposed in the binding site (52). Although many methods described above can be used to screen individual structures or enumerated libraries for potential biologic activity, they offer, with exception of the generalized binding site models, no guidance for the de novo design of new scaffolds that have an increased potential for biologic activity on a range of targets. One of the earliest concepts to offer such guidance is based on the observation, that common core scaffolds exist, which can be differentiated by modification of its side chains into ligands that are individually selective for different target proteins. Evans et al. (53), who made this observation in the case of the benzodiazepine core contained in selective ligands, proposed that the limited number of protein folds in the pro- teome also leads to a limited number of ligand-sensing cores. A ligand-sensing core is defined as the folding pattern of the protein in a sphere with a diameter of 20–30 Å around the binding site without taking the individual protein side chains into account (57). Different side chains in the ligand-sensing core can lead to a variety of diversely functionalized binding cavities, which may fulfill different functions and may occur in more than one target family. A privileged structure might be a suitable scaffold that orients its side chains in different regions of the binding pocket that is defined by the ligand-sensing core. Depending on the amino acid side chain, which is exposed in the binding pocket by the individual protein, different functional groups are required to be present on the privileged scaffold. The concept of biologically-oriented synthesis (BIOS (58)) suggests that in absence of detailed knowledge of an individual target protein’s structure, it is required to screen a diversely-functionalized library around a privileged or biologi- cally pre-validated scaffold for the ligand sensing core in order to identify the correct substitution pattern for the individual target.

Natural products

Natural products are a traditional source of biologically active compounds, which are used either as drugs themselves or have inspired the discovery of synthetic drugs (59) (see the articles “Natural Products: An Overview” and “Natural Products to Probe Biosynthetic Pathways”). They cover a wide range of chemical classes (59), and they are expected to be fine-tuned by evolution to fulfill a purpose that is often still unknown, but likely to involve the interaction with biomolecules. The capability of an organism to produce many variations of a metabolite at a low effort is considered a beneficial evolutionary trait of a species, which allows it to adapt its range of produced metabolites quickly to a changing environment. For this reason, the metabolic pathways involved in the synthesis of natural products are often branched in phyto and lead to high chemical diversity in collections of natural products (60). The synthetic complexity of natural products makes them often difficult to optimize. It has been shown, however, that not only natural products themselves, but also synthetic libraries of simplified analogs that retain only the key features of the original natural product can be applied successfully in screens for biologically active compounds. It has been demonstrated that from such a simplified natural product, core selective ligands for different targets can be derived and may be regarded as privileged structures (58).

Screening Processes and Strategies

The next question is how to make the best use of a library assembled according to these design principles. The goal of screening is to identify as many actives as in the screening library possible and to characterize them by measuring dose re- sponse curves, confirming both the chemical identity and the purity of the sample. Part of a thorough characterization of the hits is also running the the counter assays necessary to confirm that the observed activity results from the interaction of the ligand with the target and is not the result of an artifact caused by a technology incompatibility of the compound. These tasks must be executed with the lowest possible experimental effort, reagent costs, protein, and compound consumption. Al- though in basic research it may be sufficient to identity a limited number of tool compounds, in pharmaceutical industry it is of interest to identify as many of the active chemical series present in the screening library as possible and to establish intellectual property, because the attrition in the next steps of drug discovery is high.

HTS processes

Typically, the screening process (see the article “High Through- put Screening (HTS) Techniques: Overview of Applications in Chemical Biology”) begins with the production of stock solutions by dissolving powder samples and reformating the solution samples into a uniform deck of stock solution plates. These samples are then stored under controlled conditions, and from these samples the screening plates are produced by plate replication systems. Perhaps the most direct approach to screening is first to measure the dose response curves with the prefabricated assay plates that contain the compounds in the different concentrations ([Fig. 2a]). This technique has been shown to be feasible with a high level of automation for libraries up to the size of 100,000 samples. Because the same number of data points is measured
for active and inactive compounds, absence and presence of activity are determined with the same degree of reliability. This reliability is an advantage for building SAR models. In addition, the analysis of the dose response curve shapes allows some conclusion as to whether the interaction between ligand and protein is specific (61). However, in the pharmaceutical industry, larger libraries of a million or more compounds are often screened, and it is desirable to have a lower consumption of protein and compounds.
Virtual screening is a process that uses computational methods to predict the activity of compounds. The process requires the availability of a relatively small number of hits in experimental screening. Between the primary screening and validation, a confirmation screen can be performed in which, for the primary hits, the single concentration experiment is repeated, and only hits with confirmed activity are validated (Fig. 2b). In any case, it is necessary to confirm chemical identity and purity of the samples found active to avoid false positives and false negatives. For the same reasons, counter screens or secondary assays that use different read-out methods are performed to exclude an unspecific interaction of the compound with the assay system.

This process requires the capability to access large subsets of the screening library. Typically, this process step, called cherry picking, individual samples must be taken from the mother plates with stock solution and dispensed into plates for the confirmation or validation screen. In addition, dilution series for dose-response curve measurement must be produced of the cherry-picked samples for validation. Technically, cherry picking is a nontrivial task, and if all compounds with significant primary activity are to be confirmed, which can be several thousand compounds, and not only the screening capacity for confirmation screening needs to be available, but also the cherry-picking capacity for these samples must be available.

For large screening libraries, these processes can only be run with a high degree of automation in sample storage, cherry picking, screening, and chemical analytics. These automation steps must be driven by an informatics platform that tracks the contents of plates, collects the results of the different readers used for screening, and performs normalization, curve fitting, and detection of errors that may result from spillage and carry-over of compounds in the pipetting process or edge effects (62, 63). The results of these automated preprocessing steps must be presented to the screen in an appropriate visualization after each screening step for quality control and final decision making, pivotal for the primary screening and its results justify the follow-up of more compounds than can be processed, then chemoinformatics techniques such as clustering can be used to ensure appropriate representation of all chemical classes in the validation set. Also in this step, compounds can be removed that interfere with the assay technology and are unlikely to interact specifically with the target (64). The decisions taken at these steps must be handed over to the cherry-picking system for automated process. The software tools used for these different tasks must be well integrated to achieve a process that runs smoothly (65). Integration of in silico screening

Because of the large investments in the hardware and software infrastructure required for HTS, to replace the primary screening in HTS by virtual screening, followed by the validation of a relatively small number of hits in experimental screening is seen as a valuable alternative (Fig. 2c). However, this task is only feasible if either information about protein structure or reference ligands is available (33). In addition to the physically available in-house collection, virtual screening can include compounds from vendor catalogs and even enumerated virtual libraries from which the hit compounds are then purchased or synthesized. Compilations of screening compound catalogs exist both publicly, such as ZINC (compiled by the Shoichet laboratory at UCSF, San Francisco, http://zinc.docking.org), which contains docking ready 3D structures (66), or in the commercial sector such as ChemNavigator (Chemnavigator, San Diego, http://www.chemnavigator.com), which is linked to a sample procurement service. Several cases have been reported in which active ligands have been discovered successfully using such processes (67). However, if automated high-throughput experimentation is abandoned, then only small numbers of compounds can be validated (typically below 1000), whereas typical HTS setups allow the validation of a couple of more than 1000 compounds. Similar to physical HTS, but to a higher extent, virtual screening is affected by false positives and false negatives. In typical virtual screening, accumulation of more than 10% of the false positives in the top 10% ranking compounds is an excellent result that is almost never reached in practice (68, 69). Assuming an industrial HTS library of a million compounds, to validate a virtual screening hitlist that consists of only 1% of this library despite the inevitably high false-negative rate this will cause, requires HT experimentation. Data fusion of HT experimentation and virtual screening can be expected to compensate errors of each of the methods and to allow the validation of a significant number of hits. Virtual screening in this setup no longer has the purpose to save investment in HT experimentation, but to maximize the positively validated compounds over the whole process to identify as many true positive hits in the collection as possible to feed in the drug discovery process (Fig. 2d).

Sequential screening

Instead of screening the whole library in one batch, it has been proposed to screen an initial subset and use the screening results from this subset to train a statistical model to predict and to prioritize the remaining library. The remaining library is then screened, and the cycle of model building, prediction, and screening can be executed several times, which is referred to as sequential or iterative screening (70). Although this seems to be very attractive because it reduces the number of compounds that require screening, the multiple selection cycles lead to a longer overall screening time increasing the assay logistics effort. Together with the multiple cherry-picking and data processing cycles this may cause more effort than the savings from screening less compounds. In sequential screening it is necessary to choose an initial set. In the absence of reasonable knowledge for the selection of a focused subset, the initial set must be selected by diversity selection, whose limitations have been discussed above. In a compound collection that has been designed to avoid unnecessary redundancy by applying reasonable diversity selection, little can be gained by additional diversity selection. Any active compound class not represented reasonably by the initial screening set is unlikely to be recovered in the additional screening cycles, because the statistical models built on the screening results cannot make valid predictions for it. However, one can expect to identify
addition of actives in the series covered by the initial set. Recently, it has been demonstrated that screening 25% of a one million compound library selected as a diversity set based on full plates followed by one prediction and screening cycle offers a reasonable compromise between logistical efforts, numbers of compounds screened and hit series covered (Fig 2e (71)). When the screening cost per compound is high and dominates the logistics effort sequential screening can be expected to be beneficial, provided it is acceptable to identify only a limited number of tool compounds instead of as many hit series out of the library as possible.

Analysis, Reporting, and Visualization

Often, the number of validated hits in an HTS is so high that the results cannot be analyzed without computational assistance. The primary objective of HTS analysis is to identify chemical series, which are characterized by a common structural core that allows the exploration of the series in a joint chemical synthesis effort and aligns the structures of a series to derive SAR. Therefore, in HTS analysis, compounds are grouped by structure, which can be achieved either with clustering procedures, especially those based on maximal common substructures (72, 73) or with rule-based classifications (18, 39, 74). An advantage to the clustering methods is that they adapt themselves to the data set, whereas an advantage for the rule-based method is that they can be more resilient to changes in the composition of the screening library. In the beginning of HTS, hits were often prioritized by potency alone; however, nowadays it is recognized that a wide range of properties has an impact on the additional success of a chemical series. Therefore, each series must be annotated with all information relevant for their additional prioritization, such as potency and ligand efficiency on the screening target, activity on other targets, and calculated or measured physical chemical properties. It is also of interest to identify inactive compounds that contain the substructure of an active series. The data set generated in this way is very information-rich, which requires special visualization techniques that are able to display several properties simultaneously and interactively (75, 76). Because the data visualized results from different experiments or computational procedures, its aggregation is a nontrivial task, which requires highly modular and interactive (77). Because the data visualized results from different experiments or computational procedures, its aggregation is a nontrivial task, which requires highly modular and interactive computational chemistry. Rev. Comput. Chem. 2002;18:1–40. Bartlett PA, Entzeroth M, eds. 2006. RSC Publishing, Cambridge. pp. 137–159.

References

Chemical Libraries: Screening for Biologically Active Small Molecules


Further Reading


See Also

Combinatorial Libraries: Overview of Applications in Chemical Biology
Compound Handling
Computational Approaches in Drug Discovery and Development
High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology
Lead Optimization in Drug Discovery
Natural Products to Probe Biosynthetic Pathways
Natural Products: An Overview
Small Molecule Combinatorial Libraries
Small Molecule, Drug Target Databases
Target Family-Biased Compound Library: Optimization, Target Selection and Validation
Cyclooxygenase Inhibition, Recent Advances in the Mechanisms of

Melissa V. Turman, Anna L. Blobaum, Mary E. Konkle and Lawrence J. Marnett, A.B. Hancock Jr. Memorial Laboratory for Cancer Research, Departments of Biochemistry, Chemistry, and Pharmacology, Vanderbilt Institute of Chemical Biology, Center in Molecular Toxicology, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee
doi: 10.1002/9780470048672.wecb108

Cyclooxygenases (COXs) are key enzymes of bioactive lipid metabolism that catalyze the committed step in prostaglandin biosynthesis. Inhibition of COX catalysis is a principal mode of action of nonsteroidal antiinflammatory drugs (NSAIDs). The combination of structural and functional analysis of COX-inhibitor interactions has provided important insights into the molecular determinants of inhibition. Although COX protein structure does not change significantly on inhibitor binding, a surprising variety of binding modes and molecular interactions are observed with different NSAIDs. This article focuses on recent advances in our understanding of these interactions with emphasis on arylcarboxylic acid-based inhibitors and their derivatives.

Cyclooxygenases (COX-1 and COX-2) are bifunctional enzymes that carry out two sequential reactions in spatially distinct but mechanistically coupled active sites—the double dioxygenation of arachidonic acid to prostaglandin G_2 (PGG_2) and the reduction of PGG_2 to PGH_2 (Eq. 1) (1). A richidonic acid oxygenation occurs in the cyclooxygenase (COX) active site, and PGH_2 reduction occurs in the peroxidase (POX) active site. The CDXs are homodimers of 70 kDa subunits, and dimerization is required for structural integrity and catalytic activity (2). The first crystal structure solved for a COX enzyme (ovine COX-1 with the bound inhibitor flurbiprofen) revealed an asymmetric unit containing two identical monomers that exhibit extensive contacts in a large subunit interface; each subunit contains the inhibitor bound only in the COX active site (3). Although it has been assumed that both subunits are active simultaneously, recent work suggests that substrate or inhibitor binding in the COX active site at one subunit precludes the binding of another molecule at the other subunit (4). This cooperativity between subunits is consistent with the observation that CDX-2 dimers only bind a single molecule of flurbiprofen tightly, and this is sufficient to inhibit all COX activity (4). The molecular details of the communication between subunits are not yet understood, but their elucidation will have a significant impact on defining COX-inhibitor interactions.

Ligand Binding to COX Enzymes
Each monomer of COX consists of three structural domains: a short N-terminal epidermal growth factor domain, a membrane-binding domain, and a large, globular C-terminal catalytic domain (Fig. 3a) (3). The COX and POX active sites are located on opposite sides of the catalytic domain with the heme prosthetic group positioned at the base of the peroxidase.
site. The epidermal growth factor domain and catalytic domain create the dimer interface and place the two membrane-binding domains on the same face of the homodimer about 25 Å apart (5, 6). The membrane-binding domain of cyclooxygenase is composed of four amphipathic α-helices, with hydrophobic and aromatic residues that project from the helices to create a surface that interacts with one face of the lipid bilayer (3). Three of the four helices lie in the same plane, whereas the last helix (helix D) projects up into the catalytic domain (5). The catalytic domain constitutes the majority of the COX monomer and is the site of substrate binding and NSAID action.

Substrate and inhibitor gain access to the COX active site (Fig. 1b) at the base of the membrane-binding domain, which leads into a long hydrophobic channel that extends deep into the catalytic domain interior (3). This hydrophobic channel narrows at the interface between the membrane-binding domain and the catalytic domain to form a constriction composed of three residues (Arg-120, Tyr-355, and Glu-524) that separate the "lobby" from the active site. The active sites of COX-1 and COX-2 are nearly identical in the amino acid residues that constitute the primary shell of the active site but differ in the presence of a side pocket in COX-2 bordered by Val-523 and Trp-387. A small, hydrophobic pocket formed by Val-349, Ala-527, Ser-530, and Leu-531 surrounds the 2-aryl ring of COX-2 selective inhibitors (e.g., celecoxib and rofecoxib). In addition, the last helix of the membrane-binding domain (helix D) is positioned differently in COX-2 and shifts the location of Arg-120 at the constriction site, which allows for a larger solvent-accessible surface at the interface between the membrane-binding domain and the COX active site in COX-2.

**Time-Dependent Inhibition of COX by Indomethacin and its Ethanolamines**

Kinetic studies of COX inhibition reveal that many NSAIDs interact with COX through a multistep mechanism in which a rapid, reversible step is followed by one or more slow steps that may be poorly reversible (Eq. 2) (9, 10):

\[
E + I \overset{k_1}{\rightarrow} EI \overset{k_2}{\rightarrow} EI^* \overset{k_d}{\rightarrow} E + I^* \quad (2)
\]

Although dissociation of the E–I* complex is often too small to be measured, none of the NSAIDs other than aspirin cause covalent modification of COX enzymes. Rather, the final step of binding for many NSAIDs corresponds to the formation of a tightly bound and functionally irreversible protein–inhibitor complex, E–I*. Indomethacin (INDO) is a highly potent, time-dependent inhibitor of COX (10). Although its time-dependent kinetics were first reported over 30 years ago, the mechanism of its tight binding was not understood. A crystal structure of the complex of INDO with murine COX-2 provides a detailed view of the binding orientation and critical protein–ligand interactions (8). The carboxylic acid of INDO binds at the constriction site of COX, forming an ionic bond with Arg-120. The heterocyclic scaffold occupies the oxygene-nase channel, with the methoxy group projecting toward Val-523 and the para-chlorobenzoyl moiety projecting toward Tyr-385 and Trp-387. A small, hydrophobic pocket formed by Val-349, Ala-527, Ser-530, and Leu-531 surrounds the 2'-methyl group on the indole ring (Fig. 2b). The importance of the interaction with the 2'-methyl group was revealed by a combination of protein mutagenesis and inhibitor modification (11). Site-directed mutants of COX-2 were generated, which altered the size of the hydrophobic pocket via
Structures and inhibitory activities of indomethacin, sulindac sulfide, and their derivatives demonstrate the importance of the 2′-methyl group for inhibition of COX by indomethacin and sulindac sulfide. (a) Indomethacin bound in the active site of COX-2. The 2′-methyl group of indomethacin, shown with its Van der Waals surface, sits in a hydrophobic pocket that consists of Val-349, Ala-527, Ser-530, and Leu-531. Hydrogens are included on the residues of the pocket (V349A) did not significantly alter the potency of INDO. Although INDO exhibited diminished potency toward V349I, the binding kinetics for V349I were similar for wild-type COX-2 and all Val-349 mutants. Dissociation of this complex, as judged by k−2, is not measurable for wild-type COX-2, V349A, or V349I. However, a measurable and, in fact, rapid k−2 was observed with V349L. As a result, most of the inhibitor was competed off within a few minutes after the addition of arachidonic acid. This finding suggests that the reversibility of the second step in the time-dependent mechanism is the primary determinant of the potency of COX inhibition by INDO and that this is largely defined by interactions between the 2′-methyl group and the hydrophobic pocket.

A derivative of INDO that lacks the 2′-methyl group was synthesized, and its interactions with COX-1 and COX-2 were studied (Fig. 2b) (12). This compound, des-methyl INDO (DM-INDO), is a poor inhibitor of wild-type COX-1 and COX-2 and all Val-349 COX-2 mutants. Again, potency is dependent on the reversibility of the second step of the binding mechanism. Curiously, DM-INDO displays COX-2 selectivity, despite complete sequence conservation within the hydrophobic pocket in COX-1 and COX-2. Although selectivity may be related to the magnitude of k−2, which is an order of magnitude faster for COX-1 than for COX-2, the structural basis of selectivity of DM-INDO inhibition has not been elucidated.

Sulindac sulfide, the bioactive metabolite of sulindac, is structurally very similar to INDO and is a slow, tight-binding inhibitor of COX. As with INDO, removal of the methyl group from sulindac sulfide results in loss of COX-2 and COX-2 inhibition (14). However, it should be noted that the benzylidine double bond of des-methyl sulindac sulfide (DM-SS) exists in the E-conformation, whereas sulindac sulfide exists in the Z-conformer. Many clinically relevant NSAIDs exert off-target effects unrelated to their ability to inhibit COX enzymes. For example, INDO and SS induce apoptosis of tumor cells and modulate γ-secretase activity (15, 16). INDO also activates the nuclear transcription factor PPARγ (17). The complexity of in vivo pharmacologic effects makes it a challenge to separate the contribution of COX inhibition from other effects in a given pharmacologic response. Thus, the removal of COX inhibitory activity by a minor modification, such as the removal of a methyl group, provides an opportunity to dissect COX-dependent and COX-independent effects of certain NSAIDs. In fact, DM-INDO and DM-SS activate PPARγ in HCA-7 cells with dose responses similar to those of the parent drugs (14). Likewise, the des-methyl compounds exhibit potency similar to the parent compounds in their ability to induce apoptosis in RKO cells, a human colon cancer cell line, and to activate PPARγ in cellular reporter assays.

Gastrointestinal toxicity is a classic side effect of INDO and other potent, nonselective COX inhibitors. It is widely accepted that this toxicity arises from inhibition of PGE2 production in the gastric mucosa; however, other mechanisms have been proposed (15, 18). C57BL/6 mice were administered INDO or DM-INDO at 5 mg/kg, a dose above the LD50 of INDO. Whereas the gastric mucosa of INDO-treated animals exhibited significant tissue necrosis, the gastric tissue from DM-INDO-treated animals was identical to that of healthy control animals (14). Furthermore, the LD50 of DM-INDO in...
Cyclooxygenase Inhibition: Recent Advances in the Mechanisms of

C57BL6 mice was 20-fold higher than that of INDO, which suggests that toxicity of INDO is closely associated with inhibition of COX.

Understanding how a drug works on the molecular level is critical to elucidating its pharmacologic effects. INDO provides a striking case study of how subtle interactions, revealed by protein structure-function analysis, can be exploited to dissect the off-target effects of drugs. DM-INDO and DM-SS may provide scaffolds for probing or fine-tuning the beneficial side activities of the parent drugs, while reducing dose-limiting toxicity.

For many NSAIDs, including flurbiprofen and INDO, the ionic bond between Arg-120 and the carboxylic acid of the inhibitor is absolutely required for time-dependent inhibition of COX-1 (10, 19); the methyl esters of INDO and other NSAIDs were shown to be very poor inhibitors of COX-1 (10). However, this interaction is not a universal requirement for inhibition of COX-2 (20). Conversion of the INDO carboxylate group to neutral esters and amides provides a general and facile method for generating COX-2-selective inhibitors (21).

A notable exception is a series of α-substituted ethanolamides of INDO, which exist as R/S-enantiomeric pairs (22). Across a range of α-substitutions, the R-enantiomers are consistently COX-2-selective. However, the S-enantiomers efficiently inhibit both COX-1 and COX-2.

Crystal structures of the pair of α-ethyl-substituted enantiomers bound to COX-1 provide a surprising structural basis for the enantioselectivity of inhibition (23). The R-enantiomer (8) binds in a conformation similar to that of INDO bound to COX-1 or COX-2 (Fig. 3). The para-chlorobenzoyl group is oriented toward Tyr-385 and Trp-387, and the methoxy projects toward the side pocket, and the ethanolamide is positioned at the constriction site. The hydroxyl group of the ethanolamide forms a hydrogen bond to the guanidinium of Arg-120. To accommodate the hydroxyethyl substituent, Arg-120 must rotate away from Tyr-355 and toward Glu-524. The ethyl group extends through the constriction.

The S-enantiomer (9) adopts a very different conformation (Fig. 3). In this case, the methoxy group projects toward the apex of the channel, with the para-chlorobenzoyl wedged into a groove below Leu-531 and the 2′-methyl group positioned above Tyr-355. The ethanolamide moiety inserts into the side pocket, which consists of His-90, Gin-192, Leu-517, Phe-518, and Ile-523. The substituted ethanolamide of 9 makes several hydrophobic interactions with Phe-518 and Ile-523. The hydroxyl group can hydrogen bond to His-90 and Gln-192. In COX-2, the sulfone or sulfonamide moiety of the diarylhetarocycles use this side pocket, which is enlarged by conservative amino acid substitutions in the primary and secondary shells, and provide the basis for COX-2-selectivity of this class of inhibitors (7, 8). Notably, attempts to model 8 in a conformation similar to the noncanonical conformation of 9 indicate that the R-enantiomer cannot take advantage of the same network of interactions and could not avoid steric clashes with protein side chains.

The kinetics of 8 and 9 binding and inhibition with COX-1 reveal an interesting profile. Pre-steady state kinetic measurements indicate that binding of 8 and 9 proceeds by a multistep mechanism like that of INDO (Eq. 2). Although $K_i$, the ratio of $k_1$ to $k_2$, is similar for 8, 9, and INDO, formation of E–I* proceeds at a much faster rate (10-fold) for 8 and 9 than for INDO.
Cyclooxygenase-Inhibition, Recent Advances in the Mechanisms of

Inhibition by Aspirin and Salicylate

The development of aspirin in the late 1890s is a landmark in the evolution of the pharmaceutical industry (24). Many decades later, aspirin was shown to inhibit prostaglandin biosynthesis by covalently modifying COX (25, 26). Inactivation results from acetylation of Ser-530, which is located above Arg-120 and across from Tyr-385 in the COX active site (27). Mutagenesis of Tyr-385 to Phe in COX-2 reduces acetylation of Ser-530 by greater than 90%, which suggests the importance of hydrogen-bond formation in acetylation (28). Indeed, a crystal structure of COX-1 acetylated by bromoacetyl salicylic acid reveals a hydrogen bond between Tyr-385 and the carbonyl oxygen of the bromoacetyl group (29). It has been proposed that hydrogen bonding by Tyr-385 stabilizes the incipient anion state of the heme that oxidizes Tyr-385 to a tyrosyl radical in the active enzyme. This reduction seems to be caused by the peroxide oxidation of the heme prosthetic group (and presumably the oxidation of Tyr-385) because substitution of Mn3+ protoporphyrin IX for heme eliminates hydroperoxide antagonism of salicylate and acetaminophen inhibition (31, 32). Hydroperoxide-antagonism of COX inhibition may be related to the purported differential sensitivity of COX-1 and COX-2 to certain NSAIDs because COX-1 requires approximately a 10-fold higher concentration of peroxide than COX-2 to achieve full activation (33, 34).

Acetylation of COX-1 completely abolishes cyclooxygenase activity presumably by blocking the binding of arachidonic acid in the COX active site. However, acetylated COX-2 can oxidize arachidonic acid to a 15(R)-hydroperoxy derivative (35). This ability is because of the greater size of the COX-2 active site. Mutagenesis of the side pocket residues, Val-523, Arg-513, and Val-434, to the COX-1 variants (Ile-523, His-513, and Ile-434) eliminates the ability of acetylated COX-2 to oxidize arachidonic acid (36). This finding indicates that acetylated COX-2 can accommodate arachidonic acid in its active site in a binding mode that allows oxygenation at carbon-15 to 15-hydroperoxyeicosatetraenoic acid. This finding may be pharmacologically important because 15-hydroperoxyeicosatetraenoic acid is a substrate for oxygenation by lipoxygenases to lipoxin A4 (37). Lipoxin A4 and related compounds have potent antiinflammatory activity, and the contribution to its formation by acetylated COX-2 may account for some of the antiinflammatory activity exhibited by aspirin.

Diclofenac and Lumiracoxib: Novel Mechanisms for Inhibition and COX-2 Selectivity

The importance of Tyr-385 in aspirin acetylation of Ser-530 suggests that it may play a broader role in interacting with electron-rich centers. In fact, the crystal structure of a catalytically inactive mutant of COX-2 with bound arachidonic acid reveals the arachidonate bound in an inverted, noncanonical orientation in which its carboxylate group is hydrogen-bonded to both Tyr-385 and Ser-530 (38). This conformation is not a productive conformation for arachidonate oxygenation, but
or leucine alter the size of the pocket and lead to changes in lumiracoxib and diclofenac exhibit large differences in the selectivity of COX-2 inhibitors (e.g., celecoxib and rofecoxib) and does not contain structure of the diarylheterocycle class of COX-2 selective inhibitors. Lumiracoxib lacks the tricyclic short plasma half-life (41, 42). Lumiracoxib displaces a 500-fold greater selectivity for COX-2 than COX-1 with its carboxylic acid moiety coordinated to Ser-530 and Tyr-385 at the top of the active site and not ion-paired nor hydrogen-bonded to Arg-120 and Tyr-355 (39). This binding mode seems to be the binding mode responsible for inhibition because mutation of Ser-530 to Ala abolishes diclofenac inhibition whereas mutation of Arg-120 to Gin is without effect (39). Interestingly, diclofenac also seems to make use of the small hydrophobic binding pocket composed of Ala-527, Val-349, Ser-530, and Leu-531 that INDO uses to insert its 2′-methyl group. Mutations of Val-349 in this pocket to alanine or leucine alter the size of the pocket and lead to changes in the potency of indomethacin and diclofenac (11, 40).

A close structural analog of the non-selective COX inhibitor diclofenac, lumiracoxib displays a 500-fold greater selectivity for COX-2 than COX-1 in vivo and exhibits a unique pharmacologic profile that includes rapid absorbance and a relatively short plasma half-life (41, 42). Lumiracoxib lacks the tricyclic structure of the diaryheterocycle class of COX-2 selective inhibitors (e.g., celecoxib and rofecoxib) and does not contain a sulfonamide or sulfone group. Although structurally related, lumiracoxib and diclofenac exhibit large differences in the selectivity of COX-2 inhibition, and the molecular basis for this difference in the selectivity of COX inhibition only recently was elucidated.

As expected from its structural resemblance to diclofenac, lumiracoxib binds to COX-2 in an inverted orientation similar to that of diclofenac with hydrogen-bonding interactions between the carboxylate of the inhibitor and Ser-530/Tyr-385 at the top of the active site (43). A comparison of this crystal structure with a model of lumiracoxib bound to COX-1 leads to the conclusion that the COX-2 selectivity of lumiracoxib arises from the insertion of the methyl group on the phenylacetic acid ring of the inhibitor into a small groove provided by the movement of a primary shell leucine residue (Leu-384) in the COX-2 active site. The movement of Leu-384 is thought to be restricted in the active site of COX-1 because of the presence of bulky secondary shell residues lying behind it (Ile-525 and Phe-531) that prevent the movement of Leu-384 with inhibitor bound.

Prior structure-activity studies with diclofenac analogs indicated that methyl or chlorine substituents on the lower aniline ring of the inhibitor in the ortho position are required for potent inhibition of COX (44). Analogs with halogen substitutions (fluorine or chlorine) at the 5′-position of the phenylacetic acid ring demonstrate an even higher potency for COX inhibition (44).

Although the selectivity of lumiracoxib for COX-2 has long been demonstrated in vivo, the chemical and structural basis for the balance that exists between potency and COX-2 selectivity was determined only recently for this inhibitor (40). An extensive structure-activity analysis with a library of lumiracoxib analogs indicates that the 5′-methyl group on the phenylacetic acid ring of the inhibitor is the major determinant for COX-2 selectivity, and that the chemical nature of the substituents in the ortho positions on the lower aniline ring exert the major influence on the potency of COX inhibition (2,6-dichloro, 2,6-dimethyl, or 2-chloro-6-methyl substitutions are preferred). Interestingly, the structure-activity study also suggests a contributory role for the chlorine atom of lumiracoxib in COX-2 selectivity, a key finding not divulged by the crystal structure determination. Site-directed mutagenesis of the small hydrophobic binding pocket valine residue (Val-349 to Ala, Ile, or Leu) in COX-2 also supports a role for lumiracoxib binding interactions in this region of the active site (40).

The resultant structure-activity analysis of lumiracoxib analogs with the V349 mutants indicates a preference for a chlorine substituent on the lower aniline ring for potent inhibition of COX-2, with F, Cl, CI-CI, or CI-CH3 substitutions allowed at the 2,6 position to maintain potency. These data suggest that the chlorine atom on lumiracoxib (similar to diclofenac) binds in the small hydrophobic pocket that consists of V349, S530, A527, and L531 and that this interaction may contribute to the potent inhibition of COX.

The structure-activity studies with diclofenac and lumiracoxib reveal that the COX-2 inhibitor activity of lumiracoxib results from a fine balance between potency and selectivity. The principal determinant of potency is the dihaloarylamine ring, whereas the meta-methylphenylacetic acid ring primarily controls selectivity. An F → Cl substitution in the dihaloarylamine ring of lumiracoxib increases the tightness of binding (potency of
Cyclooxygenase Inhibition: Recent Advances in the Mechanisms Of


References


2. Xiao G, Chen W, Kilmac R. Effect of substrate, especially at elevated concentrations of arachidonic acid, on aspirin-induced COX-2 selectivity but reduces potency. In kinetic analyses, understated differences exist in side-chain conformations, displacement of main-chain protein atoms, and kinetics of inhibitor binding. Different inhibitors use different sets of interactions to optimize their binding in the COX active site, which can afford the selectivity and potency of inhibition in vivo. The interactions of a variety of selective and nonselective inhibitors with COX-1 and COX-2 have been probed by a combination of crystallography, mutagenesis, and chemical modification. Although the macromolecular protein structure is essentially the same in all COX-inhibitor complexes, seemingly different interactions exist in side-chain conformations, displacement of main-chain protein atoms, and kinetics of inhibitor binding. Different inhibitors use different sets of interactions to optimize their binding in the COX active site, which can afford significant perturbations of selectivity and potency (4S). This situation gives rise to a striking variety of bound orientations that illustrate the impact of subtle interactions on ligand-protein association. Problems remain to be solved with this fascinating class of enzymes that should provide additional insights into the potency and selectivity of different classes of inhibitors. The challenge will be to use this information to create novel classes of inhibitors or to address the pharmacological deficiencies of extant inhibitors (e.g., gastrointestinal toxicity, cardiovascular toxicity, and renal toxicity).

© 2008, John Wiley & Sons, Inc.
Cyclooxygenase Inhibition: Recent Advances in the Mechanisms of


Forward Chemical Genetics

Young-Tae Chang, National University of Singapore and Singapore Bioimaging Consortium, A*STAR, Singapore
doi: 10.1002/9780470048672.wecb178

With the successful results of the Human Genome Project, we are now faced with the problem of handling numerous gene targets whose functions remain to be studied; this challenge is being undertaken by the field known as functional genomics. Chemical genetics is an emerging new paradigm to attack this challenging problem, of which there are two approaches: forward and reverse. Although reverse chemical genetics uses a "cause-to-effect" approach, forward chemical genetics uses an "effect-to-cause" approach. As opposed to conventional genetics where gene knock-outs or overexpression, forward chemical genetics uses a small-molecule library to produce a novel phenotype that eventually is employed in the elucidation of gene function. Compared with conventional genetics, chemical genetics holds several unique advantages. A successful forward genetic study provides not only knowledge about a novel gene's function, but also it provides a small molecule on/off switch that can regulate biologic processes. These small molecules will be extremely useful biologic probes as well as potential new drug candidates. Three main components make up forward chemical genetics: 1) chemical toolbox generation, 2) phenotypic screening, and 3) target identification. Although all three components require additional refinement, target identification poses the greatest challenge. Herein, the general concepts of chemical genetics, with a focus on the forward approach, and the technical overviews for each component are described.

Chemical genetics, in the simplest of terms, can be defined as a "genetics" study using "chemical" tools (1). Elucidating the function of every gene from the sequence data of tens of thousands of genes (so-called functional genomics) is the next major step for the human genome project. Geneticists conventionally have investigated the function of unknown genes by comparing the normal phenotype with that of knockouts, or through the overexpression of target genes. A novel approach using chemical library screening to find interesting phenotypic changes through the targeting of a specific gene product, a protein, has emerged as an alternative tactic—so-called chemical genetics. In chemical genetics, one chemical compound may specifically inhibit or activate one (or multiple) target proteins. Therefore, the compound is equivalent in conventional genetics to a gene knockout or to the overexpression of a gene. As in classic genetics, chemical genetics is divided into two approaches: forward and reverse.

Genetics and Chemical Genetics
Forward and reverse approach

Forward genetics (FG) operates "from effect to cause" or "phenotype (physically apparent characteristic) to genotype (genetic sequence)" and requires no specific gene target at the onset. It studies changes in phenotype(s) such as morphology, growth, or behavior resulting from random genomic DNA mutations or deletions induced from radioactive or chemical mutagenesis, and then it identifies the gene responsible through mutation mapping. Forward chemical genetics (FCG) mimics FG by substituting random mutagenesis with a collection of a library of typically unbiased (not targeted) compounds as protein function regulators in place of mutagens (2). The first step in both FG and FCG is to screen for changes induced by either the inhibition or the stimulation of a protein’s function, and both go on to identify the genetic cause but in different ways. FG goes after genetic mutation, which is a permanently retained marker, but FCG needs to identify the protein partner for the small molecule. This target identification is one of the greatest challenges in chemical genetics (3).

With the help of molecular biology techniques, reverse genetics (RG) was a later development in genetics and operates "from cause to effect" (genotype to phenotype) (4). Reverse genetics begins with selecting a gene of interest, manipulating it to produce an organism harboring the mutated gene, and characterizing the phenotypic differences between the mutant and the
lethality, thus providing a partial knockout phenotype (13). Allows the use of sublethal doses of the ligand and avoids full gene because the deletion may be lethal (12). Chemical genetics others intact (11). Additionally, if a gene is essential for survival and dissect particular functions of that protein while leaving multiple functions, and chemical genetics can potentially isolate the protein (10). It is always possible that one protein may have from those that develop from merely a particular function of protein entirely from the organism (9). Therefore, it is difficult to determine the effects that develop from the deletion separate from those that develop from merely a particular function of the protein (10). It is always possible that one protein may have multiple functions, and chemical genetics can potentially isolate and dissect particular functions of that protein while leaving others intact (11). Additionally, if a gene is essential for survival or development, a total knockout such as in classic genetics, may abolish the chance to study the later stage function of that gene because the deletion may be lethal (12). Chemical genetics allows the use of sublethal doses of the ligand and avoids full lethality, thus providing a partial knockout phenotype (13).

Advantages of forward chemical genetics

In FCG, the starting point is a selection of the protein of interest using previous information. The chemical libraries are mainly tested for only one selected protein target. This article is a more focused study of the known (most likely well-validated) target protein usually focused on improving on the chemical probes, which in a sense narrows down the scope of the work. In contrast, FCG studies whole cells or organisms, and thus, the compounds are screened against multiple potential targets simultaneously. A successful FCG will identify a novel gene product (target protein) and its on/off switch, the small-molecule complement. Therefore, FCG promises an efficient “two-birds-with-one-stone” approach. However, the unpredictable pharmacologically usable results have kept pharmaceutical companies from making a full commitment to FCG, leaving the field to be developed by the academic community thus far.

Compared with classic genetics, FCG offers several advantages and provides access to previously unstudied biologic space. Use of chemical tools offers greater ease and flexibility than does classic genetic modification. Classic genetic techniques are relatively difficult to employ especially in mammals because of their diploid genome, physical size, and slow reproduction rate. On the other hand, FCG studies may be conducted on any complex cellular or animal models without the need for any time-consuming genetic modifications that may prove lethal or in which the cell (animal) can mask the phenotype through related gene functional compensation for the mutation. Especially important is FCG’s promise in operating in the relevant context of human cells at physiologic conditions that has strained traditional genetics techniques. Therefore, FCG fills a major gap in genetic studies where no, or suboptimal, model systems exist. Additionally, chemical genetics also allows for the possibility of “multiple knockouts” by adding multiple specific ligands, which is a situation often described as a “nightmare” for the geneticist (8).

Second, classic genetic knockouts, in principle, delete the protein entirely from the organism (9). Therefore, it is difficult to determine the effects that develop from the deletion separate from those that develop from merely a particular function of the protein (10). It is always possible that one protein may have multiple functions, and chemical genetics can potentially isolate and dissect particular functions of that protein while leaving others intact (11). Additionally, if a gene is essential for survival or development, a total knockout such as in classic genetics, may abolish the chance to study the later stage function of that gene because the deletion may be lethal (12). Chemical genetics allows the use of sublethal doses of the ligand and avoids full lethality, thus providing a partial knockout phenotype (13).

Another advantage of chemical genetics is real-time control. Chemical genetics allows for this control by the ability to introduce rapidly a cell-permeable ligand at any stage that may yield the desired phenotype as quickly as diffusion-limited kinetics will allow. The chemical genetics inducer-protein effect is a “switch” that can turn the event under study ON or OFF in real-time and allows for kinetic in vivo analysis, which is something not usually possible in classic genetics. A temporal control is available in classic genetic studies through conditional alleles, such as temperature-sensitive mutations, they often have unwanted broad side-effects that may interfere with the desired result (4). The antisense oligonucleotide and RNA interference (RNAi) are other popular alternatives for conditional knockouts (10) that work by inhibiting the synthesis of the target protein from mRNA. However, as their effects are delayed until all existing proteins are degraded, they are particularly ill-suited to time-sensitive studies, such as signal transduction, that occur on the milliseconds-to-hours time scale.

Chemical genetics and classic genetics are techniques that compliment each other well (14). One of the greatest advantages of classic genetics is the incredible specificity of a gene knockout. Although some chemical ligands can be specific switches with specificity approaching that of a gene knockout’s, the low specificity of many ligands often give “off-target” effects in which the probe may interact with proteins other than the protein(s) targeted. This low specificity makes defining specific protein functions very difficult because these off-target effects may lead to toxicity or false or unwanted positive/negative biologic results. In addition to this lack of specificity, chemical genetics cannot yet match the generality of genetics. Geneticists can, in theory, “knockout” any gene as long as the genomic sequencing is completed in the desired species, which is an ability that at this point exists as nothing more than a dream for the chemical geneticist (8). These situations are the perfect place for the integration of chemical and classic genetics (15).

Importantly, unlike drug development where specificity is tantamount, in chemical genetics, ligands need not be completely specific, so long as they give an identifiable phenotype that allows for the deciphering of the target protein’s function and that its side effect are relatively small. Whereas one may desire compounds with affinity in the subnanomolar range capable of producing the desired effect, in reality, compounds of low micromolar affinities are often accepted as good-to-reasonable candidates in chemical genetics (3). Despite this issue, and with an understanding of the necessary medicinal chemistry follow-up modifications and studies required, chemical genetics still has the advantage of immediately offering a potential drug lead, rather than simply a target gene or protein, as in classic genetics. Interestingly, a lead compound developed in drug discovery that may not possess pharmacokinetic properties suitable for therapeutic purposes may still be used as a probe in chemical genetics studies (16). In fact, the lower pharmacokinetic property requirements for chemical genetics probes compared with drugs allows for the use of a greater variety of functional groups and for a maximization of the chemical space in library constituents (15).
Components of Forward Chemical Genetics

With the advent of combinatorial library techniques that facilitate the synthesis of a large number of molecules (Fig. 1a, in combination with the success of HTS [high-throughput screening]), a large number of compounds can be easily and rapidly screened to discover a novel small molecule (hit compound). Once the hit compound is selected in a cellular or organism system (Fig. 1b), the next step is to identify the target protein and the biochemical pathways involved. An on-bead affinity matrix or tagged molecule (photoaffinity, chemical affinity, biotin, or fluorescence), which is obtained through the modification of the lead compound (Fig. 1c), is commonly used in the identification of the target protein. The protein is identified by “fishing out” the target through the exploitation of the binding affinity of the proteins toward the immobilized molecule (Fig. 1d), followed by gel separation (Fig. 1e) and mass-spectrometry. Some small molecules and their target proteins identified FCG thus far are summarized in Table 1 (16–46). In summary, the forward chemical genetics procedure is composed of 1) preparation of chemical toolboxes, 2) phenotypic screening, and 3) target identification and validation.

Chemical toolboxes

Collections of compounds, so-called libraries, are the absolute starting point for any chemical genetics study. These library compounds can be either natural products or synthetic compounds with various designs.

Natural products

Natural products have evolved alongside the various life forms on earth. Each surviving natural product may have its own reason for existence in the context of biologic function, so-called, biologically validated. That is why many natural products have been used for medicinal purposes for so long. For example, colchicine has been used as a drug for over 2000 years. Therefore, it is generally accepted that collections of natural products have a higher probability of delivering pharmacologically useful compounds than a typical synthetic combinatorial library, (9, 47). Natural products are obtained from sources such as plants, soils, and marine sponges. The most difficult in natural product discovery is isolating any active components out of the whole mixture extract. The typical and time-consuming isolation route is known as bioassay-guided purification. The purification procedure involves iterative processes in which compounds undergo multiple rounds of extraction/chromatography guided by the screening results of the successive crude extracts. These studies are hampered by overlooking potentially highly active low-abundance compounds, the cytotoxicity of one component masking the desired effect of another component, and bioactivity resulting from complicated synergistic effects (2, 48). In addition, determining the structure of the compound is another laborious and challenging task. Over the past decade and a half, these drawbacks led to a retreat from natural products by most pharmaceutical companies. However, the disappointment in the number of drugs originating from combinatorial chemistry and the continued benefits of natural products are luring many companies back to them. Currently, major pharmaceutical companies are re-embracing natural products and natural product-like libraries. Natural products need not have their use restricted to therapeutics, but they can be of great value to chemical genetics studies. Extremely exciting is the integration of traditional natural products and their scaffolds with modern combinatorial and HTS tools. Perhaps success will lie in a balance of the old and the new.

Natural product-like compounds

Natural product-like libraries offer a highly desirable middle ground between those who seek powerfully bioactive compounds from natural libraries and those who seek the ease of synthesis found in libraries composed of small organic molecules. Natural products are typically chiral, are extremely complex, and contain many stereogenic centers. These structures are often highly potent and serve as attractive leads for drug development (49). Natural products can be considered privileged structures in a biologic context and an excellent starting point for library design with a high probability for biologic activity. These compounds often contain sets of related and homologous pharmacophoric groups throughout families of natural compounds (50). Natural product-like libraries are those collections of compounds whose structures are based on or share high structural homology with natural products. These libraries may be designed to generate derivatives of a natural product scaffold. In addition, some have sought to generate natural product-like libraries not to improve on known activity but to expand a molecule’s functionality into a previously unknown area of biologic space (51). The synthesis of natural product-like libraries,
<table>
<thead>
<tr>
<th>Small molecule</th>
<th>Compound source</th>
<th>Library size</th>
<th>Activity</th>
<th>Target protein</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encephalazine</td>
<td>DOS of Triazine</td>
<td>1,536</td>
<td>Inhibition of brain-eye development in zebrafish</td>
<td>Ribosomal subunit proteins (S5, S13, S18, L28)</td>
<td>(16)</td>
</tr>
<tr>
<td>Small molecule inhibitors of rapamycin (SMIRs)</td>
<td>Commercial</td>
<td>16,320</td>
<td>Inhibition of rapamycin activity</td>
<td>Tmplp (PTEN) &amp; Ybr077cp</td>
<td>(17)</td>
</tr>
<tr>
<td>Splitomicin</td>
<td>Commercial</td>
<td>6,000</td>
<td>Yeast growth inhibition</td>
<td>Sir2p (HDAC)</td>
<td>(18)</td>
</tr>
<tr>
<td>SFK1</td>
<td>Commercial and in-house small-molecule library</td>
<td>100,000</td>
<td>Suppression of FK506 activity</td>
<td>Par1p (VDAC)</td>
<td>(19)</td>
</tr>
<tr>
<td>SFK2,3,4</td>
<td>Commercial and in-house small-molecule library</td>
<td>100,000</td>
<td>Suppression of FK506 activity</td>
<td>Ald6p (aldehyde dehydrogenase)</td>
<td>(20)</td>
</tr>
</tbody>
</table>
Forward Chemical Genetics

DOS 1,600 Auxin signal increase, Sirtuin (21, 22)

DOS Single compound

DOS Single compound

Sirtinol Bi-Blas alcohol dehydrogenase inhibition in plant

Brassinazole Dos Single compound BR biosynthesis inhibition in plant

DWF4 (cytochrome P450) DWF4 (cytochrome P450)

Sirtinol Known inhibitor (vascular endothelial growth factor inhibitor)

PTK787/ZK222584 Single compound Block blood vessel formation, VEGFR inhibitor (24)

> 100 Slow embryonic development in zebrafish (25)

Tubulyzine A, B, C Tubulyzine A, B, C

Tubulyzine A, B, C Tubulyzine A, B, C

Tubulyzine C Tubulyzine C

Sirtinol Known inhibitor (vascular endothelial growth factor inhibitor)

DOS 1,600 Auxin signal increase, Sirtuin (21, 22)

Brassinazole Dos Single compound BR biosynthesis inhibition in plant

DWF4 (cytochrome P450) DWF4 (cytochrome P450)

Sirtinol Known inhibitor (vascular endothelial growth factor inhibitor)

PTK787/ZK222584 Single compound Block blood vessel formation, VEGFR inhibitor (24)

> 100 Slow embryonic development in zebrafish (25)

Tubulyzine A, B, C Tubulyzine A, B, C

Tubulyzine A, B, C Tubulyzine A, B, C

Tubulyzine C Tubulyzine C

Tubulyzine A, B, C Tubulyzine A, B, C

Tubulyzine C Tubulyzine C

Sirtinol Known inhibitor (vascular endothelial growth factor inhibitor)
<table>
<thead>
<tr>
<th>Small molecule</th>
<th>Compound source</th>
<th>Library size</th>
<th>Activity</th>
<th>Target protein</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP576</td>
<td>Commercial</td>
<td>50,240</td>
<td>SARS-CoV inhibition</td>
<td>SARS-CoV protease</td>
<td>(26)</td>
</tr>
<tr>
<td>HE802</td>
<td>Commercial</td>
<td>50,240</td>
<td>SARS-CoV inhibition</td>
<td>SARS-CoV helicase</td>
<td>(26)</td>
</tr>
<tr>
<td>Myoseverin</td>
<td>DOS of purine</td>
<td>Single compound</td>
<td>Myotube disassembly</td>
<td>tubulin</td>
<td>(27, 28)</td>
</tr>
</tbody>
</table>
MP C11

DOS of Triazine 1.536 Pigmenting albino melanocyte

Mitochondrial F1F0-ATPase

Melanogenin

DOS of Triazine 1.170 Pigmenting melanocyte

Prohibitin

Syntol A

Commercial 16,320 Mitosis perturbation tubulin

HDAC6 (32)

Tubacin

DOS of 1,3-dioxane library 7,392 Inhibition of β-tubulin acetylation

7,392 Inhibition of β-tubulin acetylation

Hoechst A

DOS of Triazine 1,536 Pigmenting albino melanocyte
<table>
<thead>
<tr>
<th>Small molecules</th>
<th>Compound source</th>
<th>Library size</th>
<th>Activity</th>
<th>Target protein</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monastrol</td>
<td>Commercial</td>
<td>16,320</td>
<td>Block mitosis</td>
<td>Mitotic kinesin Eg5</td>
<td>(33)</td>
</tr>
<tr>
<td>ICG-001</td>
<td>DOS</td>
<td>5,000</td>
<td>Downregulation of β-catenin</td>
<td>cAMP response element-binding protein</td>
<td>(34)</td>
</tr>
<tr>
<td>TWS119</td>
<td>DOS</td>
<td>&gt; 100,000</td>
<td>Neuronal differentiation</td>
<td>GSK-3β</td>
<td>(35)</td>
</tr>
<tr>
<td>PNRJ-299</td>
<td>DOS</td>
<td>12</td>
<td>AP-1 inhibition</td>
<td>REF-1</td>
<td>(36)</td>
</tr>
</tbody>
</table>
Hh-Ag 1.1

DOS 140,000 Inhibition of Smo Smo (37)

MOL-294

DOS of Triazine 1,120 Daf-2 insulin signaling in C. elegans (44)

Aminopurvalanol

Commercial 16,320 Inhibition of actin assembly

NF-κB inhibition thioredoxin (45, 46)

Wiskostatin

GAPDH (44)

DOS Not specified

Inhibition of actin assembly
<table>
<thead>
<tr>
<th>Small molecules</th>
<th>Library size</th>
<th>Activity</th>
<th>Target protein</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diminutol</td>
<td>DOS of purine library 150</td>
<td>Cell cycle arrest at G2/M</td>
<td>CDK1 (cyclin dependent kinase)</td>
<td>(40)</td>
</tr>
<tr>
<td>Ubistatin A, B</td>
<td>DOS of purine library 1,561</td>
<td>Inhibition of mitotic spindle assembly</td>
<td>NQO1 (quinate oxidoreductase)</td>
<td>(41)</td>
</tr>
</tbody>
</table>
Diversity-oriented synthesis

After the combinatorial boom of the 1990s, a well-documented disappointment in the number of quality leads has set in (52). The compounds coming out of early combinatorial libraries simply did not perform up to the designer’s expectations. Guidelines such as Lipinski’s rules are helpful for generating more drug-like small molecules, but it has been argued that compounds from solid-phase organic synthesis may not be chemically diverse enough to generate the desired selectivity and potency (53).

Diversity-oriented synthesis (DOS) is a new term for a method of library construction in chemical genetics. Advocates of the DOS approach point to the archetypical case of a flat aromatic or heterocyclic core dotted with various diverse appendages that have largely failed to deliver the promised drugs. The discrepancy between the success of natural products and the traditional combinatorial libraries’ as yet disappointing rate of lead generation has forced the question of whether natural products occupy regions of chemical space evolutionarily fine-tuned to be the most potent and active compounds (54). Thus, one of the stated goals of DOS is to create extremely diverse libraries populating the maximum amount of chemical space in order to explore the greatest amount of biologic space.

DOS draws its name from its contrast to the traditionally used approach of target-oriented synthesis (TOS). Beginning with a known target (often natural product), TOS uses retrosynthetic analysis to work backward from a complicated product to simple and available starting materials. The goal of TOS is to get to a precise region of chemical space—a single target or a few closely related derivatives. Retrosynthetic analysis is not applicable to DOS because no target structure is available, and thus, the library cannot be targeted. Therefore, new thinking and planning strategies are required, and the idea of forward synthetic analysis has been proposed, which moves in the direction of simple to complex, or reactants to products, in contrast to TOS. DOS strategies focus on maximizing diversity by using branched and divergent pathways where the products of one reaction are common substrates for the next and where any follow-up chemistry must be efficient and systematic (55, 56). Molecules that occupy a greater degree of 3-D space, or those that are more globular/spherical, are a popular goal of DOS, as opposed to the relatively flat or circular molecules traditionally used in combinatorial chemistry (54).

Tagged libraries

Tagged libraries represent unique opportunities in library design (57). In this approach, libraries may be synthesized by any means and designed around any type of scaffold, but they must contain some functional tags integrated into the library. These tags are incorporated into the library compounds from the beginning, and they bestow some additional function into the molecules. The most common example of a tag is a fluorophore, but many more variations exist that allow the compounds to have functions extending beyond their biologic activity that may aid in areas such as target identification or ligand assembly. Peptide nucleic acids (PNAs) are oligonucleotide-like molecules that have their DNA backbone removed and reattached with an achiral polyamide backbone that can hybridize with DNA through strand displacement (58). PNA tags have been used in the development of tagged libraries that allow for the spatially addressable localization and identification of probes on a DNA microarray.

Click chemistry offers unique tags that can be used to generate library members through strategies such as target-guided ligand assembly. Sharpless developed click chemistry as a rapid reaction that proceeds in a short amount of time through “spring-loaded” highly exothermic, irreversible reactions to form carbon–hetero bonds (59). One of the most common bonds is a (2 + 3) Huisgen 1,3-dipolar cycloaddition between an azide and an alkyn moiety, which is described as an energetically “near-perfect” reaction. Click chemistry’s main advantage is the swift and clean nature of these reactions that can greatly assist library synthesis and lead discovery. Click chemistry–based tags have also been incorporated into studies of activity-based protein profiling (ABPP) (60). ABPP is a chemical genetic method that uses probes to monitor and visualize changes in protein functions/levels in the cell, especially changes that occur on a posttranslational level. In a traditional ABPP study, the probe carries a tag/reporter, such as biotin or fluorescence, whose size and physical properties may adversely affect probe-target interaction, cellular uptake, and probe distribution. In the click chemistry–based approach, the probe is functionalized with a tag bearing a simple azide moiety, which prevents much less of a disturbance in the system. After screening and binding to the protein target, an acetylene bearing reporter is added, and this facilitates the conversion of a hit compound to affinity matrix without the need for a time-consuming structure-activity relationships (SAR) study (44). The tagged strategy is applicable to any type of library scaffold and allows for rapid transfer from biologic screening to target identification.

Instead of tags being external to the active portion of the molecule, internally tagged fluorescent compound libraries were also demonstrated in a chemical genetic study. In many cases, the addition of an exogenous reporter or tag can alter the effect or cellular distribution of the original molecule. Tagging the molecule intrinsically avoids this problem. In this strategy, the probes being studied are all intrinsically fluorescent through their known fluorescent scaffold, and no additional tagging is required to visualize cellular localization (61) or to visualize a bound protein (62).

Dynamic combinatorial libraries

Traditional combinatorial libraries are synthesized primarily by parallel or split-pool techniques as static pools of discrete
molecules. Although using compound mixtures fell out of favor in combinatorial chemistry, they found renewed use in supramolecular chemistry in the form of dynamic combinatorial chemistry (DCL). The dynamism of DCLs results from the reversible interchangeability possible with their components. In these systems, every member of the library and the targets themselves affect all other members of the library, particularly in terms of library composition (63). Any stabilization of one member will result in an equilibrium shift and thermodynamic redistribution, by LeChatelier’s principle, of the library mixture favoring the best binder. Advantageously, library construction and screening can be combined in one step, because amplification of the best binder can be analytically detected. Therefore, the use of DCLs is as much an application of HTS as is library development. Whereas traditional combinatorial libraries rely on their sheer numbers, DCLs offer an alternative approach through self-replication and amplification.

The use of dynamic combinatorial libraries requires three steps: 1) the collection/design of a library capable of undergoing reversible constituent interchange, 2) conditions whereby the library may undergo interconversion, and 3) a step that selects the “fittest” binder and possibly involves its amplification. When a template or ligand is used to amplify the concentration of a member of the library, called “survival of the fittest,” the Darwinian implications of this often lead to it being referred to as “molecular evolution.” In a molecular evolution system, the fittest binder is amplified with each successive round of screening, whereas the poor binders’ concentration will either be unaffected or decrease. Although widely applied for many other purposes, DCLs have also been used to identify protein ligands, which make them a promising tool in chemical genetics library development.

Annotated chemical libraries

Generally, an assay is required to design followed by extensive screening. Truly daunting is the follow-up work to identify a hit compound’s mechanistic mode of action or target protein. An alternative approach to designing and building large collections of compounds of which no known biologic activity information exists is the use of an annotated chemical library (ACL). An ACL is a collection of compounds of diverse structure from various sources possessing experimentally bona fide biologic activities and mechanisms. An ACL contains compounds with diverse sets of biologic activity, whereas a typical library would even in a best-case scenario be composed of just a fraction of a percent of active compounds. One need not synthesize even a single compound to generate an ACL; one merely needs to identify, collect, and annotate. ACLs operate by assigning previously reported biologic activity to each compound without any required regard for the pathway or phenotype under study.

Phenotypic screening

Once libraries of compounds are assembled, the next step in chemical genetics is evaluating them for their biologic activity (i.e., assay or screening). Although virtually any kind of organisms can be used for phenotypic screenings in FCG, the most popular model systems have been Saccharomyces cerevisiae (budding yeast), Arabidopsis (plant), zebrafish embryo, drosophila (fruit fly), Caenorhabditis elegans (worm), and various mammalian cell cultures.

A stays must be designed with maximum sensitivity, selectivity, reproducibility, and cost-effectiveness in mind (64). Given the size of today’s libraries, the option of doing anything other than HTS is something less than desirable. Screening technologies have progressed in miniaturized steps from high-throughput (384 well plates) to ultra-high-throughput (3456 well plates) screening (u-HTS), and their limits have not yet been reached. However, although screening methods have matched the enormity of many libraries, the screens must be reliable and reproducible, and the data must also be manageable. Not only have the number of assays increased, but assays have increased in the dimensionality of the data produced, so-called “high-content screening.” Therefore, a significant challenge in HTS, and especially in u-HTS, is the development of tools that allow for data management and analysis in chemical genetics.

Yeast

Budding yeast (S. cerevisiae) is a robust and powerful tool in chemical genetics. Three significant advantages of yeast as a model organism in understanding cellular responses to chemical perturbation include its ease of growth, high genetic conservation with humans, and the size of the collection, up to 6000 genes, which makes their use a very high-throughput screen.

Plants

Plants offer an attractive platform for phenotypic screening in chemical genetics: 1) All known plant growth regulators are small molecules, the experimental protocols for analyzing plant growth regulators are well defined and can be easily adapted to unbiased chemical genetic screens; 2) the genomes of the most common systems are already sequenced and annotated, and the activities and mechanisms. An ACL contains compounds with diverse sets of biologic activity, whereas a typical library would even in a best-case scenario be composed of just a fraction of a percent of active compounds. One need not synthesize even a single compound to generate an ACL; one merely needs to identify, collect, and annotate. ACLs operate by assigning previously reported biologic activity to each compound without any required regard for the pathway or phenotype under study.

Zebrafish

Zebrafish (Danio rerio) have become a promising whole-organism screening method in chemical genetics for many of the same reasons they are popularly used in developmental biology and genetics. Whole-organism screening is preferred in some cases over target-based screening because it allows for a more unbiased discovery in a relevant physiologic context. However, although mammals provide an excellent relevant screening context, their use is expensive, requires a great deal of space, large quantities of compounds, strict regulations, is laborious, and often raises ethical questions. Because of these limitations in mammals, systems such as zebrafish have been popularized as a result of the many advantages they present. First, zebrafish, unlike other systems such as yeast or round worms, are vertebrates with discrete organs such as the brain, sensory organs, heart, muscles, and bones. These organ systems are very close to their human counterparts, in terms of their high level of structure, and this aids in their suitability for chemical genetics and drug discovery. Additionally, zebrafish are small enough in their
early embryonic stages to live in a well of a microtiter plate. They are also prolific reproducers, which allows for the screening of large libraries. Lastly, zebrafish embryos are a desirable model because of their complete transparency that allows for the multiple observations of dynamic processes in every organ and structure without the need for dissection or for sacrificing the animal.

**Drosophila**

Drosophila has been used by geneticists in many studies. Their short life cycle and low cost make them a desirable animal model. Although they bear less genes than humans, cases exist of one Drosophila gene representing several human genes, and this kind of feature makes it a popular model system for studying human disease pathways. Additionally, Drosophila is an excellent compliment to *C. elegans* RNAi studies in which some knockouts are unavailable, such as genes expressed in the nervous system.

**C. elegans**

*C. elegans* has had little use by researchers in chemical genetics thus far, but it is of growing use in drug discovery and is poised to make a large contribution. It is merely a matter of time until it becomes a widespread tool in chemical genetics. *C. elegans* was the first multicellular organism to have its genome completely sequenced. The knowledge and the experience that comes with it will prove invaluable in future *C. elegans* applications in chemical genetics. The worm is small and transparent, which allows for full visualization of its developmental processes and contains complex structures such as a digestive tract, nervous system, and muscles. Additionally, it has a short life cycle and produces many progeny, which makes it more compatible with high-throughput screening. An advantage, RNAi screens were first used and developed in *C. elegans* and the large amount of experience in this field will greatly aid multipronged chemical genetic approaches, particularly target identification.

**Target identification and validation**

Target identification is an integral part of chemical genetics studies and is its most rate-limiting and challenging step (57). Much of the difficulty originates from the weakly binding compounds often identified in FCG screening. The possibility of success in target identification greatly increases with increasing binding affinity. Whereas nM or pM binding constants will make target identification much easier, mid-µM results are much more common. Some simple approaches exist such as "the guess and test" where one hypothesizes a target and performs in vitro tests to validate it, or "guilt by association" identification where targets are implied or hinted at through studies such as mRNA transcription profiling (12). However, as in most FCG, without a clue toward the identity of the target protein, an ab initio target identification technique is required.

**Affinity matrix**

The most commonly used tool in target identification is the pull-down experiment, which is also known as the classic "fishing experiment" using an affinity matrix. This procedure typically involves the attachment of the "hit" molecule to a solid-phase resin such as agarose beads. The solid-linked material is then exposed to a cell extract. This process is commonly performed by passing the extract over a column of the immobilized material.

Although commonly used, affinity experiments are beset with drawbacks. First, the compounds need to be derivatized to include a handle for attachment to the resin, unless they intrinsically bear some functional tag that allows for it. The so-called tether effect can alter the activity of the compound, and tedious structure-activity relationships (SAR) studies are required to optimize the attachment point. As a solution, this SAR work could be avoided by introducing the intrinsic linker tag strategy (57). Two other requirements are needed for any hope of success in affinity matrix experiments: 1) high-affinity ligands and 2) high abundance of target. The end result of the affinity matrix relies on the multiplication of affinity and on the abundance. Also, during the cell extract preparation, a serious dilution occurs, at least 100 times. It is totally possible that the target protein exists in a complex with other protein(s) in intact cells, but it may be segregated into a monomer in the cell extract, and thus, it no longer binds to the small molecule. Therefore, a new technique, which can be performed in vivo without breaking the cellular integrity, is highly desirable.

**Photoaffinity**

Another interesting alternative is photoaffinity labeling, which involves attaching a photoaffinity moiety and a reporter tag. Although this method does require significant SAR knowledge, it does not require immobilization of the compound on solid support (3). Photoaffinity simply uses a photo-activated cross-linking group that forms a covalent bond during irradiation and a reporter, such as a radioactive isotope or biotin, that allows for isolation or ease of identification. Libraries have been designed bearing photoaffinity groups to exploit this approach.

**Tag for mass spectrometry**

Labeling proteins with "heavy" and "light" tags and screening the "hit" compound versus an inactive control, followed by mass spectrometric comparison of the two samples, is another approach that avoids many of the common pitfalls in affinity methods (3). Techniques such as stable isotope labeling with amino acids in cell culture and isotope-coded affinity tagging (ICAT) exemplify these techniques.

**Yeast three-hybrid system**

Genetic approaches also exist to tackle this problem. A promising approach is the yeast three-hybrid system (37). This work evolved from yeast two-hybrid screens and has grown in use (65). Three components are required for these studies: 1) a protein containing a DNA binding domain fused to a small molecule with a ligand binding domain, 2) a protein with a transcriptional activation domain fused to another ligand binding domain, and 3) a bivalent small molecule. The bivalent small molecule is composed of a known ligand with an affinity for the protein containing the DNA binding domain, a probe portion of the molecule that is being tested for novel protein binding, and a
linker connecting the two faces of the bivalent small molecule. Should the probe portion of the small molecule bind to the protein bearing the transcriptional activation domain, that protein is brought into a proximal relationship with the DNA and allows for the activation of the downstream reporter gene, which thus indicates successful target identification (66). The advantages of the Y3H system are that the identification of ligand binding proteins is linked to the selection of the cDNAs that encode the proteins, that phenotype and genotype are closely linked, and that these systems are explored in vivo. However, one drawback of this approach is that it is limited to simple, unicellular organisms, but work is underway to overcome this hurdle (3).

Drug western
In this approach, tagged small molecules are used to probe electrophoretically resolved cell extracts or cDNA expression libraries. Drug westerns involve bacteriophages infecting bacteria grown in a Petri dish with a cDNA library. Lysis caused by a viral infection leads to a clearing, called a plaque, containing a single member of the library. The proteins in the plaque are transferred to nitrocellulose where they are screened with tagged small molecules. Any hit plaques are isolated, a single protein–(GFP)–fused protein library was expressed in mammalian cells. When mixed together, the target proteins bind to the nanoparticle. An exogenously applied magnetic field assembles the GFP fused proteins whereby the signal can be visualized in vivo using a magnetic field assemblage (67).

Protein microarray
Protein microarrays are derived from and are complimentary to DNA microarrays. Here, collections of proteins are immobilized on a microarray surface and are probed for specific binding with tagged small molecules or by comparing the profiles of healthy versus diseased tissues (68). Although a very direct approach, protein microarrays suffer from the lack of large numbers of purified and stable proteins available for immobilization on the microarray surface. Although they have found great utility in proteinomics, their development has been slow because of several technical challenges.

Magnetism-based interaction capture (MAGIC)
A unique in vivo target identification method in mammalian cells has been reported recently. The Hit compound was attached covalently to magnetic nanoparticles and a green fluorescent protein–(GFP)–fused protein library was expressed in mammalian cells. When mixed together, the target proteins bind to the nanoparticle. An exogenously applied magnetic field assembles the GFP fused proteins whereby the signal can be visualized (69). Although a chemical modification of the Hit compounds is still required, this is the only technique reported so far to visualize a small-molecule–target-protein interaction in mammalian cells.

Display cloning
A phage display-based direct cloning of cellular protein has been demonstrated as a possible technique for target identification (70). The surface protein library was constructed using human brain cDNA, and a biotinylated FK506 probe molecule was used as a chemical bait. During the affinity selection, the FK.BP12 (FK506-binding protein) gene emerged as the dominant library member and was the only sequence identified after the second round of selection. Although the approach has been demonstrated successfully by isolating a full-length gene clone of FK.BP12, its broad application potential to novel target identification remains to be explored.

References
52. Ortholand JY, Gangean A. Natural products and combinato-
53. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental
and computational approaches to estimate solubility and perme-
ability in drug discovery and development settings. Adv. Drug
54. Burke MD, Schreiber SL. A planning strategy for diversity-
55. Lo MM-C, Neumann S, Nagayama S, Peristel EO, Schreiber
SL. A library of spirooxindoles based on a stereoselective three-
component coupling reaction. J. Am. Chem. Soc. 2004;126:
16077–16086.
56. Burke MD, Berger EM, Schreiber SL. A synthesis strategy yield-
ing skeletal diversity small molecules combinatorially. J. Am.
57. Mitsoupolous G, Walsh DP, Chang YT. Tagged library approach
2004;8:26–32.
58. Nielsen PE, Egholm M, Berg RH, Buchardt O. Sequence-Selective
Recognition of DNA by Strand Displacement with a Thymine-
59. Klib HC, Sharpless KB. The growing impact of click chemistry
60. Speers AE, Cravatt BF. Chemical strategies for activity-based
61. Rosaria GR, Lee W, Ding L, Yang HS, Chang YT. Combinatorial
approach to organosilane-targeted fluorescent library based on the
Solid-phase synthesis of styril dyes and their application as amy-
63. Ramstrom O, Banayatzoonsoot T, Lohmann S, Lehn JM. Chemical
64. Verkman AS. Drug discovery in academia. Am. J. Physiol.
65. Licitra EJ, Liu J. A three-hybrid system for detecting small
93:12877–12881.
66. Key J. Chemical dimersizers and three-systems: scanning the
67. Tanaka H, Ohtsuka N, Hidaka H. Isolation of cDNAs encoding
cellular drug-binding proteins using a novel expression cloning
68. MacBeath G, Schreiber SL. Printing proteins as microarrays for
high-throughput function determination. Science 2000;289:1760–
1763.
nanoprobe technology for detecting molecular interactions in live
70. Sich PP, McEwen KM, White JD, Austin DJ. Display cloning:
functional identification of natural product receptors using cDNA-

Further Reading

Crews CM, Splittgerber U. Chemical genetics: exploring and control-
1999;24:117–120.

Khenonsky SM, Chang YT. Strategies for facilitated forward chemical

See Also

Chemical L libraries: Screening for Biologically Active Small Molecules
Diversity-Oriented Synthesis of Small Molecules
Reverse Chemical Genetics
Small Molecule-Biological Target Interactions. Tool to Study
Small Molecules to Elucidate Biological Function
Advanced Article

Article Contents

• Metallointercalator-DNA Interactions and Site-Specific Targeting

• Metallointercalators as Probes of DNA-Mediated Charge Transport

• Implications and Conclusions

Metallointercalators as Probes of DNA Recognition and Reactions

Katherine E. Augustyn, Valerie C. Pierre and Jacqueline K. Barton, California Institute of Technology, Pasadena, California
doi: 10.1002/9780470048672.wecb328

Here we describe studies of metallointercalators bound to DNA. These octahedral transition metal complexes primarily bind noncovalently by stacking within the DNA helix. Given the rich photophysics and photochemistry of the ruthenium and rhodium complexes we employ, we have used a variety of biophysical studies to characterize their interactions with DNA. X-ray crystallography has also provided atomic resolution detail as to their binding to the duplex. Complexes have been designed that target DNA with high specificity. We have, for example, designed metal complexes that bind specifically to mismatched sites in the DNA duplex, and these have found application in the detection of single nucleotide polymorphisms and studies of mismatch repair deficiency. The photophysical properties of the metal complexes along with their intercalative stacking have been useful in particular as tools to characterize long-range charge transport in DNA. Using metallointercalators tethered to the duplex, oxidative damage to DNA from a distance has been demonstrated. The metallointercalators may serve as models for DNA-binding proteins, not only in binding DNA sites with high specificity, but also in carrying out electron transfer chemistry mediated by the DNA base-pair stack. Certainly these metallointercalators have proven to be powerful probes of this chemistry.

Our laboratory has focused on studies of metallointercalators that bind to duplex DNA through an ensemble of noncovalent interactions. Here we describe some of our studies with these complexes, including experimental design and applications. We focus attention in particular on the utility of these complexes in probing DNA-mediated charge transport chemistry. Our review here is not intended to be exhaustive but instead is focused on some examples of work from our laboratory to illustrate the effectiveness of these complexes in probing recognition and reactions with DNA (see Reference 1 for additional references). Indeed these complexes in several respects can be regarded as small mimics of DNA-binding proteins, but ones where their photophysics and photochemistry, as well as their inherent stability, allow us to sensitively probe their chemistry with DNA.

Investigations of the interactions of metallointercalators with DNA must start with a range of photophysical, nuclear magnetic resonance (NMR), and crystallographic studies. The complexes that we use are all substitutionally inert so that no direct coordination with the DNA bases occurs. The primary interaction involves intercalation of one ligand into the DNA base stack from the major groove. There is a concomitant doubling of the base-pair rise, from 3.4 Å to 6.8 Å, and the base pairs separate to accept the intercalating ligand. In the stack, the intercalator seems like a new base pair (2). A bulky metallointercalator that is not so easily accommodated in this stacked structure binds through an alternative means, where the sterically demanding ligand inserts instead from the minor groove side at a thermodynamically destabilized mismatch site, with ejection of mispaired bases. In this case, the stacked ligand replaces a base pair. Both with intercalation at B-DNA sites and insertion at mismatched DNA sites, chiral discrimination in binding of the octahedral complexes is evident; that is, there is a necessity sterically to match the chirality of the metal complex to that of the double helix (Fig. 1). Indeed, the $\Delta$ enantiomer favors interaction with right-handed B-DNA, whereas the $\Lambda$ enantiomer preferentially binds left-handed Z-DNA. These enantiomeric preferences...
Metallointercalators as Probes of DNA Recognition and Reactions

ancillary ligands

$\[\text{Rh} \text{(MGP)}_2 \text{phi}\]$ $^{3+}$ $\Delta$-isomer

$\[\text{Rh}(\text{DPB})_2 \text{phi}\]$ $^{3+}$ $\Delta$-isomer

$\[\text{Ru}(\text{bpy})_2 \text{dppz}\]^{2+}$ $\Delta$-isomer

$\[\text{Ru}(\text{bpy})_2 \text{dppz}\]^{2+}$ $\Lambda$-isomer

Figure 1 Examples of chiral metallointercalators that bind DNA with little site selectivity (above) and with high specificity for the targeted sites shown (below). In the center is shown schematically the basis for enantiomeric discrimination in stacking in the right-handed DNA helix. For the $\Delta$-isomer, the ancillary ligands have a right-handed orientation in the DNA groove, whereas for the left-handed $\Lambda$-isomer, steric clashes between the ancillary ligands and phosphate backbone can develop.

control further the reactions of complexes on the DNA helix. DNA-mediated charge transport involving $\Delta$-isomers similarly is more efficient than with left-handed complexes that are not as well coupled electronically into the right-handed double helix.

Metallointercalator/DNA Interactions and Site-Specific Targeting

Early photophysical and photochemical studies

Major groove intercalators bind DNA with high affinity ($K_a > 10^6 \text{ M}^{-1}$) and, in some cases, high sequence specificity. Indeed, an extended aromatic system on the ligand outward from the metal center, as in the case of the phi (9,10-phenanthroquinone dimer) or dppz (dipyridyl[3,2-a:2',3'-c]phenazine) ligands, favors its intercalative stacking between the base pairs of the double helix. The intercalating ligand of these complexes thus behaves as a stable anchor in the major groove, oriented parallel to the base pairs, and directing the orientation of functionalized ancillary ligands with respect to the DNA duplex. Photophysical studies first provided support for intercalation (4). Extensive NMR studies and a crystal structure detailed the nature of the intercalation for the metal complexes via the major groove of the DNA (2). Interestingly, although not all of these complexes are sequence specific, they still demonstrate chiral discrimination: The $\Delta$ enantiomer interacts preferentially with right-handed B-DNA through stacking within the DNA duplex.
Importantly, these complexes possess rich photochemical and photophysical characteristics that have been exploited advanta-
giously both to probe interactions with DNA and to understand
further aspects of DNA chemistry. A well-studied example is
[Ru(bpy)(dppz)]²⁺, which has found many uses as a molecular
light switch (5). The Ru complex shows solvatochromic lumini-
scence in organic solutions. In aqueous solution, however, it
does not luminesce because of the ability of water to deactivate
the excited state through hydrogen bonding with the phenazine
nitrogen atoms of the intercalating ligands. Upon intercalation
in DNA in aqueous solution, it is brightly luminescent, reflect-
ing the shielding of the intercalating ligand from bulk solvent.
This is akin to introducing the complex into a local organic sol-
vent that shields the ring nitrogens on the intercalating ligand
from protonation.

If ruthenium complexes have shown uses as molecular light
switches, rhodium analogs have been proven to be efficient
agents for photoactivated DNA strand cleavage (6). This re-
activity enables us to mark directly the site where the metal
complex intercalates in the double helix and thus character-
izes the recognition properties of the complex. The observed
reactivity upon excitation at short wavelengths (313-325 nm)
leads to radical formation on the intercalating ligand with sub-
sequent hydrogen atom abstraction from the adjacent deoxyri-
bosine ring. Degradation of the sugar radical then leads to direct
DNA strand cleavage. In the absence of oxygen, photolysis of
[Rh(DPB)(phi)]³⁺ bound to DNA results in the formation of 3′
phosphate and 5′ phosphate termini, as well as free bases. In the presence of oxygen, different products re-
sult: Direct strand cleavage is observed, but products include,
together with the 5′ phosphate termini, base propanoic acid,
and a 3′ phosphoglycaldehyde end. These results are consistent with
the previously described radical chemistry at the C3
position. However, because both the crystal and the NMR structure of the
major groove intercalator revealed that the C2′H of the sugar
is closer to the intercalating ligand than the C3′H, we propose
that initial reaction of the photo-exicted intercalator occurs with
the C2′H abstraction followed by H-migration to form the C3′
radical and subsequent degradation of the sugar ring (2).

Although rhodium intercalators efficiently cleave DNA upon
photoactivation, DNA cleaving agents that do not require pho-
toactivation have the advantage of being more convenient to
use in many research laboratories. Bisfunctional agents have
thus been developed in which metal coordinating peptides were
covaletly tethered to a metallointercalator (7). In these bi-
functional conjugates, [Rh(phi)bpy]³⁺ behaves as a targeting
vector that delivers the metal ions to the sugar phosphate back-
bone. The Zn(II) and Cu(II) centers of the metallopeptide, once
delivered, promote DNA strand cleavage.

In a similar approach, luminescent DNA cross-linking probes
were designed by conjugating short peptides to ruthenium in-
tercalators (8). In this case, [Ru(phen)(bpy)(dppz)]²⁺ delivers
the peptide to the oligonucleotide and oxidizes it upon irradi-
ation in the presence of an oxidative quencher. This then enables
the nearby tethered peptide to cross-link with the oxidized sites
of the DNA. A delivery of the peptide by the metal-
lointercalator is not essential for cross-linking, this technique
advantageously yields cross-linking adducts that are lumines-
cent and are thus easily detectable. Furthermore, these cross-links
may resemble those found in vivo under conditions of oxidative
stress.

**DNA recognition based on shape and functionalities**

One of the earlier and important findings with the metalloin-
tercalators is the importance of matching the chirality of the
metal complex with that of the double helix. The discrimi-
nation basically depends on the size of the ancillary ligands
relative to that of the DNA groove. Although some selectivity
is observed for intercalation into B-DNA with metal complexes
containing phenanthroline or bipyridine as ancillary ligands,
the most striking stereospecificity is observed with metal complexes
with bulky ancillary ligands such as [Rh(DPB)(phi)]³⁺ (9). The
enantiotopic rigidity of the rhodium complex selects one peri-
cleaves the sequence 5′-CTCTAGAG-3′ upon photoactivation.
However, no intercalation and cleavage is observed with the \(\lambda\)
enantiomer even with 1000-fold excess of metal complex.

The remarkable specificity of \(\Delta-\text{[Rh(DPB)(phi)]³⁺}\) enables
the efficient inhibition of the restriction endonuclease XbaI.
Notably, no comparable inhibition of XbaI is achieved with other metallointercalators and \(\Delta-\text{[Rh(DPB)(phi)]³⁺}\) also does not
inhibit restriction enzymes that bind to alternative se-
quences. Thus, this coordination complex effectively mimics
a DNA-binding protein.

Sequence-selective metallointercalators were also designed
de novo by matching the functionality of the ancillary lig-
ads positioned in the major groove with that of the targeted
base pairs. Targeting of 5′-CG-3′, for instance, is achieved
with the complexes [Rh(NH₃)₄(phi)]³⁺, [Rh[(R,R)-Me₂trien](phi)]³⁺, and
\(\Delta-\text{[Rh(en)(phi)]³⁺}\) through hydrogen bonding between the ax-
ial amines of the metallointercalators and O6 of guanine (11).
The predictive design of sequence-specific metallointercalators
was expanded with \(\Delta-\alpha-\text{[Rh}(\text{R, R})-\text{Me}_{2}\text{trien})(phi)]³⁺\), which is a complex that directly reads out the sequence 5′-TGCA-3′
(Fig. 2). The targeting of this site was based on predicted
hydrogen bonding contacts between the axial amine and the
O6 of guanine, as well as of the methyl groups on the metal complex and the methyl
groups on the flanking thymines. A high-resolution NMR so-
lution structure followed by the first high-resolution crystal
structure of a metallointercalator bound to DNA later revealed
atomic resolution detail of the interaction 3. As predicted, this
DNA unwinds to enable complete and deep intercalation of the
phi ligand of the metal complex, which intercalates through the
major groove, thereby causing a doubling of the rise in the in-
tercalation site. The metallointercalator thus behaves as a newly
inserted base pair, which the DNA accommodates with minimal
structural perturbation.

Metal-intercalators have also been designed to target specific
sequences based on their shape and on their functionalities.
A derivative of [Rh(phen)(phi)]³⁺, called 1-[Rh(MGP)(phi)]³⁺,
contains pendant guanidinium groups on the ancillary phan-
ethroline ligands, and it was designed to bind a subset of

---

**Metallointercalators as Probes of DNA Recognition and Reactions**

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
sequences recognized by the former. As predicted, the \( \Delta \) enantiomer recognizes the sequence 5'CATATG-3' specifically. Surprisingly, however, the \( \Delta \) enantiomer does bind DNA and recognizes the sequence 5'-CATATG-3' despite the large steric size of the ancillary ligands (11). Plasmid unwinding assays and NMR studies established that the \( \Delta \) enantiomer of the metallointercalator binds DNA by unwinding it up to 70\( ^\circ \). It is in this conformation that the complex can span the entire six base-pair binding site and contact the N7 position of the flanking guanines with the pendant guanidinium groups. Replacing these flanking guanines with deazaguanine demonstrated that the absence of the N7 nitrogen removed selectivity for the site. It is therefore the guanidinium functionalities of the ancillary ligands that are responsible for the recognition of the guanine, whereas it is the shape of the metallointercalator that enables the recognition of the "twistable" central 5'-ATATG-3' sequence.

This peculiar binding of a bulky \( \Delta \) enantiomer and unwinding of the DNA has found biological application in inhibiting transcription factor binding to DNA. 5'-\( [\text{Rh}[(R,R)\text{Me}_2\text{trien}]\phi]_3 \)-1 has been used to site specifically inhibit a transcription factor from binding to a modified activator recognition region (12). In competition experiments with yeast Activator Protein 1 (yAP-1), the metal complex was able to compete with the protein at concentrations as low as 120 nM. This work illustrates the potential applicability of the complexes as therapeutic agents, in inhibiting transcription factors sequence-specifically.

Detection of mismatched DNA sites

In contrast to the sequence specificity of major groove metallointercalators that is achieved through interactions of the ancillary ligands, the site-specificity for mismatched DNA is conferred by the intercalating ligand. DNA mismatch detection requires an extended or bulky intercalating ligand too wide to intercalate readily in well-matched B-DNA. The chrysene quinone diimine (chrysi) ligand of [\( \text{Rh}[(\text{bpy})_2\text{chrysi}]_3 \)]\( ^{1+} \) is 0.2\( \AA \) wider than the span of matched DNA and 2.1\( \AA \) wider than the intercalating phi ligand (Fig. 3). The resulting rhodium complex recognizes and cleaves upon photoactivation over 80% of mismatch sites in all possible single base-pair sequence contexts around the mispaired bases (15). The extremely high selectivity of [\( \text{Rh}[(\text{bpy})_2\text{chrysi}]_3 \)]\( ^{1+} \) for thermodynamically destabilized sites was demonstrated through the ability of the metallointercalator to recognize and photocleave a single mismatch within a 2725 base-pair plasmid (16).

The bulky rhodium complex binds single base mismatches with binding affinities of 0.3-20 \( \times \) 10\(^{5}\) M\(^{-1}\) as determined by quantitative photocleavage titrations. Importantly, the mismatch-specific binding affinities directly correlate with independent measurements of thermodynamic destabilization of the single base mismatch, thereby supporting the hypothesis that helix destabilization is a crucial factor in determining the binding affinity of the metal complex for the mismatched site (15). This hypothesis was recently confirmed by NMR studies and the crystal structure of the bulky metallointercalator bound to its target single base mismatch (14). Significantly, in contrast to phi complexes, \( \Delta [\text{Rh}[(\text{bpy})_2\text{chrysi}]_3 \)]\( ^{1+} \) inserts into the DNA stack via the minor groove and ejects both mismatched bases out of the double helix; the bulky chrysi ligand thus replaces a destabilized base pair. Nonetheless, the metallointercalator only minimally distorts the DNA, which accommodates insertion of the extended intercalating ligand simply by opening its phosphate backbone. Furthermore, as opposed to major groove intercalation, where DNA strand cleavage involves abstraction of the C2'H of the deoxyribose ring, \( \Delta [\text{Rh}[(\text{bpy})_2\text{chrysi}]_3 \)]\( ^{1+} \) preferentially abstracts the closer C1'H of the sugar adjacent to the mismatched site, resulting in different cleavage products (17).

Similar selectivity for thermodynamically destabilized sites in DNA is achieved with the use of other extended intercalating ligands. An analog of the chrysi complex, [\( \text{Rh}[(\text{bpy})_2\phi]_3 \)]\( ^{1+} \), also targets with high selectivity single base mismatches and promotes direct DNA strand scission upon photoactivation (18). The phi complex binds its target with higher affinity than...
Metallointercalators as Probes of DNA Recognition and Reactions

Figure 3 Monofunctional (top) and bifunctional (bottom) bulky metallointercalators that target single base mismatches in DNA. In the center is shown a view of the crystal structure of the complex inserted into the DNA from the minor groove at the mismatched DNA site, with ejection of the mismatched bases (14).

does the chrysi analog; site-specific photocleavage is evident even at nanomolar concentrations. This increase in affinity is attributed to greater stability in the stacking of the heterocyclic intercalating ligand with the flanking base pairs upon insertion. Notably, the increased affinity is not detrimental to the high selectivity of the metal complex, which binds mispaired versus well-paired sites in the same ratio as [Rh(bpy)2chrysi]3+. Metallointercalators that selectively and efficiently target single base mismatches have found several applications both as biologic probes and as potential chemotherapeutic agents. For instance, [Rh(bpy)2phzi]3+ was used to probe the relative frequency of mismatched sites in cell lines deficient versus proficient in their mismatch repair machinery (18). The relative cleavage observed with the phzi complex in healthy cell lines was low compared with that in cancer cell lines that carried mutations in essential repair proteins. These results support previous studies on the association of mismatch repair deficiency and cancer.

We have also investigated the design of fluorescent probes for mismatched DNA. The luminescent ruthenium complex, [Ru(bpy)2(tactp)]2+, containing a bulky intercalating ligand that is analogous to the dppz ligand of the popular molecular light switch was prepared and showed luminescence enhancement with mismatched DNA (19). A more efficient fluorescent probe for mismatched DNA was later accomplished by tethering a charged fluorophore to the bulky metallointercalator (20). In the Rhodium–Oregon Green conjugate, ion pairing between the cationic rhodium and the anionic fluorophore moieties dramatically quenches the fluorescence of the conjugate in aqueous solution and in the presence of matched DNA. However, with mismatched DNA, the bulky rhodium complex binds the DNA polyanion, and the resulting electrostatic repulsion with the anionic Oregon Green fluorophore drives the latter away from the rhodium center so as to reduce intramolecular quenching. The fluorescence of the conjugate is thus increased over 300% upon binding to a mismatch site.

Mismatch targeting metallointercalators have also been applied to the discovery of single nucleotide polymorphisms (SNPs). SNPs are the largest source of genetic variation in humans; yet their detection remains difficult as current methods have poor signal-to-noise ratio and yield many false positives. In this regard, mismatch selective metallointercalators have proven to be valuable new tools (21). When pooled genomic samples containing low-frequency SNPs are amplified, denatured, and annealed, mismatches are statistically generated at the polymorphic DNA sites. With photoactivation, these DNA mismatches are cleaved selectively by [Rh(bpy)2chrysi]3+ or [Rh(bpy)2phzi]3+. Fluorescent labeling of the cleaved products...
and separation by capillary electrophoresis thus permits rapid identification with single-base resolution of the SNP site. This method is remarkably sensitive, and minor allele frequencies as low as 5% can be readily detected.

Can these mismatch recognition agents be targeted inside cells? Intracellular delivery was first achieved by tethering a cell-penetrating peptide such as D-octaarginine to the rhodium complex (22). The resulting conjugate binds and with photocatalysis it selectively cleaves DNA neighboring single-base mismatches, although the presence of the oligoarginines is found to increase non-specific binding of the conjugates for both matched and mismatched DNA. Noticeably, the peptide does not affect the selectivity of the rhodium-induced photocleavage of the mismatch site. Similarly, the rhodium complex does not interfere with the delivery properties of the cell-penetrating peptide and the conjugates rapidly localize in the nucleus of HeLa cells.

A different strategy for the cellular uptake of metallointercalators consists of increasing the “greasiness” of the ancillary ligands (23). In agreement with studies on cis-platin analogs, increasing the lipophilicity of dopp complexes of ruthenium favors their passive uptake by HeLa cells. Importantly, the metal complexes are stable to the intracellular environment. Indeed no degradation in luminescence is evident, as would be expected based on changes in complex coordination.

A mentioned, deficiencies in the mismatch repair machinery (MMR) of cells are associated with an increased susceptibility to cancerous transformation. We have developed bifunctional metallointercalators as potential chemotherapeutic agents in which the rhodium complex serves as a targeting vector toward mismatches. For instance, the bulky metallointercalator was tethered to an aniline mustard known to form covalent adducts to 5′-GNC-3′ sites (24). The bifunctional agent demonstrates preferential alkylation of mismatched over fully matched DNA at concentrations where unmodified organic mustards show little reaction. Notably, the tethered alkylation does not inhibit binding of the intercalator at the mismatch site, and similarly, the metallointercalator does not hinder alkylation of the DNA. The site-selective alkylation at mismatched DNA thus renders these conjugates useful tools not only for the covalent tagging of DNA base-pair mismatches but also as new chemotherapeutic agents.

Similarly, a potential chemotherapeutic drug was designed via a bimetallic conjugate that combines a metallointercalator specific for DNA mismatches tethered to a reactive cis-platinum analog that coordinates DNA and inhibits transcription and replication (25). The recognition of a DNA mismatch by the bulky rhodium intercalator directs the reactivity of the platinum unit to a site close to the mismatch that may or may not be the preferred site for platinum coordination. Indeed, in the latter case, the rhodium targeting dominates over the platinum reactivity. This ability to tune the reactivity of the cis-platinum analog could lead to therapeutic agents for MMR-deficient cell lines.

Interestingly, the parent monofunctional metallointercalators have also shown promise as potential chemotherapeutic agents targeted to MMR-deficient tumor cell lines. Indeed, both [Rh(bpy)2chrysi]3+ and [Rh(bpy)2phzi]3+ inhibit cellular proliferation differentially in MMR-deficient cells compared with cells that are MMR-proficient (18). Significantly, the inhibition of cellular proliferation depends strictly on the mismatch repair deficiency of the cell and thus correlates with the ability of the bulky metallointercalators to target DNA mismatches. For instance, it is the α enantiomer of [Rh(bpy)2chrysi]3+ that is active both in targeting the mismatches and in inhibiting DNA synthesis; neither mismatch binding nor inhibition of cellular proliferation is observed with the β enantiomer. A additionally, the cellular response is enhanced with photocatalysis, which is an effect that correlates with the ability of the rhodium intercalators to promote strand cleavage at the mismatch site upon photocatalysis. Targeting DNA mismatches may thus provide a cell-selective strategy for chemotherapeutic design.

Metallointercalators as Probes of DNA-Mediated Charge Transport

Early photophysical studies

Rhodium and ruthenium intercalators have served as powerful probes of DNA-mediated charge transport. Since DNA-mediated charge transport depends so sensitively on π-stacking, it is reasonable that a probe that intercalates into DNA, with optimum π-stacking, might also serve as a powerful probe of this chemistry. Our earliest studies of DNA-mediated charge transport employed a ruthenium complex containing dopp as the photooxidant and a rhodium complex containing phi as the electron acceptor. Assemblies containing 5′-labeled [Rh(phen)2(dcpp)]2+ (phen′ = 5-amido-glutarate-1,10-phenanthroline) with and without 5′-labeled [Rh(phen)2(phen)]2+ were designed where fluorescence quenching of the photooxidant was observed only in the presence of the electron acceptor (26) (Fig. 4). Given a 0.75-eV driving force, little spectral overlap between the excited state of ruthenium and the ground state of rhodium, and the fact that the tethered complexes are well separated on a 15-mer duplex, the observed results were consistent with DNA-mediated charge transfer. This work set the stage for many varied experiments using metallointercalators to characterize this interesting chemistry.

A additional experiments spectrally identifying the ruthenium (I) intermediate confirmed that the quenching mechanism was caused by charge transfer (27). When bound to DNA, [Ru(DMP)(dppz)]2+ decays with two lifetimes corresponding to the two orientations of the intercalating dopp ligand. After excitation at 480nm of intercalating [Ru(DMP)(dppz)]2+ (DMP = 4,7-dimethylphenanthroline) in the presence of [Rh(phen)2(bpy)]2+ a negative transient was observed on the microsecond timescale at 440nm after the initial bleach corresponding to the decay of the ruthenium excited state. As the rhodium concentration was increased, the observed decrease in luminescence intensity but not in lifetime of the ruthenium excited state indicated that the quenching and hence the rate of charge transport was fast relative to the measurement. The same transient was also observed when [Ru(NH3)6]2+ was used as the quencher, albeit with a slower rate of formation. As expected, the decay kinetics of the transient was similar for both quenchers. Differences in quenching kinetics...
between \([\text{Rh}(\phi_2)_{2}(\text{bpy})]^{3+}\) and \([\text{Ru}(\text{NH}_3)_6]^{3+}\) are attributed to intercalation. The intercalative \([\text{Rh}(\phi_2)_{2}(\text{bpy})]^{3+}\) exhibits static quenching, whereas the diffusional \([\text{Ru}(\text{NH}_3)_6]^{3+}\) shows dynamic quenching.

**Long-range oxidative damage from a distance**

Oxidative conditions within the cell can lead to damage of the DNA bases. Guanine, which has been experimentally determined to have the lowest oxidation potential of the naturally occurring bases, is the most easily damaged (28). Upon oxidation, the neutral guanine radical can react with water or oxygen to form permanently damaged products such as 8-oxo-G, oxazolone, or imidazalone (29). Many organic photooxidants such as anthraquinone, riboflavin, and napthalimide have been shown to specifically damage the 5′-guanine of a guanine doublet (30).

The rhodium intercalator, which is tethered to the terminus of an oligonucleotide, was first employed to demonstrate oxidative damage to DNA from a distance through DNA charge transport (30, 31) (Fig. 5). With an excited state potential greater than 2.0 eV versus NHE, the rhodium intercalator serves as a potent photooxidant. Irradiation at 365 nm of 5′-radioactively labeled DNA duplexes containing the tethered rhodium intercalator results in oxidative damage through long-range hole transport that can be revealed by gel electrophoresis after treatment in hot piperidine; piperidine promotes strand breaks neighboring the base lesion 29. If these same duplexes are irradiated at shorter wavelength (313 nm), hydrogen abstraction leads to direct scission of the DNA backbone, indicating the exact position of intercalation. Comparison of the irradiation products of 15-mer oligonucleotides containing two sets of guanine doublets with a 5′-tethered rhodium intercalator versus one where the photooxidant is intercalated noncovalently reveals that the damage patterns show little distinction in oxidation of the proximal and distal guanine doublets 31. However, the 5′ guanine in both doublets was more susceptible to damage than the 3′ guanine. This finding is consistent with ab initio molecular orbital calculations that have indicated that the HOMO is localized on the 5′ guanine of a guanine doublet 32. This preferential reaction at the 5-G of guanine doublets has become a signature for one-electron DNA oxidation 30. Irradiation at 313 nm reveals that the covalently tethered rhodium intercalates three bases from the tethered end, whereas the noncovalent complex intercalates throughout the duplex. As the covalently bound rhodium intercalates far away from the observed damaged guanines, oxidation must therefore occur through long-range DNA-mediated charge transport. A detailed analysis of the damaged products by high performance liquid chromatography (HPLC) after enzymatic digestion showed that the primary damage product was 8-oxo-G.

Oxidative DNA damage has also been studied using the ruthenium intercalator, \([\text{Ru} \text{phen} \text{dppz} \text{bpy}]^{3+}\). When excited by visible light and quenched by non-intercalating quenchers such as \([\text{Ru}(\text{NH}_3)_6]^{3+}\), \([\text{Co}(\text{NH}_3)_5\text{Cl}]^{2+}\), or methyl viologen, a powerful ground state oxidant, \(\text{Ru(III)}\), is generated. With a 1.5 V potential versus NHE, \(\text{Ru(III)}\) can oxidize DNA bases in a similar manner to photoexcited \([\text{Rh}(\phi_2)_{2}(\text{bpy})]^{3+}\) (30). This flash quench technique, originally developed by Gray and coworkers to study electron transfer reactions in proteins, can also be applied to study DNA charge transport (33). DNA duplexes containing 5′ tethered \([\text{Ru} \text{phen} \text{dppz} \text{bpy}]^{3+}\) were
Metallointercalators as Probes of DNA Recognition and Reactions

Figure 5  Schematic representation of a metallated duplex designed to probe long-range oxidative damage in DNA. Oxidative damage is found at the 5’-G of the guanine doublet.

irradiated at 442 nm in the presence of either methyl violo-
gen or [Ru(NH3)6]3+, and preferential damage was observed at the 5’-guanine of a guanine doublet, which is consistent with one-electron oxidation chemistry. HPLC analysis showed 8-oxo guanine also as a primary oxidation product. To rule out guanine damage caused by singlet oxygen sensitization, duplexes containing tethered ruthenium were irradiated in the absence of quencher. Damage was only observed at guanines near the ruthenium intercalation site, and was not 5’-specific. Furthermore, this damage without quencher increased when the experiments were performed in D2O, a characteristic of singlet oxygen chemistry.

Spectroscopy of charge transfer intermediates

Spectroscopic studies provide a means to characterize DNA-mediated charge transport in more detail and to provide a link to biochemical observations. The flash quench technique was first used in experiments involving the synthetic oligonucleotide poly(G-G-C) (30). In the presence of poly(dG-dC), the negative absorbance at 440 nm, caused by excited state quenching of intercalated [Ru(phen)(dppz)2+] by non-intercalating [Ru(NH3)6]3+, disappeared concomitantly with a rise in a positive signal at 390 nm, which is consistent with the formation of the guanine radical. A difference spectrum of this species was obtained with strong positive features at 390 and 550 nm, indicative of the deprotonated neutral guanine radical. Importantly, this signal was not observed in the presence of the synthetic oligomer poly(dA-dT) nor in the absence of quencher in poly(dG-dC). Formation of the radical occurs in less than 10−7 s or within the time scale of quenching of the ruthenium excited state.

Spectroscopic studies of long-range DNA charge transport were then carried out on assemblies containing tethered [Ru(phen)(dppz)2(Lpy)2]2+ as the oxidant and the artificial base 4-methylindole as a guanine analog (34). The oxidation potential of 4-methylindole is lower than that of guanine, and the higher extinction coefficient of its radical at 600 nm renders it particularly amenable to spectroscopic studies of DNA-mediated charge transport. Excitation of ruthenium with oxidative quenching induces charge injection into the duplex. Hole migration to the methylindole base gives rise to a positive absorbance at 600 nm. However, if a mismatch is introduced into the intervening π stack of the duplex, DNA-mediated charge transport is disrupted, which leads to a complete attenuation of methylindole radical formation.

The effect of sequence on charge transfer rate was assessed with assemblies containing a tethered ruthenium photoxidant separated from the methylindole hole trap by a series of A-T base pairs of increasing length. Over distances of 17-37 Å, methyl-indole radical formation is found to occur concomitantly with quenching of the Ru(II) excited state. Thus the rate of radical formation over this distance through AT tracts is greater than 107 s−1, and over these distances, charge transport through the DNA is not rate-limiting. Furthermore, guanine radical formation can compete with that of the methylindole radical as charge equilibrates across the duplex (35).

Distance dependence of charge transport

Our earliest results had indicated that DNA charge transport might be significant over long molecular distances and certainly over longer distances than had been demonstrated in studies of protein electron transfer. The tethered rhodium intercalator was employed in the design of assemblies to examine relative oxidative damage at two guanine doublets within a 28-mer duplex.
Metallointercalators as Probes of DNA Recognition and Reactions

...the distal double guanine sites then served as an indicator of the proximity of the rhodium intercalator, whereas the distal doublet was placed in two base-pair increments at increasing distances from the photooxidant. The damage ratio between the proximal and distal double guanine sites then served as an indicator of charge transport efficiency. Remarkably the damage ratio did not show significant diminution as a function of distance over 75 Å. DNA charge transport thus shows a very shallow distance dependence that is not a result of helical phasing.

Similar studies were also carried out using the ruthenium intercalator, [Ru(phen)(dppz)(bpy')]2+. A comparison of oxidative damage with the rhodium and ruthenium photooxidants was carried out in a 63 base-pair duplex containing six sets of guanine doublets arranged at 19 base-pair increments. Both complexes were able to oxidize all six sets of double guanines, indicating that DNA charge transport chemistry could be observed at distances up to 197 Å away from the intercalation site. Oxidative damage to DNA in fact can develop over biologically significant distances.

Cyclopropyl amine-substituted bases have provided very fast traps for DNA charge transport using an irreversible ring opening reaction associated with oxidation (37). Guanine is in fact a poor oxidative trap, because the guanine radical reacts only on the microsecond-to-millisecond timescale with water and oxygen to form irreversible products; model studies suggest that irreversible ring opening of the N2-cyclopropylguanine radical occurs on the picosecond time scale. Recently our laboratory has designed a series of rhodium-tethered duplexes, in which the oxidative trap is N2-cyclopropylguanine (38). With intervening adenine tracts as the bridge between the intercalator and the trap, a shallow distance dependence is observed, now with prominent periodic features. Interestingly, if the same experiments are repeated monitoring guanine damage, the periodicities are not apparent. These periodicities were also absent in assemblies containing AT base-pair bridges, and they were less pronounced when the intervening bridge consisted of AI, ATC, or ATC (I = inosine) repeats.

An analogous hole trap, cyclobutylycycloisine, has also been used to monitor charge transport through the higher energy pyrimidine bases (39). Rhodium-tethered duplexes containing a distally placed cyclobutylycycloisine were irradiated and ring opening was observed. This striking result indicated that rhodium-induced charge transport can oxidize not only guanine bases but also cytosines. These findings indicate that charge transport through DNA must involve all the DNA bases, not only the low energy guanines we observed using gel electrophoresis.

Taken together, our results from gel electrophoresis and cyclobutylycycloisine ring opening can be rationalized by considering a novel mechanism for DNA-mediated charge transport: conformationally gated domain hopping. Two general mechanistic extremes are used to describe DNA-mediated charge transport, including superexchange, in which the donor and acceptor are lower in energy than the orbitals of the intervening DNA base-pair bridge, and hopping, in which the donor and acceptor have similar energies to the DNA base-pair bridge. In the superexchange mechanism, the charge tunnels from the donor to the acceptor without actually occupying the DNA bridge.

In the hopping mechanism, the charge hops from the donor to the acceptor, transiently occupying discrete sites on the bridge. A superexchange mechanism would exhibit an exponential distance dependence as a function of donor acceptor separation, but a hopping mechanism would result in a more shallow distance dependence so long as the hopping rate exceeds the rate of radical trapping. Although our results cannot be explained by either of these mechanistic extremes alone, we propose a mechanism in which the charge migrates through the DNA by hopping between transiently generated delocalized domains, defined by base sequence and dynamics. We describe a domain as a series of four to five bases acting in concert, over which a charge can delocalize. As the gel electrophoresis experiments measure guanine radical trapping on the millisecond timescale, contributions from base dynamics are not easily revealed. However, when the faster assay of oxidation-induced ring opening is used, additional effects from base dynamics can be discerned. Base motions, occurring on picosecond time scales, contribute to conformational gating of the charge transfer events, both limiting and facilitating the migration of charge between domains (40).

Sensitivity of charge transport to DNA conformation and dynamics

The importance of DNA conformation to DNA-mediated charge transport was evident also in many of our early photophysical studies. In fluorescence quenching experiments using assemblies containing the organic intercalator, ethidium, tethered to one end of the duplex, and the rhodium intercalator tethered to the other end, quenching by photoinduced electron transfer was found with a well-matched duplex, but no significant quenching was observed with an intervening CA mismatch (41). A well \( \pi \)-stacked array of heterocyclic aromatic bases is essential to the efficient transport of charge over a duplex. Perturbations in the intervening \( \pi \)-stack inhibit long-range oxidative damage. Assemblies containing a series of single base mismatches located between proximal and distal guanine doublet sites relative to a tethered ruthenium photooxidant were designed to explore the effect of stacking disruption on charge transport yield (42). These studies showed the dependence of charge transport efficiency on the dynamics of a mismatch; those mismatches that are relatively well stacked, as in purine-purine mismatches, cause only small attenuations in charge transport yield, whereas disruptive mismatches cause significant attenuations. In addition to mismatches, bulges can also disrupt the integrity of charge transport (43). Duplexes containing a tethered rhodium photooxidant and an AT bulge positioned in between a proximal and a distal double guanine site showed a drastic decrease in damage at the distal guanine doublet site, again underscoring the necessity of a well-stacked duplex. Indeed, the sensitivity in charge transport yield to intervening perturbations in base stacking is two important consequences: 1) the path of charge transport must be through the bases rather than through the sugar-phosphate backbone, and 2) the reaction can report sensitively upon the integrity of the DNA duplex.

An interesting study using the base flipping enzyme Methyltransferase Hhal (M.Hhal) showed that disruption of the \( \pi \)-stack by protein binding with insertion of a nonaromatic amino acid...
side chain can also significantly attenuate charge transport. M.HhaI performs its alkylating reaction on DNA after flipping out the central cytosine in the 5′-GGCG-3′ sequence and inserting a guanine residue in its place. An assembly containing a covalently tethered rhodium photooxidant and proximal and distal 5′-GG-3′ doublets separated by the M.HhaI target site was used to investigate charge transport yield in the presence versus absence of the enzyme. Site-specific binding of the enzyme to its target sequence was effective in eliminating oxidation at the distal doublet guanine site. Moreover, when a mutant enzyme containing tryptophan in place of guanine in the wild type was used instead, insertion of the aromatic amino acid served to restore the base-pair stack, leading to extensive damage at the distal site. From these studies it seems that the binding of DNA-binding proteins can both inhibit and activate long-range DNA charge transport.

Biological opportunities for DNA charge transport

We have extensively studied the effect of sequence, structure, and distance dependence of DNA-mediated charge transport. However, the ultimate question remains: is DNA-mediated charge transport an issue, indeed perhaps even a useful reaction, within the cell? We already knew that charge transport chemistry could occur over long enough distances to be biologically relevant, and that DNA-binding proteins could modulate the chemistry, but we needed also to determine whether this chemistry could occur within the tightly packed nucleosome structures found in cells. Within these structures, the DNA is highly bent, wrapped around a positively charged histone octamer, containing a S′-tethered rhodium was therefore constructed (45, 46) (Fig. 6); upon phototransactivation of the tethered rhodium, oxidative damage at a distance to guanines within the core particle was observed. In fact, the efficiency of damage was similar to that observed on the same DNA in the absence of bound histones. We have also shown that rhodium can induce DNA damage in the nucleus of HeLa cells 47. Moreover, in these studies, if we compare sites of rhodium binding with those of strong oxidative damage, we determine that oxidative damage can occur at a distance within the cell nucleus. Long-range charge transport through DNA does develop within the cell.

A additional evidence indicating that DNA charge transport may be biologically relevant comes from studies with DNA repair proteins such as MutY and EndoIII that contain [4Fe-4S] clusters (48). Bound to DNA, the redox potential of the [4Fe-4S] cluster in these proteins is found to be shifted so that the protein is more easily oxidized. By comparing potentials both bound to DNA and free, we estimate that the binding affinity of the oxidized form is at least three orders of magnitude higher than that of the reduced form. Based on these studies, we have proposed a model in which base excision repair enzymes can locate damaged DNA using DNA-mediated charge transport. A repair protein in its reduced state can bind to DNA, becoming more easily oxidized so as to transfer an electron to another DNA repair protein bound at a distal site, reducing the distally bound protein, and promoting its dissociation. But this DNA-mediated reaction can only occur if the intervening DNA base stack is intact and well stacked; if not, the protein remains associated with the DNA and on a slower time scale can progressively migrate to the damaged site. DNA-mediated charge transport thus serves to redistribute the repair proteins in regions of the genome near damage. The redistribution of repair proteins to a damage site using DNA-mediated charge transport essentially provides a way for the proteins to scan large regions of DNA without physically binding to each base. Recognizing that damage is particularly important under conditions of oxidative stress, when guanine radicals are generated. We also used the flash quench method with ruthenium intercalators to show that guanine radicals can provide the first signal for repair, promoting the oxidation also of the DNA repair proteins in a DNA-mediated reaction (49).

Implications and Conclusions

These experiments hopefully serve to illustrate the use of metallointercalators in probing recognition and reactions on the DNA helix. Starting with relatively simple coordination complexes that contain a wealth of photochemical and photophysical properties, functionalizing second and third generation derivatives, we have designed metallointercalators of high affinity, high specificity, and high usage in targeting and reacting with DNA. These complexes can serve as mimics of DNA binding proteins, competing effectively with them for DNA sites and perhaps even carrying out comparable electron transfer chemistry. These experiments, however, provide only a sampling of what might be considered in the future. Can these metal complexes, for example, serve as the basis for new chemotherapeutic designs targeted selectively to cancer cells? Perhaps the experiments we have described will inspire a new generation of complexes to follow with even more powerful applications to biology and medicine.

Figure 6. A nucleosome core particle is shown containing a 146 base-pair DNA duplex wrapped around a histone octamer with a rhodium intercalating photooxidant tethered to the DNA terminus. This particle was constructed to probe DNA charge transport through a nucleosome (45). Seven sets of guanine doublets are located at the red positions along the duplex. Oxidative damage initiated by rhodium photoactivation is observed at the guanine doublets, demonstrating long-range oxidation within the nucleosome.
Acknowledgments
We thank the NIH for financial support of this work. We thank also our many coworkers in these studies for their hard work and critical insights.

References

Metallointercalators as Probes of DNA Recognition and Reactions

We thank the NIH for financial support of this work. We thank also our many coworkers in these studies for their hard work and critical insights.

Acknowledgments
We thank the NIH for financial support of this work. We thank also our many coworkers in these studies for their hard work and critical insights.

References
Metallointercalators as Probes of DNA Recognition and Reactions


Further Reading


See Also

Inorganic Chemistry in Biology

Oxidative DNA Damage, Chemistry of

Physico-Chemical Properties of Nucleic Acids

Nucleic Acid Recognition by Peptides and Drugs

DNA Damage and Carcinogenesis
Phosphate Mimics, Cyclic Compounds as

Amanda C. Nottbohm and Paul J. Hergenrother, Department of Chemistry, Roger Adams Laboratory, University of Illinois, Urbana, Illinois

doi: 10.1002/9780470048672.wecb641

Compounds that contain phosphate and diphosphate moieties are not ideal biologic probes. Not only does their ionic character inhibit cell membrane permeability, but also, once inside a cell, the ester and the anhydride functionalities are likely targets for enzymatic cleavage. Thus, replacements for the phosphate motif are important as enzyme inhibitors, DNA or RNA analogs, phospholipid mimics, and phosphorylated metabolite analogs. To date, several classes of phosphate mimics have been developed that have been grouped into four categories: phosphorus-containing, sulfur-containing, dicarboxylates, and the novel cyclic mimics, which will be the focus of this review.

From nucleic acid phosphodiester linkages to reversible post-translational modifications, the phosphate motif is one of life’s most versatile functional groups. Cellular energy is stored in phosphorylated compounds such as ATP, and coenzymes often contain the pyrophosphate moiety. Protein interactions, stability, and activity are often mediated by the controlled introduction or removal of phosphate groups; indeed, it has been estimated that over 30% of human cellular proteins contain covalently bound phosphate groups (1). This modification has been implicated in the regulation of cell cycle progression, transcription control, protein synthesis, glycogen metabolism, and intracellular transport (2, 3). A host of disease states including cancer, inflammation, diabetes, atherosclerosis, immunodeficiency, and the bubonic plague (4–7) have been associated with the disruption of the phosphorylation machinery within the cell.

Although the ionized state of the phosphate moiety at physiologic pH likely allows for more efficient retention of phosphorylated intermediates within the cell wall and organelles (8, 9) and slows hydrolysis greatly (10), this charged state also prevents compounds that contain phosphate and diphosphate moieties from being useful biologic probes. Not only does their ionic character hinder cell membrane permeability, but also, once inside a cell, both the phosphate-ester and -anhydride functionalities are likely targets for enzymatic degradation. Thus, much effort has been directed toward the development of novel replacements for the phosphate motif, as such compounds might become new enzyme inhibitors, DNA or RNA analogs, phospholipid mimics, and isosteres of phosphorylated metabolites.

Therefore, the challenge is to design functional groups that mimic both the electronic properties and the spatial arrangement of the phosphate group while retaining cell permeability and stability. This task is extremely difficult considering the unique ability of the tetrahedral phosphate moiety to connect two modular subunits while possessing a negative charge. Additionally, this anion usually plays a major role in enzymatic binding and recognition events (11), as it often forms ionic contacts with both metal ions and lysine/arginine residues. Thus, phosphate replacements that lack ionization run the risk of not binding to their protein target. The ionization states and $pK_a$ values for free phosphate in solution are shown in Fig. 1 (12, 13). Interestingly, it has been demonstrated that the $pK_a$ of the phosphate group can change drastically on binding; in the binding of SH2 domains to phosphotyrosine-containing peptides, the phosphate in phosphotyrosine undergoes a large downward shift in $pK_a$ (from a free $pK_a$ of 6.1 to a bound $pK_a$ of 4.5) on binding, and additional studies indicated that 25% of the free energy of binding is caused by the second negative charge on the phospho monoester (14).

Current classes of phosphate isosteres have been developed to interact with target proteins in the same fashion as the phosphate moiety, but with increased bioavailability and stability in a physiologic environment. These phosphate motif replacements can be grouped broadly into four categories: three “traditional” replacement motifs (phosphorus-containing, sulfur-containing, and carboxylate linkages) that have been reviewed previously (3, 15), and the more unique cyclic mimics, which will be the major focus of this review (See Fig. 2 for general structures of mimics discussed in this review).
Phosphate Mimics, Cyclic Compounds as...

Common Phosphate Isosteres: Phosphonates, Sulfones, and Carboxylates

Phosphonates

Phosphonates, sulfones, and dicarboxylates are the most prevalent phosphate mimics in medicinal chemistry. Although the phosphate moiety makes the smallest perturbation, these compounds often do not solve the problem of cell membrane permeability (16-21), and can be limited in their biologic activity. Because of the high ionization constant of phosphonates compared with phosphates (%207), phosphonates are not di-ionic under physiologic conditions, and the oxygen-to-methylene substitution removes a potential hydrogen bonding moiety, which often leads to greatly reduced binding affinities (3). Nonetheless, development of phosphate inhibitors of glycosyltransferases (23-25), fucosyltransferases (26), and squalene synthetase (27) has led to the introduction of phosphonate-containing inhibitors with IC50 values of 100 nM, whereas the phosphonate-containing inhibitor with the same nonfluorinated analogs of UDP-galactose inhibited galactosyltransferase activity competitively with Kᵢ's of 62 to 969 μM (31). As indicated by the relatively weak activities of phosphonates, often these compounds are not adequate mimics of the phosphate group because of their increased pKa relative to the parent phosphate (Table 3). Through the introduction of one or two fluorine atoms, one can decrease the phosphonate pKa slightly to match a standard phosphate more closely; increasingly potent inhibitors have been developed through the application of this concept. For example, when incorporated into a hexameric peptide sequence, phosphonodifluoromethyl phosphenylalanine inhibitors of protein tyrosine phosphatase 1B have been developed with IC50 values of 100 nM, whereas the nonfluorinated analogs inhibit at 200 μM (28) (Fig. 3). In general, the biologic activity of (α-monofluoro)alkylphosphates remains less explored, which is surprising given the fact that their pKa is almost identical to that of the phosphate group (Table 3). Few studies have compared the activities of phosphonate-containing inhibitors with both their nonfluorinated and difluorinated equivalents, yet it has been reported that the introduction of a new stereochemical center with monofluorinated phosphonate analogs of glucose 6-phosphate is 11 times less potent than the 7(R)-monofluorophosphonate analog of glucose 6-phosphate isomer in studies of glucose 6-phosphate dehydrogenate substrates (32). Of course, because of the increased hydrophobicity...
Traditional phosphate mimics

- Phosphonate
- Phosphonate ester
- Dicarboxylate
- Sulfide

Novel cyclic phosphate mimics

- Tetrionic acid
- Squaric acid
- Thiazolidinones
- Rhodanine

Table 1

<table>
<thead>
<tr>
<th>X</th>
<th>Bond angle θ °</th>
<th>Bond length Å</th>
<th>( pK_{a2} )</th>
<th>% Dianion at pH 7.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>118.7</td>
<td>1.59</td>
<td>1.43</td>
<td>6.4</td>
</tr>
<tr>
<td>CH₂</td>
<td>121.1</td>
<td>1.81</td>
<td>1.51</td>
<td>7.5</td>
</tr>
<tr>
<td>CH₃</td>
<td>123.3</td>
<td>1.83</td>
<td>1.50</td>
<td>6.5</td>
</tr>
<tr>
<td>CF₂</td>
<td>116.5</td>
<td>1.85</td>
<td>1.50</td>
<td>5.6</td>
</tr>
</tbody>
</table>

*Only the monofluorophosphonates closely match the \( pK_a \) of the parent phosphate.

Sulfur-containing mimics

To eliminate problems associated with the stability and the membrane impermeability of phosphates in vivo, incorporation of sulfur-associated functionalities has been explored. Oxidized...
Phosphate Mimics, Cyclic Compounds as pTyr Pmp

IC50 200 μM

F2Pmp
IC50 0.1 μM

Figure 3 Phosphonomethyl phenylalanine (Pmp) and phosphonodifluoromethyl phenylalanine (F2Pmp) are nonhydrolyzable phosphotyrosyl (pTyr) mimetics. Hexameric peptide sequences that incorporate either Pmp or F2Pmp inhibit protein tyrosine phosphatase 1B. The difluoro analogs are most potent, presumably because of their reduced pKa relative to the phosphonate.

Forms of sulfur can provide isosteric replacements for a phosphate, but they will lack the full negative charge. Although a direct comparison with the phosphate moiety is difficult to make, studies have indicated that the oxygen of DMSO carries an effective charge of ~0.63 (31), whereas the Mulliken atomic charge on a diaryl sulfone is ~0.616 (32). On the other hand, computational examination of the ESP charges on the oxygens of phosphotyrosine molecules have found that they carry charges of ~0.73 or ~1.06 for the monoanionic and dianionic forms, respectively (33). It must be noted that these cannot be considered direct comparisons because Mulliken charges are calculated based on the linear combination of atomic orbitals method, and ESP charges represent the electrostatic potential around an atom.

Multiple efforts have been made to replace phosphorous-containing linkages with sulfur-containing isosteres in the context of enzyme inhibition. In a search for nonionic transition state analog inhibitors of restriction enzymes, Blätter et al. (34) found that nucleic acid duplexes that incorporate a dimethyl sulfone in place of a phosphodiester have distorted backbones similar to those in restriction enzyme bound DNA. Chimeric DNAs that incorporate sulfone linkages were synthesized, and depending on the location of the dimethylene sulfone linker, either between the first AT unit or the second AT unit in the EcoRV recognition site, Ki values were 20 nM and 120 nM.

Figure 4 (a) Analogs of DNA that contain a dimethyl sulfone group have a distorted backbone compared with the natural biopolymer, which makes them useful as inhibitors of restriction enzymes. (b) A sulfonylbenzoyl nitrostyrene inhibitor of tyrosine protein kinase was designed based on a transition state model of the reaction of ATP with tyrosine residues and exhibits good activity. (c) Based on salicyl-AMP, a sulfamoyl-containing compound exhibits remarkable activity in the inhibition of siderophore biosynthesis.

4 WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Cyclic Phosphate Mimics

In an effort to improve cell permeability/bioavailability, recently developed phosphate isosteres often have decreased overall charge. In the past, it has been proposed that reduced p-tency is often a trade-off with increased cellular permeability of phosphate replacements (3), but these trends are beginning to change. A recent review of enzymatic phosphate recognition (11) pointed out that the phosphate-binding mode depends greatly on orientation within the protein. A thorough burial binding pockets often are filled with neutral amino acid residues, cationic residues play a larger role in phosphate recognition and binding closer to the surface. Additionally, over one third of phosphate binding sites do not contain a metal or other cationic phosphate binder (11); thus, the chances of using a neutral, more drug-like compound are increased in these cases. Next, cyclic phosphate mimics including tetronic acid, squaric acid, thiazolidinone, rhodanine, perfluorooxy- and sulfhydantoin-based derivatives (see Fig. 2) will be discussed.

Cyclic phosphate mimics: tetronic acid derivatives

Using a 3-acyltetronic acid as a phosphate-mimicking core (Fig. 6a), potent inhibitors of cdc25B (a dual specificity phosphatase [DSP]) and vaccinia VHR-related phosphatase (VHR) were identified recently (43). Drawing on general natural product inhibitors of protein tyrosine phosphatases (PTPs), Sodeoka et al. (43) searched for a novel core structure that could interact specifically with the active site loop of DSPs. Although previous attempts to find PTP inhibitors focused on phosphonates and carboxylates as phosphate core isosteres (44), the acidic 3-acyltetronic acid group of the known VHR inhibitor RK-682 (Fig. 6b) looked promising as a general replacement for the phosphate moiety. A library of 36 tetronic acid derivatives was synthesized, and some selective inhibitors of both cdc25B and VHR were identified. In particular, a tetronic acid derivative with a diazomalonyl group was found to be extremely potent against cdc25B (Fig. 6b), with a 30-fold selectivity for cdc25B over VHR. In general, the library showed good selectivity for VHR and cdc25B over general tyrosine phosphatases. Most crucially, modeling studies indicated that the dissociated 3-acyltetronic acid anion can act as a hydrogen-bonding phosphate mimic, whereas the R and R2 groups provide two positions for additional derivatization, which allows for increased specificity.

Cyclic phosphate mimics: squaric acid-based motifs

Over the years, squaric acid and its derivatives have found several uses in medicinal chemistry: Squaric acid inhibits glyoxy-

sulfone-3-cyclobutenedione) is an inhibitor of pyruvate dehydrogenase and transketolase (46); and other derivatives serve as antagonists of the N-methyl-D-
Phosphate Mimics, Cyclic Compounds as farnesyl pyrophosphate natural substrate

![Diagram of Phosphate Mimics]

Chaetomellic acid A

(a)

(b)

(c)

K_D = 42 μM

K_D = 0.10 μM

K_D = 7.0 μM

K_D = 2.5 μM

K_D = 1.0 μM

Figure 5

(a) Chaetomellic acid A is a novel dicarboxylate-containing natural product that is thought to mimic farnesyl pyrophosphate and thus to inhibit the enzyme FTPase. This compound can exist in either the dicarboxylate or anhydride form. (b) As peptides that contain aromatic oxamic acids are good inhibitors of tyrosine kinase p56Lck SH2 domains, it is believed that the oxamic acid moiety may mimic the phosphate group. (c) A symmetric 3,5-disubstituted benzoate analog of S3P (substrate) and EPSP (product) inhibits EPSPS.

Searching for an isostere with similar charge distribution to that of the phosphate group for the creation of oligodeoxynucleotide analogs, Sato et al. (48, 49) also examined squaric acid derivatives. Squaric acid contains two acidic hydroxyl groups (pK_a values of 0.54 and 3.48, respectively) as well as two highly polarized carbonyl groups (50). Ab initio calculations of electrostatic charge distributions indicate that although a dissociated dimethyl phosphate has –0.84 charges on each of the oxygen atoms, N-isopropyl-N'-methylsquaryldiamide has similar polarizing patterns, with –0.47 and –0.51 charges on the oxygen atoms of the carbonyl groups (49). Furthermore, similar to phosphate functionality, the squaric acid motif contains carbonyls that can function as hydrogen-bonding partners and binding sites for divalent metal ions (Fig. 7a) (51, 52).

On the basis of these calculations, a modified 3′-5′ thymidine dimer derivative (Fig. 7b, TsqT) was synthesized (49). Based on both UV and CD spectra of TsqT and the corresponding phosphate derivative (TpT), the dimer that contains the squaryldiamide linkage is similar structurally to the natural TpT. Thus, the TsqT dimer structural motif was incorporated into oligodeoxynucleotides using standard phosphoramidite chemistry, producing 5′-CGCATsqTAGCC-3′ and 5′-GACGCA sqTAGCCGAT. Enzymatic digestion of squaryl-
In a similar manner, a 2-oligodeoxynucleotides may have a variety of applications (49).

Furthermore, squaryldiamide-modified oligodeoxynucleotides (ODNs) demonstrated A–T base pairing was preserved, which indicates that the squaryldiamide-modified oligomers can mimic the phosphate group, particularly that of phosphotyrosine. Developing such a more robust ODN analog that contains the squaric acid motif as a nonhydrolyzable phosphotyrosine mimetic enzymes tested, including RmlB, RmlC, and RmlD. The fact that all three enzymes modify a phosphorylated rhamnose intermediate important for Mycobacterium tuberculosis (MTB) cell wall synthesis was a possible indication that substituted rhodanines might mimic the phosphate motif, but it was unknown whether such inhibitors actually bind in the same fashion as the natural substrates. A later computational follow-up study by K. Karanadjeff et al. (57) indicated that the third rhodanine inhibitor binds at or near the phosphate-binding region of MTB. Core structure electron-density isosurfaces for these antmycobacterial compounds were generated, further hinting at the common features shared between substituted rhodanines or thiazolidinones. Core structure electron-density isosurfaces for the thiazolidinones and the diphosphate moiety.

Core structure electron-density isosurfaces for these antmycobacterial compounds were generated, further hinting at the common features shared between substituted rhodanines or thiazolidinones. Core structure electron-density isosurfaces for the thiazolidinones and the diphosphate moiety.

Cyclic phosphate mimics: thiazolidinones and rhodanines (thiothiazolidinones)

MurB, an enzyme that reduces enolglycyl uridine diphosphate N-acetylglucosamine (EP-UNAG) to uridine diphosphate N-acetylMuramic acid (UNAM) (54), is an attractive antibacterial target (55). Using crystallographic data of MurB bound to EP-UNAC, Andres et al. (55) sought a suitable surrogate for the diphosphate moiety of the natural substrate, and focused eventually on 4-thiazolidinones. Modeling studies indicated that a carbonylic acid moiety at the R1 position potentially could mimic essential interactions of the diphosphate with a lysine residue (56). Additionally, it was believed that the thiazolidinone core would orient the side chains so that they might occupy the space normally reserved for the uridine and glucosamine motifs. Although it is unlikely that the thiazolidinone core can mimic all interactions of the diphosphate, elimination of the corresponding rotatable bonds of the natural substrate was also predicted to have a positive entropic effect to binding of the thiazolidinone-based inhibitor. Only thiazolidinones that contain an n-butyl group at the R1 position exhibited any activity, with diastereomers derived from D-norleucine that exhibited a 4-fold increase in activity over those synthesized from L-norleucine (55).

In a recent screen of 8,000 compounds to identify potential drugs that target the synthesis of DTP-rhamnose from glucose-1-phosphate and dTTP, 11 inhibitors were identified; three of these contained a rhodanine (Fig 2) structural motif (56). Inhibition was found for all three dTDP-rhamnose synthetase enzymes tested, including RmlB, RmlC, and RmlD. The fact that all three enzymes modify a phosphorylated rhamnose intermediate important for Mycobacterium tuberculosis (MTB) cell wall synthesis was a possible indication that substituted rhodanines might mimic the phosphate motif, but it was unknown whether such inhibitors actually bind in the same fashion as the natural substrates. A later computational follow-up study by K. Karanadjeff et al. (57) indicated that the third rhodanine inhibitor binds at or near the phosphate-binding region of MTB. Core structure electron-density isosurfaces for these antmycobacterial compounds were generated, further hinting at the common features shared between substituted rhodanines or thiazolidinones and the diphosphate moiety.

High-throughput screening for compounds that might compete with the UDP-GlcNAc substrate of MurG also led to the identification of a rhodanine core structure as a phosphotide mimic (Fig 8c) (58). Using a fluorescein derivative of UDP-GlcNAc, compounds that compete with the binding of the fluorescent substrate were identified on the basis of anisotropy changes on incubation with the putative inhibitors. Over 48,000 compounds were screened, and of the 277 compounds that were identified as possible inhibitors of MurG,
Phosphate Mimics, Cyclic Compounds as

Squaric acid

(a)

(b)

UsqT

TsqT

(c)

phosphotyrosine

IC50 = 47 μM

IC50 = 56 μM

Figure 7  (a) Dialkylphosphate similarity to squaric acid derivatives; both possess two negatively charged oxygen atoms in their tautomeric form. (b) Modified oligodeoxynucleotides that contain a squaryldiamide moiety are remarkably similar to natural oligodeoxynucleotides. 3′-5′ linked dimers can form Watson-Crick base pairs with adenine, whereas the 2′-5′ linked dimers can form both Watson-Crick base pairs with adenine and wobble base pairs with guanine. (c) Squaric acid-based inhibitors of PTPase from Yersinia are designed to mimic phosphotyrosine.

11 compounds exhibited greater inhibitory effects than the natural inhibitor UDP. 7 compounds contained a rhodanine core structure. A similar screen also revealed several compounds with rhodanine-like structures. It should be noted that these 1,3-substituted compounds are conjugate acceptors and thus potentially reactive as electrophiles; however, additional kinetic analysis indicated that no covalent modification of the enzyme was occurring. Additionally, most compounds were selective for MurG over closely related enzymes that use similar substrates. The high percentage of inhibitors with a similar core may possibly hint at a common binding mode, which reveals the ability of rhodamines to potentially serve as diphosphate mimics.

Although it is unknown whether either thiazolidinones or rhodanines actually mimic the phosphate binding site or simply have shapes that occupy the nucleotide-sugar binding site favorably, as proposed by Carlson, studies of systems that do not contain a nucleotide-sugar seem to indicate the former. Modeling studies first indicated that particular perfluoraryl compounds might mimic the overall spatial and electronic distribution of the diphosphate and triphosphate moiety. By appending a perfluoranyl moiety onto a guanosine substrate, Barber et al. hoped to mimic either guanosine...
Phosphate Mimics, Cyclic Compounds as

(a) The substituted thiazolidinone core. (b) Sometimes, thiazolidinone-containing compounds make good inhibitors of enzymes that use nucleoside diphosphates. For example, MurB, which uses EP-UNAG as its natural substrate is inhibited by the thiazolidinone derivative pictured. (c) Rhodanine-containing compounds also serve as excellent inhibitors of enzymes with natural diphosphate-containing inhibitors. For comparison, UDP (an inhibitor of MurG) is pictured next to a rhodanine derivative that exhibits good antibacterial activity. (d) Thiazolidinones have proven their usefulness as general replacements for natural substrates that contain the phosphate group. For example, 2-aryl-4-oxo-thiazolidin-3-yl-amides have replaced serine amide phosphates as more stable mimics of lysophosphatidic acid, and they have been shown to be cytotoxic in prostate cancer cell lines.

5′-triphosphate or guanosine 5′-diphosphate and thus modulate the activity of Ras. Barber et al. (67) succeeded in appending several perfluoraryl phosphate replacement moieties onto guanosine. Although no IC₅₀ values are given in the patent literature (66), the relative binding abilities of the guanosine nucleotide mimics were compared with that of GTP for the binding site of a mutant H-Ras protein using a cold chase experiment, in which either GTP or a perfluoraryl compound was used as a competitor to tritiated GTP. When experiments were performed with the natural substrate, GTP (Fig. 9a), at a concentration of 25µM, a 79% reduction in radioactivity was observed. Using the putative GDP mimic 5′-O-(2,3,5,6-Tetrafluoro-4-pyridyl) guanosine (Fig. 9a) at the same concentration, a 21% reduction in the mean c.p.m. was noted, which indicates that the perfluoraryl compound likely inhibits binding of GTP to Ras to some extent. Interestingly, when using a compound designed to mimic GTP, 5′-O-[(2,3,6-Trifluoro-5-hydroxy-4-nitrophenyl) guanosine (Fig. 9a), no reduction occurred in the binding ability of tritiated GTP (66).

Taking this concept a step further, general inhibitors for protein farnesyltransferase (FTase) (67), geranylgeranylttransferase I (GGTase-I), and squalene synthase (SqSase) (68) were also developed. In a posttranslational modification event, FTase transfers a farnesyl group to the Ras protein (69). Although not directly related to the function of Ras, SqSase uses the same substrate as FTase; thus, the compounds developed were also tested for SqSase inhibition (68). Finally, GGTase I controls the geranylgeranylation of Rho proteins and plays a role in cell motility and invasion, and also is a potential target for cancer therapy (70). Of 15 compounds tested, 5 compounds exhibited IC₅₀ values less than 10µM (68), with the best/most selective inhibitors for each enzyme illustrated in Fig. 9b–d.
Cyclic phosphate mimics: sulfhydantoins

Relatively little information exists about sulfhydantoins in the literature. Although they have been described previously as peptidomimetic serine protease inhibitors (71), Saunders et al. (72) were the first to report the use of sulfhydantoins as phosphate mimics. In an attempt to develop novel inhibitors for SHP-2, which is a tyrosine phosphatase, it was noted that general phosphate inhibitors either incorporate carboxylate or the fluorophosphonate moiety to mimic the negative charge of the phosphate at physiologic pH (71). Driven by the lack of drugs-like inhibitors in the literature, they searched for heterocyclic modulators of SHP-2 activity based on sulfhydantoins. In vitro competition assays were completed to determine the IC₅₀ values of several putative phosphate mimics, the most potent of which are illustrated in Fig. 10. Although exact IC₅₀ values were not reported in the patent literature, the compounds depicted exhibit IC₅₀ values between 1 and 100 µM (72).

Cyclic phosphate mimics: hydroxytropolones

Hydroxytropolones derived from puberulonic acid, a natural product isolated from Penicillium, were investigated for their ability to inhibit inositol monophosphatase (IMPase). Although bisphosphonic acids were developed previously as potent inhibitors of enzymes that use phosphate-containing substrates, other bimetallic, phosphate-binding metal complexes might also be targets of hydroxytropolone inhibition. Indeed alkaline phosphatase, which contains two catalytic zinc ions, the tri-zinc-containing phosphofructokinase C (75), and human immunodeficiency virus RT polymerase/ RNase H/Integrase (76), which contain essential Mg²⁺ ions, are inhibited competitively by hydroxytropolones as well.

Cyclic phosphate mimics: do they really mimic the phosphate group?

Many cyclic phosphate mimics have been developed recently, and only limited data exists about their ability to actually imitate the phosphate group. Thus, the term “phosphate mimic” simply refers to the ability of these compounds to act as potent inhibitors of enzymes that use phosphate-containing substrates. Nonetheless, a myriad of computational studies have attempted to showcase the structural and electrostatic similarities of certain cyclic phosphate replacements relative to their phosphate parents. Using the program SPA RTA N. 5.0 (Wavefunction, Inc., Irvine, CA), electrostatic charge distribution studies have attempted to show the structural and electrostatic similarities of certain cyclic phosphate replacements relative to their phosphate parents. Using the program SPA RTA N. 5.0 (Wavefunction, Inc., Irvine, CA), electrostatic charge distribution studies have attempted to showcase the structural and electrostatic similarities of certain cyclic phosphate replacements relative to their phosphate parents. Using the program SPA RTA N. 5.0 (Wavefunction, Inc., Irvine, CA), electrostatic charge distribution studies have attempted to showcase the structural and electrostatic similarities of certain cyclic phosphate replacements relative to their phosphate parents. Using the program SPA RTA N. 5.0 (Wavefunction, Inc., Irvine, CA), electrostatic charge distribution studies have attempted to showcase the structural and electrostatic similarities of certain cyclic phosphate replacements relative to their phosphate parents.
interact with a phenylalanine residue, a positively charged region of the protein is proposed to interact with the carbonyl moieties of the squaric derivative (Fig 12b). Unfortunately, no direct comparison of the binding mode of phosphotyrosine versus the proposed binding of these squaric acid derivatives exists in the literature; no cocrystal structures or NMR structures of these compounds are bound to proteins. Other docking studies have been completed with tetronic acid derivatives in the active site of VHR; results indicate that on dissociation, the 3-acyltetronic acid anion can bind tightly to the active site through several hydrogen bonds in a manner similar to that of phosphotyrosine (43). Finally, rhodanine and thiazolidinone cores have been shown to bind at or near the phosphate-binding site in modeling studies of various proteins (57), and they share many commonalities with compounds believed to mimic the pentavalent transition state of phosphate hydrolysis. Core structure electron-density isosurfaces for both rhodanine and thiazolidinone cores have been shown to bind at or near the phosphoimine motif in the active site of IMPase. At this time, crystal structures of proteins complexed with these compounds are bound to proteins.

Choosing the Right Phosphate Isostere

With so many phosphate isosteres to choose from, where does one begin? Despite recent advances, it would seem that the ideal isostere must still be determined on a case-by-case basis. Ease of synthesis and compatibility of the designed synthetic route with production of a combinatorial library of compounds can be a major determining factor. Additionally, one must consider the history of each of these classes as a phosphate mimic. Much data has been generated on the “traditional” phosphate replacements, and many success stories exist with these classes of compounds. However, many published examples exist in which such replacements are biologically inactive (19, 40), and as highlighted by Fressigné et al. (78), the incorporation of phosphate replacements affects the conformational behavior of some systems strongly. In contrast, several recent publications have highlighted the potential of the cyclic compounds as phosphate mimics, but only time will tell whether many failures exist as successes.

**Choosing the right phosphate isostere: good**

Although they have been around the longest, phosphorus- and sulfur-containing replacements for the phosphate moiety present the greatest synthetic challenges. The low yields associated with appending phosphonates and fluorophosphonates make the synthesis of combinatorial libraries more difficult, which has been an obstacle in the development of potent inhibitors from these classes of molecules. Additionally, although they are certainly more stable biologically than their parent phosphates, the highly polar nature of phosphonates hinders biologic activity because of limited cell permeability (17-22). For example, one of the most potent and selective inhibitors for protein tyrosine phosphatase 1B (PTP 1B) contains an aryl difluorophosphonate group (16). Because of its anionic character at physiologic pH, it is not cell permeable and exhibits no cellular activity; only when produced in a pro-drug form do such phosphonate inhibitors become useful (16). In general, the phosphonate inhibitors that exist in the literature have slightly reduced potencies compared with their phosphate parent (3); thus, other structural and electronic analogs of the phosphate moiety often will be superior.

Progress has been made in the development of potent sulfone-containing phosphate mimetic inhibitors, but it is difficult to predict in what cases a sulfone might resemble a phosphate. Whereas a select few sulfone-based inhibitors have
Phosphate Mimics, Cyclic Compounds as

Dimethyl phosphate  N-isopropyl-N'-methylsquaryldiamide

-0.84  -0.84  -0.47  -0.51

Figure 12 (a) Ab initio calculations show the similarity in charge distribution of dissociated dimethyl phosphate and N-isopropyl-N'-methylsquaryldiamide. Electrostatic charge distribution and charge localization are illustrated below (34, 49). Reprinted in part from Reference 47 with permission from the American Chemical Society. (b) Proposed interactions of squaric acid at the active site of YopH (77). Reprinted in part from Reference 75 with permission from Elsevier.

Choosing the right phosphate isostere: better

Although several phosphate isosteres discussed in the previous sections demonstrate great promise, little hard data exists about their ability to actually function as a phosphate replacement. For example, only one model system has been published for both the tetrionic acid derivatives and the sulfhydantoins, and limited structural evidence exists that demonstrates phosphate-mimicking ability. Additionally, even though the sulfhydantoin or reverse sulfhydantoin core resembles the structure of biologically active the thiazolidinones and rhamnines, the only evidence for their activity is limited to the patent literature, in which no absolute IC50 values or crystal structures are presented (72). Finally, although the hydroxytropolones exhibit promising activity in a few model systems, labeling them as phosphate mimics rather than general inhibitors of bimetallic enzymes might be premature. Whereas it is possible that these three cores also function to some extent as a general phosphate replacement, more study must be completed before this can be confirmed.

Choosing the right phosphate isostere: best

Grouped in this final category are the highly desirable drug-like phosphate mimics, which include squaric acid derivatives, rhamnines/thiazolidinones, and perfluoraryl compounds; a wide range of relatively simple methods exist for the synthesis of these derivatives, and they have been demonstrated to work as phosphate mimics for at least two different classes of enzyme substrates. Squaric acid derivatives have been used in two very different phosphate-mimicking situations: in squaryl-linked oligodeoxynucleotides (48, 49) and as inhibitors of PTPases (53). Additionally, computational studies that involve electrostatic charge distribution have added to the evidence that squaric acid derivatives might be similar spatially and electronically to the highly acidic proton on the acetylsulfamate nitrogen might mimic some of the negative charge found on the phosphate, although more information is needed on this topic.

Most successful dicarboxylate-containing phosphate mimics have been based on carboxylate-rich natural products (42) that have been shown previously to have activity against the desired enzyme. Although modeling studies have shown that the negatively charged oxygens on maleate are spatially within 0.1 Å of that in the equivalent diphosphate, some malonate-, succinate-, and tartaric acid-based phosphate mimics have not been potent inhibitors of enzymatic activity (35, 38). Although dicarboxylate-containing compounds have also been developed to mimic the six-membered ring pyrophosphate-divalent metal complex that is proposed to form in many enzymes (80, 81), in those cases monosaccharide linkages have been found to be much better mimics of this chair or boat conformation (Fig. 13) (37). Carboxylate moieties can be incorporated easily into compounds through amide linkers (37) or oxidation strategies (38), but their inability to consistently function as potent inhibitors limits their applicability.
the phosphate moiety. Thiazolidinone and rhodanine derivatives can simply be prepared through a one-pot cyclocondensation reaction, which has led to their relatively high occurrence as members of commercially available compound collections. Additionally, the thiazolidinone/rhodanine core has been shown to be a valid phosphate replacement for several classes of molecules; both nucleoside diphosphates (56, 58) and phosphorylated lipids (64) have been mimicked by these heterocyclic core structures, which hints at their possible potential to serve as a general phosphate mimetic. Perfluoroaryl compounds have been demonstrated to mimic both the diphosphate and triphosphate moiety of nucleoside phosphates (67) as well as phosphoroylized isoprenoids (66, 68). Thus, it seems reasonable to believe that these strategies for phosphate mimicity could be applied successfully to other biologic targets. In the coming years, even more testing and crystallographic or computational evidence will hopefully illuminate the manner in which these compounds act as phosphate isosteres. For now, based on a few key examples these three derivatives all seem to have great potential.

Summary

Phosphonates, sulfones, and dicarboxylates are all valid replacements for the phosphate moiety, but it has become clear that several interesting alternatives to these functional groups exist. Although the squaric acid and thiazolidinone motifs have been best characterized in terms of their ability to actually mimic both the negative charge and the positional orientation provided by a phosphate group, other cyclic compounds have surfaced additionally, the thiazolidinone/rhodanine core has been shown to be a valid phosphate replacement for several classes of commercially available compound collections. Additionally, the thiazolidinone/rhodanine core has been shown to be a valid phosphate replacement for several classes of molecules; both nucleoside diphosphates (56, 58) and phosphorylated lipids (64) have been mimicked by these heterocyclic core structures, which hints at their possible potential to serve as a general phosphate mimetic. Perfluoroaryl compounds have been demonstrated to mimic both the diphosphate and triphosphate moiety of nucleoside phosphates (67) as well as phosphoroylized isoprenoids (66, 68). Thus, it seems reasonable to believe that these strategies for phosphate mimicity could be applied successfully to other biologic targets. In the coming years, even more testing and crystallographic or computational evidence will hopefully illuminate the manner in which these compounds act as phosphate isosteres. For now, based on a few key examples these three derivatives all seem to have great potential.

References

Phosphate Mimics, Cyclic Compounds as


54. Benson TE, Walsh CT, Massey V. Kinetic characterization of wild-type and S229A mutant MtrA: evidence for the role of serine 229 as a general acid. Biochemistry 1997;36:796-805.
70. Lebowitz PF, Pragerger GC. Non-Ras targets of farnesyltransferase inhibitors: Focus on RhD. Oncogene 1998;17:1439-1447.
Small Molecule Inhibitors, Design and Selection of

Jianwei Che and Yi Liu, Genomics Institute of the Novartis Research Foundation, San Diego, California

Nathanael S. Gray, Harvard Medical School, Dana Farber Cancer Institute, Biological Chemistry and Molecular Pharmacology, Boston, Massachusetts
doi: 10.1002/9780470048672.wecb123

The discovery and optimization of small-molecule inhibitors for the treatment of human disease has been the major focus of the pharmaceutical industry for more than a century. In the last decade, they have also attracted increased attention from academia because they provide a powerful means to interrogate biological systems in an acute fashion. Although pharmaceutical companies have identified many small-molecule inhibitors for drug targets, to find a desirable small-molecule inhibitor for a specific target is still a major challenge for drug discovery and for chemical biology. Here, we present an overview of the computational approaches that can be exploited to discover small-molecule inhibitors for protein targets of interest.

Small molecule inhibitors are of special value to chemical biology and to drug discovery. Thousands of small molecules have been developed as drugs to treat various human diseases. Increasingly, small-molecule inhibitors are used to perturb the function of proteins to elucidate their biologic functions (1, 2). Complementary to the conventional genetics, small molecules allow rapid, conditional, and tunable inhibition. Furthermore, small molecules can often be identified that alter specifically one particular function of a multifunctional protein. The recent completion of the human genome sequencing project has resulted in the identification of a plethora of new protein targets for which small-molecule modulators could be developed as potential therapeutics and as biologic “tools.” Identifying a selective small-molecule inhibitor for each protein in the human genome is a formidable challenge for drug discovery and for chemical biology. Experimentally, high-throughput screening is the most widely used technology currently to identify hits for a target. On the other hand, computational methods, such as virtual screening and structure-based design, have become equally important to identify and to optimize small-molecule inhibitors. In Fig. 1, we illustrate how computational and experimental approaches can complement each other in the compound discovery and optimization process. The numbers of compounds associated typically with each method is indicated. At early stages of hit identification, high throughput methods such as virtual ligand screening (VLS) and ultra high-throughput screening (uHTS) are used frequently. At later stages, high accuracy algorithms combined with biologic assays can be used to guide the selection and the synthesis of a limited set of analogs. As shown in Fig. 1, computational and experimental approaches are not mutually exclusive processes. They are best exploited by using computation and experiment in an iterative fashion to optimize compounds progressively for a particular target of interest. Without these interactions, none of the methods can achieve its maximum potential. In this review, we focus on computational approaches depicted in Fig. 1. As will become clear later in the article, most computational methods are not associated exclusively with predefined processes or protocols. Significant redundancies exist in terms of when and what methods should be applied. The structure of the article follows the common flow of identifying small-molecule inhibitors. We will discuss the particular computational methodologies used at different stages of the process in subsequent sections.

Virtual Screening in Selecting Small Molecule Inhibitors

Virtual screening (VS) uses computational or theoretic methods to select molecules with desired biologic properties. It has broad applications in pharmaceutical industry. In this article, we limit ourselves to VS to identify biologically active small molecules. The estimated chemical space for small organic molecules is over 10^38, and the registered organic and inorganic molecules in CAS are already over 30 million. Despite advances in experimental high throughput screening technologies, currently it is impossible to screen more than a few million compounds.
With the rapid improvement of computational power and algorithms, VS methods have become an important complementary technique to experimental uHTS. It has many distinct advantages, such as investigating virtual molecules that can cover a broader area of chemical space. Compared with uHTS, VS is also a relatively inexpensive means to probe many molecular structures.

Generally, two types of virtual screening are used. One is receptor based, such as docking, and the other is ligand based, such as similarity search. In practice, both methods are often used synergistically. In Fig. 2, we illustrate how these two types of methods are used to select “hits” from a given small-molecule library.

Receptor-Based VLS

In receptor-based VLS, the atomic coordinates of receptor molecules are known experimentally or theoretically (e.g., homology models). The interactions between a small molecule and a protein are often characterized by a “lock-key” mechanism, in which high affinity small molecules often fit the protein-binding site well. The shape complementarities provide important enthalpy contributions from favorable van der Waals interactions. In addition, correctly positioned charges, hydrogen bonds, and hydrophobic contacts usually are required for higher affinity. Based on computational efficiency, receptor-based VLS can be divided into three tiers. The first tier is the popular high-throughput docking and scoring that can be computed for each ligand molecule within a few minutes on a typical desktop computer. In the rest of this article, we refer to high-throughput docking as docking and scoring for simplicity. The second tier includes the medium throughput methods that calculate the interactions more thoroughly and accurately. It often includes algorithms that account more accurately for desolvation energies, entropic penalties, and receptor flexibility. For example, detailed implicit solvent models usually are applied in these calculations. It normally takes tens of minutes for each calculation. The third tier represents the slowest method but ultimately offers more accurate results. The typical examples are atomistic molecular dynamics (MD) or Monte Carlo simulations with explicit water molecules. The details of these calculations are out of the scope of this review. Interested readers can consult recent reviews on these specific topics (3).

Among the three classes of methods, docking is the most widely used method in pharmaceutical industry. Many high throughput docking software packages are available publicly, such as Dock, Autodock, Gold, Glide, FlexX, QXP, and ICM. In addition, many companies have built their own proprietary docking technologies. A docking-based VLS calculation usually consists of three stages: preparing, computing binding modes, and scoring. Currently, most docking programs treat receptors as rigid molecules. Under this approximation, a receptor field can be calculated for the binding site. A receptor field is the interaction potential created by receptor atoms. It consists of van der Waals, electrostatic, hydrogen bonding, and solvation contributions. Without atom thermal motions, the receptor field is also static, so it can be precalculated in the absence of ligand molecules. This feature provides enormous savings of computational time over explicit atom models. Different programs employ very different algorithms to sample ligand conformations, to orient small molecules in the binding site, to retain preferred binding modes, and to increase the computational efficiency (4–7). The predicted binding configuration is selected based on molecular interaction energies or an empirical scoring function. Finally, docking ranks all ligand molecules based on binding poses that possess the highest score, and virtual “hit” molecules are selected from the top of the list. This method
The docking/scoring uses existing 3-D structures of receptors to construct binding models, and the pharmacophore searching/ranking uses known active hits, 9 of them inhibited SIGK binding with G receptor 3-D structure (i.e., docking and scoring) and the other is based on active ligand molecules alone (i.e., pharmacophore searching and ranking).

et al. (12) to identify that ranged from 100 nM to 60 µM, one chemical diversity set of 1990 compounds. Among 85 virtual actions. Bonacci et al. (11) used docking (FlexX) to screen the to discover small-molecule inhibitors for protein–protein interaction. II inhibitors (9) and BCR-ABL tyrosine kinase inhibitors (10) were also discovered by docking. Docking can also be applied to discover small-molecule inhibitors for protein–protein interactions. Banaci et al. (12) used docking (FlexX) to screen the chemical diversity set of 1990 compounds. Among 85 virtual hits, 9 of them inhibited SIGK binding with Gβδ1-2 with IC50 that ranged from 100 nM to 600 nM. It was also used by Trusset et al. (12) to identify β-Catenin inhibitors for Wnt signal pathway from 177,000 compounds. Among 22 tested virtual hits, 3 hits were confirmed as binders to β-Catenin and Tcf4 competitors. Without experimental receptor structures, docking can be carried out using carefully constructed model structures. Becker et al. (13) and Salo et al. (14) reported the successful identification of small-molecule inhibitors of several G-protein coupled receptors (GPCR) systems based on docking.

One major challenge in docking-based VLS is the accuracy of scoring functions. Conceptually, binding free energy is a rigorous measure for correct binding conformation and compound affinity. However, the inherent errors in classic force fields and the extremely lengthy computation required preclude its immediate use for VLS. In practice, an empirical function (i.e., a scoring function) normally is used to estimate free energy of binding. A few examples of commonly used scoring functions are PMF (15), Chemscore (16), Drugscore (17), Goldscore (18), and GlideScore (19). The specific mathematical forms of Chemscore, Goldscore, and PMF are shown here to highlight their similarity and difference. Other scoring functions can be found in respective references:

\[ \Delta G_{\text{chemscore}} = \Delta G_0 + \Delta G_{\text{bond}} + \Delta G_{\text{vdw}} + \Delta G_{\text{hbond}} + \Delta G_{\text{metal}}. \]

where \( \Delta G_0 \) is a fitting constant, \( \Delta G_{\text{bond}} \) is hydrogen bond energy between ligand and receptor, \( \Delta G_{\text{vdw}} \) is the energy between ligand atoms and metal ions in a receptor active site, \( \Delta G_{\text{hbond}} \) is the hydrogen receptor lipophilic interaction, and \( \Delta G_{\text{metal}} \) is for the rotatable bond penalty.

\[ \Delta G_{\text{glidescore}} = \Delta G_{\text{bond}} + \Delta E_{\text{vdw}} + \Delta E_{\text{internal}}. \]

where \( \Delta G_{\text{bond}} \) is the hydrogen bond energy between ligand and receptor, and \( \Delta E_{\text{vdw}} \) is the van der Waals energy of ligand and receptor, and \( \Delta E_{\text{internal}} \) is the van der Waals energy and torsional strain of the ligand molecule.

Figure 2. Workflow of computational algorithms used for identifying “hit” molecules from a compound library. The top panel indicates the compound collection from which the “hit” molecules are to be identified. The middle panel illustrates two virtual screening strategies. One is based on a known receptor 3-D structure (i.e., docking and scoring) and the other is based on active ligand molecules alone (i.e., pharmacophore searching and ranking). The docking/scoring uses existing 3-D structures of receptors to construct binding models, and the pharmacophore searching/ranking uses known active small molecules to construct pharmacophore models.
Small Molecule Inhibitors: Design and Selection of

\[ \Delta G_{\text{bind}}^{\text{score}} = \sum_{i<j} A_{ij}(r) + \Delta E_{\text{vdw}}, \]

where \( A_{ij}(r) = -k_B T \ln \left( \frac{\rho_{\text{seg}}(r)}{\rho_{\text{bulk}}} \right) \)

\[ \Delta E_{\text{vdw}} = -\frac{1}{3} \sum_{i<j} \frac{\rho_{\text{seg}}(r)}{\rho_{\text{bulk}}}, \]

is the ligand volume correction factor, \( \rho_{\text{seg}}(r) \) is the number density of a ligand receptor atom pair of type \( ij \) at inter atomic distance \( r \), \( \rho_{\text{bulk}} \) is the reference number density of ligand receptor pair of type \( ij \), and \( \Delta E_{\text{vdw}} \) is the van der Waals interaction between ligand and receptor. Although common features exist within different scoring functions, they often emphasize different aspects of the protein–ligand interaction. These empirical scoring functions bear significant errors because of crude approximations. Currently, no scoring functions exist that are able to rank diverse compound libraries consistently and accurately. The performance of docking based VLS varies considerably from system to system and from program to program (20). Although some programs tend to perform better than others do, it is difficult to predict the performance a priori on a specific system. In addition to ranking molecules based upon a single scoring function, combination of several scores (i.e., consensus scoring) or statistical methods (e.g., Bayesian statistics) are also used to select virtual hits.

Often, rigid receptor models provide a reasonable approximation and offer tremendous practical value. The majority of docking calculations used in industry adopt this approximation. In reality, however, receptor atoms undergo constant thermal motions, and some proteins have significant flexibility. One example is inactive and active conformations of kinases in which the activation loop moves in and out of active sites (21) with some atoms that move well over 10 Å. It is common to see a variety of conformational states for a given protein in PDB database. It is also possible that such conformations observed crystallographically represent only a fraction of the transient dynamic states that may exist in solution. Although several methods (22–24) have been developed to account for protein flexibility, considerably more computational and experimental work will be required to create algorithms that can incorporate receptor flexibility accurately. In addition to receptor plasticity, another complicating factor is the difficulty to account accurately for the enthalpy and entropic contributions of discrete water molecules in the ligand bound and unbound states. Water molecules have been observed crystallographically to play important roles in mediating hydrogen bonds between protein atoms and between protein atoms and ligands. In most docking calculations, the water molecules are removed from binding sites, which assumes that to displace these water molecules by small molecules would give favorable entropic contribution. However, the entropic gain comes at the expense of enthalpy loss if the water molecules make strong hydrogen bonds to receptor atoms or mediate ligand–protein interactions. Several programs exist that attempt to include movable water molecules in docking calculations, however, often this results in a significant loss of computational efficiency.

Besides virtual screening of ligand libraries, docking is extremely useful for lead optimization. Successful docking can elucidate the important interactions between lead molecules and the target, explain important features of the structure activity relationships (SAR), and consequently help to design new molecules. Other than docking, useful receptor based VLS tools include MD simulation, MC simulation, and QM/MM calculation.

Ligand-Based Virtual Screening

(2-D Similarity Search, 3-D Pharmacophore Search)

Another category of VLS methods is to identify ligands for targets that have unknown 3-dimensional (3-D) structures. Generally, these methods are called ligand-based VLS. The most simple and probably the most used searches involve a physical, chemical, or topological similarity search based on one or several active compounds. The similarity between two molecules can be defined in various ways. Two-dimensional (2-D) methods refer to those that only consider atom connection tables to describe molecular structures, because atom connection tables are best represented by 2-D pictures such as the ones in the top panel of Fig. 2. These methods consider topological relationships between atoms in a molecule regardless of hydrogen atoms. The simplicity of these methods allows computationally efficient comparisons. It can screen millions of molecules in seconds. Here, we outline a few representative examples.

1. Molecular 2-D fingerprints are a class of methods that use chemical substructures to characterize similarities between molecules. A fingerprint is a signature that encodes the chemical topological information of a small molecule. It is usually a long binary bit string in which each bit represents the presence 1 or absence 0 of a particular chemical substructure. For example, an ATP molecule has substructures such as a pyrimidine and imidazole heterocycle but not benzene; therefore, the bits that represent pyrimidine and imidazole would be set to “1” and the bit that represents benzene would be set “0” in the fingerprint bit string of ATP. Many variations in detailed definitions of 2-D fingerprints exist. In one straightforward way, molecular similarities are measured by the number of bits two compounds have in common.

2. Pharmacophores are defined as essential structural features in a molecule that are responsible for its binding activity, and they can also be used to generate 2-D fingerprints. Examples of commonly used pharmacophores are specific molecular fragments (i.e., an isopropyl group), hydrogen bond donors and acceptors, and negatively and positively charged atoms or groups. In two dimensions, the minimum number of bonds between two pharmacophores can be used as a distance metric. For example, the exocyclic amino
group of ATP is located 3, 4, and 5 bonds away to the three acceptor nitrogens in the purine ring. This information is coded readily in a fingerprint bit string.

3. In addition to 2-D fingerprints, the topological relationships of atoms in a molecule can be explored in high feature dimensions and nonlinear fashion by other means, such as kernel methods (25), support vector machines (26), and self-organizing maps (27).

However, molecular recognition occurs in a 3-D world. It is well known that minor chemical modifications can alter the activity of a molecule drastically. On the other hand, very different chemical structures can have similar activity against the same target. The mere existence of particular pharmacophores and/or their topological relations are not sufficient for a molecule to be active against its target. To account for 3-D information, various VLS techniques based on activities of small molecules without receptor structures have been developed, such as 3-D quantitative structure activity relationship (QSAR) models. It is commonly accepted that a small molecule must exhibit a high degree of shape complementarity to bind to its protein target with high affinity. Finding the correct superposition of active ligand molecules is the key to a successful 3-D QSAR model.

Moreover, these membrane-localized target structures are unknown with high affinity. Finding the correct superposition of active ligand molecules is the key to a successful 3-D QSAR model. Of course, shape alone does not account for all contributions of small molecules. Other factors such as locations of hydrogen bond donors and acceptors, types of functional groups, electrostatic properties, and polarizabilities all play important roles for small-molecule activities. One attempt to capture this information is 3-D pharmacophore method. Like docking, many software programs exist to develop 3-D pharmacophore models, such as DISCO (28), Catalyst (28), GASP (28), PHASE (29), and many others (30, 31). Pharmacophore features are not mutually exclusive; for instance, a hydrogen bond acceptor (i.e., donating electron pair) can be negatively charged (e.g., side chain carboxylate from Asp or Glu). Most programs provide users with the flexibility to define their own pharmacophore features. The main merit of 3-D pharmacophore models is the ability to identify lead compounds that are diverse structurally.

Although three pharmacophore models may differ in how they construct pharmacophore models, they go through the following common steps: generate conformers, enumerate pharmacophore features, generate a hypothesis, and screen the ligand library. Usually, the active conformation of a small molecule is unknown without experimental evidence. A key component of a good pharmacophore method is to sample active conformations rapidly. Studies have shown that the active conformation of a small molecule may not necessarily be the same as the lowest energy one in the unbound state (32). Therefore, a set of possible conformers within a certain energy window (e.g., 15 kcal/mol above the lowest energy conformer) are collected and clustered. The feature-enumerating step is relatively straightforward. It identifies all the possible pharmacophore features in small molecules.

Hypothesis generation is another challenging step in building pharmacophore models. In essence, this step involves finding a proper alignment of active molecules with a high degree of overlap in 3-D space in terms of their pharmacophore features. During this step, the spatial arrangements of pharmacophore features are identified so that they are common for all or a significant subset of active molecules. This step assumes that active molecules must satisfy a common pattern of pharmacophore features. However, it should not require a molecule to have all the features defined in a model to be active. The proper alignment of active molecules determines the quality of a 3-D pharmacophore model. Several algorithms have been developed to identify pharmacophore patterns (33-36) among the set of possible conformations for each active molecule. Pharmacophore hypothesis algorithms try to find and to rank the most probable pharmacophore models for known active molecules. It should also be noted that active molecules do not necessarily share a common pharmacophore pattern. Distinct binding modes have been observed in cocrystal structures (37). Moreover, different small molecules may bind to different binding sites of a receptor. Although a few algorithms exist to identify multiple pharmacophore patterns within a set of actives (29, 36), more robust methods are needed. The conformations of small molecules can be precomputed or sampled "on the fly." Once a valid pharmacophore model is constructed, it can be used to screen millions of small molecules. This strategy was used successfully by Singh et al. to identify some potent and novel VLA-4 antagonists (38). Three-dimensional pharmacophore models are used commonly for GPCR systems because generally, these membrane-localized target structures are unknown. A three-point pharmacophore model was constructed for Urotensin II receptor (39) and was used to search Avantis compound collection. Of 500 virtual "hits," 10 hits were active with the most potent compound exhibiting an IC_{50} of 400 nM. Recently, the first GPR38 specific agonist was discovered by 2-D and shape similarity in combination with a pharmacophore search (40).

Not only can active molecules be used in pharmacophore modeling, structures of inactive molecules can also provide valuable information. In common pharmacophore models, information provided by inactive compounds is captured in the concept of "excluded volume" (29, 41). A "regions that is so close to the receptor that no ligand atoms are tolerated. "Excluded volume" can be obtained automatically or defined manually. The method relies on the idea that the binding modes of the compound series are similar to each other. Although many examples from cocrystal structures and SAR have confirmed this assumption, many exceptions to this empirical observation exist. Chemical similarity does not always predict similarity in biologic activities. When active compounds consist of multiple scaffolds because of different modes of action, it is very challenging to develop pharmacophore models and to classify the actives into proper categories automatically. In addition, a compound that satisfies a good pharmacophore model is not guaranteed to have activity. For example, entropic contributions, solubility, cell permeability, and stability can also have important implications on a compound's biologic activity. Even though most pharmacophore models are built without receptor structures, binding site structures do help to construct proper pharmacophore models. The critical residues provide a complementary image of ligand pharmacophore field. This "negative image can be converted to a ligand pharmacophore model. One advantage of receptor structure-based pharmacophore models is the natural definition of "excluded volume."
Small Molecule Inhibitors: Design and Selection of...
Small Molecule Inhibitors, Design and Selection of molecular weights of between 120–250 Da to allow for subsequent additions. Usually, these compounds have weak affinities (K_d = 10 μM–10 mmol/L). To explore modifications that have the potential to increase binding affinities, the aforementioned computational methods can be applied readily. For structures with a single small molecule in a complex, the small compound can be extended automatically to maximize its interaction with target protein. During the progressive modifications, the binding mode for the initial fragment can be altered. However, certain constraints are applied to avoid drastic changes to the binding modes unless substantial modifications are introduced to the original structure. When evidence exists that multiple ligands bind to different pockets of an extended binding site in one or multiple crystal structures, computational methods can be used to propose proper linkers or fuse-disjointed fragments to yield the synergistic effects. Gill et al. (49) have used this strategy to design P38α MAP kinase inhibitors based on hits with affinities in the millimolar range. For example, they identified a lead molecule with IC_{50} = 65 nM by extending the base fragment to acquire additional interactions, and also to achieve a 100-fold increasing in potency rapidly by cojoining two overlapping fragment hits for an indole-derived compound. The structures are depicted in Fig. 4. It should be emphasized that crystal structures are extremely important in fragment-based design iterations. Because fragment hits are usually small and possess low affinity for the targets, to predict the correct binding modes by docking is difficult. In addition, the modifications to the base fragment may alter its binding modes fundamentally.

Hybrid Design Based on Known Inhibitors

Through many years of drug discovery efforts, pharmaceutical companies have synthesized millions of small-molecule inhibitors. A large amount of information has been accumulated on how these inhibitors bind to their protein targets and to the QSAR of these inhibitors. Hybrid design is an emerging technique that attempts to take full advantage of this information to design new inhibitors for existing or new protein targets (50). The essence of the hybrid design is to recombine known inhibitors for a particular target in a rational way to create new inhibitors for the same target or for a new target. In a manner analogous to the evolutionary recycling of protein domains, the hope for chemical evolution through hybrid design is to create...
new inhibitors by recombining pharmacophores derived from different classes of compounds. Hybrid design starts by using inhibitor-receptor binding information obtained experimentally from crystallography or computationally by docking to overlay multiple known ligands, and then to predict how to recombine inhibitor substructures to create new compounds. This strategy is illustrated in Fig. 5 in which a potent Aurora A kinase inhibitor can be thought of as a hybrid of two inhibitors developed originally for two other kinases. This strategy has also been applied successfully to design novel kinase inhibitors that bind to a specific inactive kinase conformation (51, 52). Gleevec is a Bcr-Abl inhibitor that has been demonstrated crystallographically to bind to the nucleotide binding cleft of Abl, which uses both the adenine binding region and an adjacent hydrophobic pocket created by the activation loop being in a unique inactive conformation (53). A hybrid strategy involved appending a 3-trifluoromethylbenzamide group to a known ATP binding site inhibitor (51). Of the hybrid compounds thus created, potent and selective inhibitors have been identified for Abl, c-Kit, p38, PDGFR, and Aurora kinases (52). Figure 5 shows how a benzamide group is attached to a quinazoline scaffold to create an inhibitor that binds to an inactive Aurora kinase conformation similar to that observed on Abl-Gleevec complex.

A computer program named BREED automates the process for hybrid design (54). The method imitates the common medicinal chemistry practice of joining fragments of two known ligands to generate a new inhibitor. The known active ligands are superimposed in their active conformation to identify all overlapping bonds, and the fragments on each side of each matching bond are swapped to generate a large set of novel inhibitors. This method has been demonstrated to have a high rate of success to identify novel inhibitors for HIV protease and protein kinases (54).

**De Novo Design**

Small molecule inhibitors can also be generated by de novo design without the prior knowledge of other active ligands. In essence, the procedures are similar to those mentioned previously. For example, LeapFrog can start without a given small molecule. It will select a core randomly and will grow a ligand from its fragment library. Although it is still rare for a computer program to generate practical and active molecules from scratch, reports of novel molecules from “raw” output of automated computer program exist. SkelGen (55) was able to generate small-molecules inhibitors for estrogen receptor binding sites without hints from known actives. Among the 17 highest scoring structures, 5 structures were active with 4 structures that possess novel structures. We have also implemented a de novo ligand design algorithm in our in-house software GRodE using an evolutionary algorithm. A molecule is represented by a unique genetic code that is subject to normal genetic operations such as mutation, addition, deletion, and translocation. The population of ligand molecules is evolved toward increasing fitness. Our retrospective studies have shown that the method can produce lead molecules that are extremely similar to experimental nanomolar inhibitors.
Small Molecule Inhibitors, Design and Selection of

If the target binding site structure is unavailable, de novo design can be carried out based on known active molecules. This circumstance may occur because none of the known actives that pose desired properties or that pose novel inhibitors are sought for patentability. However, the known inhibitors provide a basis to construct 3-D pharmacophore and other models such as Comparative Molecular Field Analysis (CoMFA) that can provide a template to guide the computational algorithm to maximize the fitness of the molecule “population” toward a given model. For example, Leapfrog can use CoMFA models in place of target binding site structures. It is also straightforward to apply pharmacophore models to novel molecular design with and without starting fragments. Novel structures are chosen based on its fitness to the given model. Unlike de novo design with known binding site structures, it is important to minimize the possibilities for resulting molecules to have steric clashes with receptor atoms. Even though these programs have produced promising results, the major drawback for such automated computer algorithms is that the “raw” computer derived molecules often are undesirable synthetically or implausible chemically. Some programs (55) have implemented chemistry rules to guide the generation of new molecules, but it is still far from satisfactory. Moreover, because a drug molecule interacts with its intended target in a complex biologic environment, a good lead molecule is selected by not only its binding affinity to its intended target, but also other properties such as crucial physical chemical properties, ability to expand SAR, selectivity relative to undesired off targets such as cytochrome P450 enzymes, and ion channels such as HERG. All are important considerations to advance a target in a complex biologic environment, a good lead molecule is selected by not only its binding affinity to its intended target, but also other properties such as crucial physical chemical properties, ability to expand SAR, selectivity relative to undesired off targets such as cytochrome P450 enzymes, and ion channels such as HERG. All are important considerations to advance a molecule toward clinical development. Various computational methods have been developed to predict these properties; however, normally they are not integrated into automated de novo design program.

Hit and Lead Optimization

Computational algorithms and modeling have advanced significantly over the last two decades in terms of accuracy, efficiency, and ease of use. Many modeling procedures are automated; however, they are far from replacing the intuition and knowledge of an experienced computational chemist. Although automated programs are used widely in early stages of lead discovery with limited human intervention, lead optimization is a process in which a computational chemist makes contributions out of his knowledge, insight, experience, and intuition. Before lead optimization, computational algorithms are designed to explore wide areas of chemical space and a wide range of activity data. During lead optimization, focus is placed on one or a few particular compound series. Small modifications are introduced to investigate local SAR and to improve other pharmacologic properties. Incremental improvements are often the objectives at this stage. Whereas the automated computational methods described previously are still important, they are not sufficiently precise to distinguish between modifications that result in 2 kcal/mol changes in binding affinities. Manual intervention, such as visual inspection and manual docking from an experienced chemist, are often of great importance for this phase.

Structure Based

As we can observe from previous discussion, protein structures have played invaluable roles in small-molecule inhibitor selection and design. Currently, over 40,000 biomacromolecular structures are deposited in RCSB protein databank. Over 5000 new structures were deposited in 2006. The atomic level structural information provides an enormous wealth of mechanistic insights with regard to protein function. A approximately 6500 protein structures in the database are in complex with an inhibitor or substrate to provide chemists with important insights into the mechanism of ligand recognition. Because biologic systems are very diverse, each system requires specialized knowledge and treatment. Successful ligand design requires an intimate knowledge of all known inhibitors, which include how their activity and selectivity is changed as a function of chemical modification (SAR), rules for which interactions are required for high affinity binding, knowledge about the potential for conformational rearrangements, and insight into what is tractable synthetically. Success therefore requires the close collaboration of computational chemists, medicinal chemists, and structural biologists. Here, we will attempt only to describe a few common techniques that computational chemists use to optimize the activity of a small-molecule inhibitor.

Visual inspection is probably one of the oldest and most powerful methods in computer modeling. Large amounts of information can be acquired by looking at the 3-D structures of an inhibitor and a protein complex, such as important hydrogen bonds, hydrophobic contacts, and strong electrostatic interactions. It allows one to design a derivative to make stronger or additional hydrogen bonds, to explore unoccupied pockets, and even to alter the scaffold while maintaining important interactions. Of course, regular molecular modeling methods are used here to ensure the proposed molecules possess sensible conformations. In contrast to the automatic molecular generating programs mentioned previously, the modifications here are proposed directly by chemists; therefore, the molecules are tractable synthetically and often follow a particular experimentally derived SAR series. The iterative use of model building, synthesis, and biologic evaluation is the most powerful and efficient means to obtain small molecules with a desired biologic activity. When multiple crystal structures with different ligands exist, one can gain even more insights into conserved interactions. This insight can be very useful to identify key interactions and to help design new compounds to maximize possible key interactions. An algorithm developed by Deng et al. (56) uses a binary string to represent ligand protein interactions. By clustering the binary strings of different small-molecule binders, the authors were able to illustrate the similarity and diversity of small-molecule binding modes. This information can be used to select correct binding modes as well as to propose new molecules.

“Anchored docking” is another common technique used in lead optimization. Unlike automated docking, “anchor” points such as particular hydrogen bonding are specified by the user to limit the configuration space sampled by docking program. This limit is to ensure that the knowledge of a chemist about the system is enforced, because usually automated docking is sufficiently accurate to find correct binding modes consistently.
The "anchor" points can be defined for a receptor, a small molecule, or both. The correct definition of "anchor" points largely relies on the users’ knowledge of the system.

**Qsar Based**

Even with an ample amount of protein structures available, many interesting protein targets without 3-D structural information exist. GPCRs are typical examples. In this case, only SAR derived from the biologic assay can be used to guide optimization. Under these circumstances, QSAR has been an important tool. It is probably one of the earliest tools for computer aided drug design. Properly choosing descriptors for QSAR models is much like an art and requires a deep understanding of the system of interest. Most common descriptors are based on physical, chemical, geometric, and topological properties. Three-dimensional QSAR models may require active ligands to be aligned properly just like the pharmacophore model building. In fact, 3-D QSAR models can be generated directly by pharmacophore models, and many pharmacophore-generating programs provide this capability. One method developed early and widely used for 3-D QSAR is CoMFA.

**Design of Selective Small Molecule Inhibitors**

For a small-molecule inhibitor to be useful as a therapeutic drug or as a tool for chemical biology study, it must have specificity for its intended target. A small-molecule drug often causes side effects or toxicity if it also modulates other proteins in the body (i.e., off-targets). Nonselective inhibitors will confound the effects or toxicity if it also modulates other proteins in the body. Therefore, the selectivity is a very important feature when considering design and selection of small-molecule inhibitors. Because ≈30,000 proteins are encoded in the human genome, which can possess multiple ligand binding sites and can be present in vastly different concentrations, finding specific inhibitors can be an extremely difficult task. Aalyzing the sequence and structural differences between a targeted protein and its closely related homologs is an effective way to design selectivity rationally.

**Targeting the Differences in the Binding Pocket**

The binding affinity of a ligand is determined by its complementarity to the size, shape, and physicochemical properties of the protein-binding site. To design selective small-molecule inhibitors rationally, the 3-D structure of the target-binding site is required. In principle, the structure of a protein-binding pocket is determined by its amino acid sequence, and it is straightforward to identify unique amino acids by sequence alignment. Once the unique residues are identified, they can be analyzed in the context of small-molecule binding to determine the residues that are important for protein ligand interaction. The inhibitors can then be designed to target them specifically for selectivity. Taunton et al. have used this approach to design selective inhibitors for p90 ribosomal protein S6 kinases (RSKs) (57). They identified two key selective determinants for RSK’s relative to other kinases. One determinant is a small "gatekeeper" amino acid, and the other is a Cys in the glycine-rich loop. Targeting these two residues, they were able to make very potent and selective inhibitors for RSKs.

Although successful for RSKs, it is often very difficult to identify unique residues suitable for selectivity for all proteins in a family. In many cases, a mono-selective inhibitor among the protein family members is not feasible, so selectivity against a subset of the family members is desired. In this case, selectivity can be achieved by targeting the sequence differences in the binding sites between the target protein and the unwanted proteins. This strategy has been applied successfully to design selective inhibitors for targets within a closely related subfamily, such as p38 kinase α, β, γ, and δ isoforms (58); protein kinase CDK2, and CDK4 (59); and COX-1 and COX-2 (60).

For protein targets that are closely related, the binding sites may look the same and no sequence difference is around the binding pocket. Therefore, a small molecule that binds to this primary site has very little selectivity. In these cases, a commonly used strategy is to search for a secondary site near the primary binding pocket. If it is less conserved, selective inhibitors can be designed by targeting the secondary site. For example, it is very difficult to develop a selective protein tyrosine phosphatase-1 B (PTP-1B) inhibitor, but a crystal structure of PTP-1B in complex with bis(para-phosphonyl) methane reveals a secondary aryl phosphate-binding pocket adjacent to the active site that can be targeted to design selective PTP-1B inhibitors (61).

Many protein structures are highly dynamic, and induced-fit effects are well known in small-molecule and protein binding (21). Although two proteins can have very similar binding pockets in static form, they can have different plasticity. Although protein flexibility is one major obstacle to predict the binding mode and the affinity of a ligand accurately, it provides a structural basis to design selective inhibitors. Usually, the conformations of the active enzymes within a family are more alike, but the inactive conformations are more different. For example, the ATP binding pockets of protein kinases in active states are highly similar, but considerable conformational variability exists among kinases in the inactive state that can be exploited for selective inhibitor design. The unique “DFG-out” inactive conformation has been applied already to the design of selective kinase inhibitors, which includes the already approved drug Gleevec and Nexavar. Lapatinib, one of the most selective kinase inhibitors, achieves its high selectivity by binding to an inactive conformation of EGFR with a methionine in the α-C helix moved away from its normal position to accommodate the beryloxy group of lapatinib (62). Exploiting differentially accessible conformational states among closely related targets has also been the structural basis for observed high selectivity of PI3K kinase inhibitors (63).
Identifying a selective small-molecule inhibitor can be time consuming. For a target in a large protein family, it is almost impossible to achieve monoselectivity. Inhibiting multiple targets by a small molecule leads to a significant complication in the analysis of target validation and studies of protein functions. To overcome the difficulty of identifying monoselective small-molecule inhibitors, a chemical genetic approach is applied to design as orthogonal receptor-ligand pairs (64, 65). In this approach, the targeted protein is mutated in the active site to create a structural distinction between the target and all other members of proteins in the same family. Then, a small molecule is designed to target the distinct structure so that it only binds the mutated protein. The first step is to study the binding mode of a small molecule to its target and identify an important interaction that is highly conserved for the ligand binding to all members in the protein family. Then, the ligand is modified in the conserved interaction region to disrupt the interaction so that it cannot bind effectively to any wild type protein, which includes the target. Two modifications are used commonly. One is to add a bulky substitute (a “bump”) at the interacting position to create steric clashes, and the other is to disrupt the critical hydrogen bonding between the ligand and the protein. The next step is to mutate the residues around the bump to smaller ones (which creates a “hole”) to accommodate the “bump” or to mutate residues to restore the critical hydrogen bonds. Although both approaches have introduced orthogonal protein-ligand pairs successfully, the “bump-hole” strategy is more suitable for design and is applicable to many different protein families. Shokat’s lab has successfully applied the “bump-hole” strategy to generate monospecific inhibitors for engineered kinases (64). The monospecific inhibitors have been used widely in kinase signaling pathway elucidation and target validation (66). Figure 6 outlines the strategies and workflows that we discussed for selective inhibitor design, in which the method details of each step in the flow chart have to be tailored towards specific systems.

**Summary**

In this review, we described briefly some common computational techniques to select and to design small-molecule inhibitors. It is not a comprehensive list of all methods available. Instead, we focused on basic methods that are used commonly in drug discovery research. Because of the size limitation, it is also impossible to discuss all the details in each method. We hope that this short review can provide some fundamental concepts used in virtual screening and rational design, and can provide the interested reader with an entry point to explore the field of computational inhibitor design and selection. Although we described the computational methods in a linear flow, the methods used in one stage of small-molecule inhibitor discovery process can be applied at different stages whenever...
appropriate. To choose the proper methods for different purposes at different stages is crucial for computational modeling to be effective. In addition, computational modeling should always interact tightly with experimental investigation in order to make significant contributions. VS and rational design is also an ever-changing field. New algorithms and strategies are emerging continuously. We believe that computational methods will play increasingly important roles in discovering new small-molecule inhibitors.

References

Small Molecule Inhibitors, Design and Selection of


Further Reading


See Also

Computation and Modeling of Molecular Recognition
Intermolecular Interactions in Molecular Recognition
Mechanisms of Enzyme Inhibition
Protein-Ligand Interactions
Structural and Functional Aspects of Molecular Recognition
Stem cells hold great promise for the treatment of many devastating diseases and will also provide new insights into the molecular mechanisms that control developmental and regenerative processes. Realization of the therapeutic potential of stem cells will require a better understanding of the signaling pathways that control stem cell fate as well as an improved ability to manipulate stem cell proliferation, differentiation, and reprogramming. Cell-based phenotypic and pathway screens of synthetic compounds have led recently to the discovery of several small molecules that can be used to control stem cell fate. Such molecules will not only provide new insights into stem cell biology but will also facilitate the development of therapeutic agents for regenerative medicine.

Given appropriate conditions, stem cells can self-renew for long periods of time while maintaining the ability to differentiate into various functional cell types in the body (1). It is these characteristics that not only make stem cells a useful system in which to study tissue and organ development but also give them great potential for regenerative medicine. Given the success of well-practiced, cell-based therapies (e.g., hematopoietic stem cell transplantation for treating hematological diseases and pancreatic islet cell transplantation for type I diabetes), it is conceivable that this approach could be applied to many other serious medical conditions where cells are lost because of disease, injury, or aging. Current challenges in cell replacement therapy include the limited source of engraftable stem/progenitor cells and a poor ability to manipulate their functional expansion and differentiation. Alternatively, drug stimulation of the body’s own regenerative capabilities (i.e., proliferation, differentiation, migration, or even reprogramming of endogenous cells to replace the damaged cells/structures) may represent a more desirable therapeutic approach to fulfilling the ultimate goal of regenerative medicine. Although tremendous efforts have been put into these areas, it is clear that a better understanding of stem cell biology is required before these approaches can be realized.

Small molecules have long been associated with biological discoveries. Our understanding of biological processes often develops from discovering or designing ways to perturb a given process and observing the effects of the perturbation. Although genetic approaches have been widely used for this purpose, the small-molecule approach clearly offers some distinct advantages. For example, small molecules provide a high degree of temporal control over protein function, which generally acts within minutes or even seconds, and their effects are often reversible, which facilitates both rapid inhibition and activation. Their effect can also be finely tuned by varying concentrations of the compound of interest. Moreover, because of the inherent difficulty of genetic manipulation for many types of stem cells (e.g., low transfection efficiency or poor clonal expansion), small-molecule tools are especially useful for the stem cell field. In this article, recent developments in the use of small molecules (Fig. 1) on the various aspects of stem cell biology such as self-renewal, differentiation, and reprogramming will be reviewed.

Biological Background of Stem Cells

Stem cells normally are classified, based on their origin and differentiation capacity, as either embryonic or adult stem cells (1). Embryonic stem cells (ESCs) are derived from the inner cell mass of the blastocyst. ESCs can self-renew indefinitely and are pluripotent—the ability to differentiate into all cell types in the embryo proper. Adult stem cells are undifferentiated (unspecialized) cells that are found as differentiated, or specialized, tissue. They have limited self-renewal capability and generally can only differentiate into the specialized cell types of the tissue in which they reside. These cells function as the “reservoir” for cell/tissue renewal during normal homeostasis or tissue regeneration. Sources of adult stem cells have been found in most tissues, including bone marrow, blood stream, cornea...
Small Molecules to Control Stem Cell Fates

and retina of the eye, dental pulp of the tooth, liver, skin, gastrointestinal tract, lung, pancreas, heart, and brain. For example, the bone marrow is a mesoderm-derived tissue that consists of a complex hematopoietic cellular system supported by stromal cells embedded in a complex extracellular matrix. A growing body of evidence suggests that bone marrow contains at least two types of stem cells: hematopoietic stem cells (HSCs) and mesenchymal stem cells (MSCs), both of which are multipotent. HSCs have the capacity to provide life-long reconstitution of all blood-cell lineages after transplantation, whereas MSCs

Figure 1 Chemicals that control stem cell fate.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

- 1972 -
have the ability to differentiate, both in vitro and in vivo, into a variety of adult mesenchymal cell types, including osteoblasts, chondrocytes, and adipocytes.

Stem cell fate in vitro is under strict control from both intrinsic and extrinsic factors, and loss of this control has been postulated to be a key step in degenerative and carcinogenic processes (2). Emerging evidence suggests cancer initiation results from an accumulation of oncogenic mutations (intrinsic loss of control) in long-lived stem cells or their immediate progenitors, followed by modification of the surrounding microenvironment (loss of extrinsic control) (3). Cancer stem cells have been detected in leukemia, breast, and brain tumors (4–7). They may originate from resident stem cells or develop as a result of a gain of self-renewal capacity in tissue progenitor cells. Therefore, the characterization of a cancer stem cell profile within diverse cancer types, and a better understanding of the biology of its counterpart, the normal stem cell, may open up new avenues for cancer treatment.

Stem cell expansion and differentiation in vitro commonly are controlled by culturing cells in a specific configuration (e.g., an attached monolayer or as suspended aggregates) with cocktails of growth factors and signaling molecules as well as genetic manipulations. However, most of these conditions are either incompletely defined or are nonspecific and inefficient at regulating the desired cellular process. Such conditions often result in inconsistency in cell culture and mixed populations of cells that would not be useful in studying specific cellular processes or in cell-based therapies. More efficient and selective conditions for homogeneous stem cell self-renewal and differentiation into specific cell types need to be developed before the various applications of stem cells can be realized. Towards this end, chemical approaches serve as excellent tools to investigate the underlying mechanism and to control the specific stem cell fate.

### Chemical Approaches to Stem Cell Research

Conceptually, two approaches exist for discovering chemical compounds in stem cell biology. The target-based approach involves development of chemical compounds for specific biological targets and the application of these compounds in order to link their targets’ modulation to the produced pharmacologic effect in cells or organisms. In the phenotype-based approach, small-molecule libraries are screened in high-throughput functional assays (in cells or whole organisms) to identify compounds that produce a desired phenotype, followed by elucidation of the molecular targets or pathways that they engage.

The success of both target-based and phenotype-based methods relies heavily on the qualities of the chemical libraries used. Although combinatorial technologies allow the synthesis of a large number of molecules with immense structural diversity, it is impossible to saturate the chemical space, which has been estimated to contain more than $10^{60}$ molecules (8). Because biological space interacts with only a fraction of chemical space, synthetic attempts to increase randomly the molecular diversity of a chemical library by introducing a high level of structural variability/complexity drastically reduce the libraries’ fitness to a given biological selection/screen, which results in most molecules being inactive. Furthermore, screening of even larger chemical collections to increase the chance of finding hits can compromise assay functionality, sensitivity, and fidelity for practical reasons (9). As a result, the screening of a highly defined set of compounds in stem cell biology becomes a critical aspect of combinatorial synthesis.

The notion of “privileged structures” describes selected structural motifs that can provide potent and selective ligands for biological targets (10). Privileged structures typically exhibit a greater tendency of interacting with biological targets and good “drug-like” properties (11, 12). One of the most straightforward and productive ways to generate “privileged” chemical libraries is to use key biological recognition motifs as the diversity elements for combinatorial synthesis (13). In this approach, a variety of naturally occurring and synthetic heterocycles that are known to interact with proteins involved in cell signaling (e.g., kinases, cell surface and nuclear receptors, or enzymes) are used as the core molecular scaffolds. Then, a variety of substituents can be introduced into each of these scaffolds to create a diverse chemical library. Using this method, a diverse heterocycle library that consists of over 100,000 discrete small molecules (representing over 30 distinct structural classes) was generated with an average purity >90%, which has proven to be a rich source of biologically active small molecules targeting various proteins involved in a variety of signaling pathways.

The screening method is another important factor for the phenotypic approach. Although the use of simple reporter systems or enzymatic activity assays would allow higher throughput in cell-based assays, the monitoring of more complex phenotypic changes (e.g., cell morphology, multiple biomarker expression and localization, and cell physiology) by high content imaging methods substantially reduces false analysis from primary screens and provides broader assay versatility (14, 15). Furthermore, informatics on the matrix of assays/compounds/genes has facilitated bioactive compound identification and deconvolution of their mechanism.

### Small-Molecule Regulators of Stem Cell Fate

#### Self-renewal

Self-renewal is the process by which a stem cell divides to generate one (asymmetric division) or two (symmetric division) daughter stem cells with identical developmental potentials as the mother cell (16). The self-renewal ability of stem cells is central to development and the maintenance of adult tissues.

Self-renewal of stem cells can be regarded as a combined phenotype of cellular proliferation and inhibition of differentiation and cell death. Consistent with this notion are the findings that self-renewal of murine embryonic stem (mES) cells can be achieved in the absence of feeder cells and serum in a chemically defined medium condition by the combined activity of two key signaling molecules: LIF/interleukin 6 (IL6)
family members and bone morphogenetic protein (BMP) (17). LIF activates STAT (signal transduction and activation of transcription) signaling through a membrane-bound gp130-LIFR (LIF receptor) complex to promote proliferation and to inhibit mesoderm and endoderm differentiations of mES cells via a Myc-dependent mechanism. BMP induces expression of Id (inhibitor of differentiation) genes via Smad signaling and inhibits differentiation of mES cells to neuroectoderm.

To gain a better understanding of ESC self-renewal, a cell-based screen for small molecules that can promote mESCs self-renewal was carried out using an established Oct4-GFP reporter mESC line and examining pluripotency marker expression (Oct4) as well as morphological analysis (undifferentiated ESCs having compact colony morphology). A novel, synthetic small molecule named pluripotin was discovered in this screen that is sufficient to propagate mESCs in the pluripotent state under chemically defined conditions in the absence of feeder cells, serum, and LIF. Long-term cultures of mESCs with pluripotin can be differentiated into cells in the three primary germ layers (ectoderm, mesoderm, and endoderm) and they can also generate chimeric mice and contribute to the germ line in vivo. Affinity chromatography using a pluripotin-immobilized matrix identified ERK1 and RacGAP as the molecular targets of pluripotin. Additional biochemical and genetic experiments confirm that pluripotin is a dual-function, small-molecule inhibitor of both ERK1 and RacGAP, and that simultaneous inhibition of both protein activities is necessary and sufficient for pluripotin’s effects on mESCs. Because pluripotin’s mechanism of action is independent of LIF, BMP, and Wnt signaling, this study suggests that ES cells may possess the intrinsic ability to self-renew and that inhibiting differentiation pathways would be sufficient for maintaining the undifferentiated state. Thus, not only did the discovery of pluripotin provide interesting insights to the mechanism of ESC self-renewal, but it also exemplified the power of small-molecule screens in that it is possible to modulate more than one target by a small single molecule to achieve a desired complex biological phenotype. In addition to the unbiased phenotypic screen described above, based on known mechanistic insights, p38 (e.g., SB203580), MEK (e.g., PD98059), or GSK3 (e.g., BIO) inhibitors have been shown to enhance self-renewal of mESCs in the presence of additional signaling inputs (18–20).

Consistent with the self-renewal model, recent studies on Notch signaling exemplified the contribution of cell survival signaling to stem cell self-renewal. Andrountelis-Theotokis et al. (19) found that activation of Notch promotes fetal neural stem cell (NSC) survival, most likely through activation of AKT, STAT3, and mTOR; this survival signal is antagonized by JAK and p38 MAPK, and applying JAK and p38 inhibitors significantly improves survival of NSCs. Interestingly, this mechanism also functions in human ESCs. Moreover, p38 inhibitors have also been shown by another independent study to promote HSC life span, which confirms even more the role of p38 in stem cell self-renewal (21). Although the details of the molecular mechanism still need to be worked out, these inhibitors clearly have had a beneficial effect on stem cell culture.

Most stem cells can divide via either asymmetric or symmetric modes (22). Because each asymmetric division generates one daughter with a stem-cell fate (self-renewal) and one daughter that differentiates, it was postulated that asymmetric division could be a barrier to stem cell expansion. Therefore, conversion of asymmetric division to symmetric division would promote self-renewal and long-term culture of stem cells. Indeed, this idea was confirmed by using one small molecule, the purine nucleoside xanthosine (Xs) (23). This small molecule promotes guanine nucleotide biosynthesis that reversibly converts cells from asymmetric division kinetics to symmetric division kinetics. It was found that Xs derived from stem cell lines exhibit Xs-dependent symmetric kinetics, and this derived stable line shows enhanced self-renewal. This study underscores the importance of balance between the two modes of division, both to stem cell expansion and to the regenerative capacity of adult stem cells.

One major challenge in hESC culture is a low survival rate after cell dissociation. Genetic manipulation (e.g., gene-targeting), and to a lesser extent, routine culture and directed differentiation are all reliant on clonal survival and/or cell association. In a small screen, a selective small-molecule inhibitor of p160 rho-associated coiled-coil kinase (ROCK), Y-27632, was found to increase hESC survival. The mechanism of action by which Y-27632 inhibits apoptosis, which is yet to be identified, should yield insights as to the causes of poor survival after dissociation (24).

Recent studies have identified multipotent isl1+ cardiovascular progenitors (MICPs) in mouse embryos, and they have also been cloned from mESCs (25). MICPs can produce the three main cell types of the heart: cardiac muscle, smooth muscle, and endothelial cells. To gain a better understanding of mechanisms that regulate the self-renewal and differentiation of MICPs, a high-throughput screen was completed to identify small molecules capable of inducing expansion of isl1+ MICPs (26). Several molecules were identified that significantly increased MICP expansion, with the potent GSK-3beta inhibitor BIO being one of the strongest. Furthermore, replacement of BIO and CMCs with a Wnt3a-producing feeder layer also facilitated MICP expansion, whereas treatment with Dickkopf-1 (DKK-1), which is an extracellular inhibitor of Wnt signaling, decreased the number of isl-1+ progenitor cells. Interestingly, BIO treatment also induced expansion of human isl1+ progenitors, which suggests a conserved role of Wnt signaling in self-renewal of MICPs.

**Lineage-specific differentiation**

Differentiation is the developmental process by which early multipotent cells acquire the features of late-stage, mature cells such as neurons, hepatocytes, or heart muscle cells. Currently, few examples of devised, highly selective, and efficient conditions for stem cell differentiation into specific homogenous cell types have been reported because of a lack of understanding of stem cell signaling at the molecular level. Small-molecule phenotypic screens provide another means to generate desired cell types in a controlled manner. Several small molecules have been identified by this method that modulate specific differentiation pathways of embryonic or adult stem cells.
Neural and neuronal differentiation

Retinoic acid (RA), forskolin, and HDAC inhibitors have been shown to induce neural differentiate of hippocampal adult neural progenitor cells. However, these factors are either pleiotropic or undefined physiologic relevance: RA has neuronal subtype patterning activity and also been reported to induce cardiac differentiation (27) as well as pancreatic differentiation (28). HDAC inhibitors are nonspecific; and forskolin activates protein kinase A and serves to increase the cellular levels of the general signaling molecule cAMP (29).

Neuropathiazol, which is a substituted 4-aminothiazole, was identified recently from a high-content imaging-based screen of chemical libraries that specifically induces neural differentiation of multipotent adult hippocampal neural progenitor cells (30). More than 90% of the neural progenitor cells treated with neuropathiazol differentiated into neuronal cells as determined by immunostaining with βIII tubulin and the characteristic neuronal morphology. Interestingly, neuropathiazol can also inhibit astroglial differentiation induced by LIF and BMP2, whereas RA cannot, which suggests that neuropathiazol functions by a different mechanism and has more specific neurogenic inducing activity. The precise molecular target(s) of neuropathiazol is still unknown, although its identification will surely provide mechanistic insights into neuronal differentiation. Additionally, neuropathiazol can be used as a specific inducer to generate homogeneous neuronal cells, and it may serve as a starting point for development of small-molecule drugs to stimulate in vivo neurogenesis.

In an elegantly devised approach, functional motor neurons were generated from mESCs by sequential treatments with RA (neutralizing and catalyzing mESCs) and a specific small-molecule agonist (Hb-Ag1.3) of Hedgehog (Hh) signaling (ventralizing the caudalized cells) (31). This experiment underscores the importance of following developmental progression through sequential induction and combinatorial factors for generating a late-stage cell type from early stem cells.

Differentiation of mesenchymal stem/progenitor cells

Mesenchymal stem cells (MSCs) are capable of differentiating into all mesenchymal cell lineages, such as bone, cartilage, adipose, and muscle, and play important roles in tissue repair and regeneration. Many small molecules have been used to control the differentiation of mesenchymal stem/progenitor cells to a variety of cell types. For example, 5-aza-C (a DNA demethylating agent) can induce C3H10 T1/2 cells (a mouse mesenchymal progenitor cell line) to differentiate into myoblasts, osteoblasts, adipocytes, and chondrocytes by enhancing cell differentiation competence via epigenetic modifications. Dexamethasone (a glucocorticoid receptor agonist) is another kind of epigenetic modifier, and its combination with other small molecules, such as ascorbic acid, β-glycerophosphate, isobutylmethylxanthine (IBMX, a nonspecific phosphodiesterase inhibitor), or peroxisome proliferator-activated receptor γ (PPARγ) agonists (such as rosiglitazone) have been used widely to modulate osteogenesis or adipogenesis of MSCs under specific conditions (32, 33). To identify small molecules that selectively induce osteogenesis of MSCs, a high-throughput screen of chemical libraries in C3H10 T1/2 cells using an enzymatic assay of alkaline phosphatase (a specific osteoblast marker) led to the identification of a novel synthetic small molecule, purmorphamine (13). Genome-wide expression profiling in conjunction with systematic pathway analysis were used to reveal that the Hh signaling pathway is the primary affected biological network by purmorphamine (34). Additionally, studies, including chemical epistasis using known Hh pathway antagonists (cyclopamine and forskolin), have confirmed that purmorphamine’s mechanism of action on osteogenesis is through specific activation of the Hh pathway, and it acts at the level of Smoothened (Smo) (35). More recently, it was shown that purmorphamine directly targets the protein Smoothened (Smo) through competitive displacement assays using fluorescently-tagged cyclopamine. Recently, a small-molecule screen completed in zebrafish embryos identified prostaglandin E2 (PGE2) synthesis as a novel modulator of vertebrate HSC function (36). Molecules that up regulated PGE2 synthesis increased HSC function, whereas those that lowered PG synthesis decreased HSC function. Moreover, treatment with a stabilized 16,16-dimethyl derivative of PGE2 improved kidney marrow recovery after irradiation in the adult zebrafish. Furthermore, PGE2 also induced an amplification of multipotent progenitors derived from mESCs, which demonstrates a conserved role in vertebrate species. These studies suggest that modulation of the PGE2 pathway could potentially be used to treat patients undergoing bone marrow transplant or to treat anemia.

Regeneration

Terminally differentiated, post-mitotic mammalian cells are thought to have little or no regenerative capacity, as they are already committed to their final specialized form and function and have permanently exited the cell cycle (36). However, it is conceivable that appropriate stimulation of mature cells to re-enter the cell cycle and proliferate may provide new therapeutically approaches for treating various degenerative diseases and injuries. The mammalian cardiomyocyte is one such mature cell type that has attracted substantial research efforts toward its regeneration. Using a target-based approach, a p38 MAPK inhibitor, SB203580, and a GSK3 inhibitor, BIO, have been shown independently to promote proliferation in both neonatal and adult cardiomyocytes indicated by BrdU incorporation and histone 3 phosphorylation (37, 38). The proliferation in adult cardiomyocytes was also observed to be associated with transient dedifferentiation of the contractile apparatus. Activation of canonical Wnt signaling by BIO also promotes proliferation of cardiac progenitor cells. The pancreatic β cell is another highly sought cell type for regeneration, the transplantation of which, in conjunction with simultaneous prevention of their immune destruction, may represent a “cure” for type 1 diabetes. Recent phenotypic screens of large chemical libraries have identified several classes of small molecules that can promote proliferation of human β cells, among which p38 inhibition was identified as the mechanism of action for one class of molecules. A major challenge lying ahead for proliferation of mature cell types is that the process typically is associated
with loss of the cell phenotype (e.g., proliferated β cells under
dergo epithelial-to-mesenchymal transition (EMT) to acquire
fibroblast-like features, and they typically do not redifferentiate
back to the mature β cells). Strategies for inducing functional
redifferentiation of the proliferated cells or for inhibiting loss
of cell identity while proliferating are highly desirable and are
under intense investigations. It should be noted that a synthetic
purine analog, myosentin, was identified previously from a
phenotypic screen, which can induce cleavage of multicellular
myoblasts to generate myoblast-like cells, which can proliferate
and redifferentiate into myotubes (39).

In the mammalian CNS, failure of axonal regeneration is
attributed not only to the intrinsic regenerative incompetence
of mature neurons but also to the inhibitory actions of CNS
myelin and molecules in the glial scar at the lesion sites. In a
cell-based screen to identify small molecules that can counteract
myelin and molecules in the glial scar at the lesion sites. In a
phenotypic screen, which can induce cleavage of multicellular
myoblasts to generate myoblast-like cells, which can proliferate
and redifferentiate into myotubes (39).

Cellular reprogramming

It was long thought that tissue/organ-specific stem/progenitor
cells could give rise only to cells of the same tissue type
but not to those of different tissue. In other words, they have
irreversibly lost the capacity to generate cell types of other
lineages in the body. However, recently several reports have
demonstrated that tissue-specific stem cells may overcome their
intrinsic lineage-restriction during exposure to a specific set of
in vitro culture or in vivo conditions. For example, recently the
Yamanaka group (41) has demonstrated that cells from mouse
embryonic fibroblast culture can be induced into a pluripotent
state during transduction with four genes (Oct-3/4, c-Myc, Sox2, and
Klf-4) in vitro. These results were extended recently to
demonstrate germline transmission of the induced pluripotent
cells (42). An extreme example is the reprogramming of a so-
matic cell to a totipotent state by nuclear transfer cloning, where
the nucleus of a somatic cell is transferred into an enucleated
oocyte or the extracts of the oocyte are fused with a somatic
nucleus (43, 44). Although in mammals neither transdifferentiation
nor dedifferentiation has yet been identified as a naturally oc-
curring process (except in certain disease states), the discovery
of cell plasticity raises the possibility of reprogramming a re-
stricted cell’s fate. The ability to dedifferentiate or reprogram
lineage-committed cells to multipotent or even pluripotent cells
might overcome many obstacles associated with using ESCs
and adult stem cells in clinical applications (e.g., immunocom-
patibility, cell isolation and expansion, or bioethics).

To identify small molecules that induce reprogramming of
lineage-committed myoblasts, a cell-based screen was designed
based on the notion that lineage-reversed myoblasts would regain multipotency. Specifically, dedifferentiated myoblasts
would acquire the ability to differentiate into (otherwise nonper-
mitted) mesenchymal cell types under conditions that typically
induce differentiation of only multipotent MSCs into adipocytes,
osteoblasts, or chondrocytes. A two-stage screening protocol
was used in which C2C12 myoblasts were treated initially with
compounds to induce dedifferentiation; compounds were then
removed, and cells were assayed for their ability to undergo os-
togenesis during addition of known osteogenic inducing agents.
Reversine, which was a 2, 6-disubstituted purine, was found to
have the desired dedifferentiating inducing activity (45). It in-
hibits myotube formation, and reversine-treated myoblasts can
redifferentiate into osteoblasts and adipocytes during exposure
to the appropriate differentiation conditions. In addition, the re-
programming effect of reversine on C2C12 cells (as well as some
other cell types) can be shown at the clonal level, which
suggests its effect is inductive rather than selective. Further-
more, reversine has also been shown to induce reprogramming
of primary mouse and human fibroblasts to a multipotent state,
which can be redifferentiated to functional myoblasts and my-
ocytes under myogenic conditions, which suggests that rever-
sine’s reprogramming mechanism might be general to different
cell types (46).

Affinity chromatography experiments revealed that nonmuscle
myosin II heavy chain (NMMII) and MEK1 are cellular
targets of reversine. Mechanistic analysis demonstrated that
reversine acts by inhibiting both NMMII and MEK1. Inhibi-
tion of NMMII induces G2/M phase accumulation/staging and
cytoskeletal remodeling, whereas MEK1 inhibition serves to
modulate signaling, including acetylation of histone H3. Addi-
tionally, PI3K signaling was found to be essential to reversine
activity (47).

Modulators of developmental pathways
and epigenetic modifiers

Developmental signaling pathways, such as Wnt, Hb, TGF/
BMP, and Notch, control embryonic patterning and cell be-
avior during development and play important roles in regu-
lating tissue homeostasis and regeneration in adulthood. Known
chemical modulators of these developmental pathways have been used widely, including GSK3β inhibitors as agonists
of a canonical Wnt pathway, fumagillin as an antagonist of a
noncanonical Wnt pathway, parmporphamine as an agonist and
cyclophosphamine as an antagonist of an Hb pathway, DAPT as an
antagonist of a Notch pathway, and SB431542 as an inhibitor of
TGFβ/Activin signaling (18, 34, 48–50).

Although activation of Wnt, Hb, and Notch pathways is in-
volved in various regenerative processes, their abberant activities
are also associated with cancer induction. Strategies to control
the activity of these pathways for regeneration while avoiding
tumor induction are highly attractive. One possible approach
would be enhancing the desired pathway activity by a syner-
gistic agonist, which would be effective only at where it is
needed (i.e., the pathway activity is inadequate), rather than
using a general pathway activator, which ectopically activates
the signaling. To identify novel compounds and pathways that
interact with the canonical Wntβ-catenin signaling pathway,
a reporter-based screen was carried out recently for molecules

Small Molecules to Control Stem Cell Fates

Developmental signaling pathways, such as Wnt, Hb, TGF/
BMP, and Notch, control embryonic patterning and cell be-
behavior during development and play important roles in regu-
lating tissue homeostasis and regeneration in adulthood. Known
chemical modulators of these developmental pathways have been used widely, including GSK3β inhibitors as agonists
of a canonical Wnt pathway, fumagillin as an antagonist of a
noncanonical Wnt pathway, parmporphamine as an agonist and
cyclophosphamine as an antagonist of an Hb pathway, DAPT as an
antagonist of a Notch pathway, and SB431542 as an inhibitor of
TGFβ/Activin signaling (18, 34, 48–50).

Although activation of Wnt, Hb, and Notch pathways is in-
volved in various regenerative processes, their abberant activities
are also associated with cancer induction. Strategies to control
the activity of these pathways for regeneration while avoiding
tumor induction are highly attractive. One possible approach
would be enhancing the desired pathway activity by a syner-
gistic agonist, which would be effective only at where it is
needed (i.e., the pathway activity is inadequate), rather than
using a general pathway activator, which ectopically activates
the signaling. To identify novel compounds and pathways that
interact with the canonical Wntβ-catenin signaling pathway,
a reporter-based screen was carried out recently for molecules
that synergistically activate signaling in the presence of Wnt3a. A 2,6,9-trisubstituted purine compound, QS11, was identified to synergize with canonical Wnt ligand both in vitro and in vivo (51). Affinity chromatography identified ARF-GAP as a target of QS11. Additional functional studies have confirmed that QS11 inhibits the ARF-GAP function, and as a consequence, it modulates ARF activity and β-catenin localization. Because ARFs play important roles in endocytosis, this study established another link between the endocytosis pathway and Wnt signaling, and it provides a useful chemical tool to modulate the Wnt pathway and explore novel functions of ARF-GAPs in cell culture and whole organisms.

Epigenetic modifications are central processes in stem cell differentiation and reprogramming that can control specific and heritable gene expression patterns in cells without altering the DNA sequence. Therefore, molecules directly regulating epigenetic machineries and changing the epigenetic status of cells should affect cell fate. Major epigenetic modifications include DNA methylation and histone modifications (acetylation, phosphorylation, methylation, and ubiquitination). S-adenosylmethionine and its analogs are widely used DNA demethylation agents, and they have been shown to increase cellular plasticity or induce differentiation of certain stem/progenitor cells (52, 53). HDAC inhibitors (e.g., TSA and VPA) have also played essential roles in studying histone acetylation and have been developed for its analogs are widely used DNA demethylation agents, and they have been shown to increase cellular plasticity or induce differentiation of certain stem/progenitor cells (52, 53). HDAC inhibitors (e.g., TSA and VPA) have also played essential roles in studying histone acetylation and have been developed for the treatment of cancers. Like all signaling/epigenetic modifiers, their effects are context-dependent; they have been shown to enhance self-renewal of HSCs, induce neuronal differentiation of NSCs, or promote myogenesis of muscle cells (54–56). In addition to these two widely used epigenetic modifiers, small-molecule inhibitors for protein arginine methyltransferases (PRMTs) (41), histone methyltransferases (HMTs, e.g., Su(var)3-9, Set1/Brm) (42) and histone demethylases (e.g., LSD1) have been identified recently via various approaches. With ongoing studies of stem cell biology and contribute to the development of regenerative medicine.

Conclusion

Stem cell research provides tremendous opportunities for understanding human development, regeneration, and diseases, and it offers great promise for developing cell-based or drug therapies to treat devastating diseases and injuries. Small molecules have proven to be useful probes of stem cell biology, from stimulating homogenous self-renewal or differentiation conditions to identifying molecules that control stem cell fate. A cell initiating human acute myeloid leukemia after transplantation into SCID mice. Nature 1994;370:645–648.

References


Solid-Phase Synthesis of Biomolecules

Rolf Breinbauer, University of Leipzig, Germany
doi: 10.1002/9780470048672.wecb552

The synthesis of biopolymers, tool compounds, and molecular probes lies at the heart of any effort in chemical biology. Solution-phase strategies are well established but usually require considerable efforts in time and labor during workup and purification steps. In contrast, solid-phase synthesis, in which the substrate is covalently bound to an insoluble support, allows the simple removal of excess or consumed reagents by simple filtration. In the cases of oligonucleotides and peptides, the synthetic operations can be performed by robotic systems allowing the high-throughput synthesis of these biopolymers at low cost within a short time. The solid-phase synthesis of oligosaccharides and small-molecule probes has made significant progress during the last several years.

This article gives an overview of the main technical aspects of solid-phase synthesis and reviews the application of this technique for the synthesis of oligopeptides and proteins, oligonucleotides, oligosaccharides, and small molecules.

Biologic Background

From the mid-nineteenth century onward, researchers have made tremendous strides in understanding the molecular nature of living organisms. The discovery that cellular catalysts ("enzymes") are composed of oligomers of amino acids, that genetic information is encoded in oligonucleotides, the solving of the puzzle of metabolic pathways, and the investigation of hormonal and cellular signaling have shaped the modern view of what has become molecular and cellular biology. These discoveries have been accompanied by progress in both analytical techniques as well as in the synthesis of these molecules. Even today, the investigation of cellular organisms offers new surprises, such as the catalytic nature of RNA ("ribozymes"), the pathogenic nature of proteins ("prions"), or the silencing of genes via RNA interference. The synthesis of biopolymers and biologically active molecules remains an important task in chemical biology, as it 1) gives access to natural compounds, which otherwise could be isolated only in small and insufficient amounts; 2) allows the synthesis of modified or non-natural analogs or intermediates of natural compounds; and 3) provides tool compounds for the investigation of biologic mechanisms.

Technical Background

Although Emil Fischer had already pioneered the synthesis of sugars, nucleobases, and oligopeptides at the beginning of the twentieth century, the synthesis of these biomolecules in solution involved multistep sequences with many tedious workup and purification operations. Bruce Merrifield recognized that their syntheses can be simplified if a substrate is immobilized on an insoluble polymeric support ("resin") via a linker unit (1). Once immobilized the substrate can be easily modified by adding an excess of reagents and building blocks to ensure complete conversion. Consumed and excess reagents are then easily removed by washing the support and by simple filtration. Finally, cleavage of the linker releases the desired product into the solution allowing its isolation in pure form (Fig. 1).

Solid Support

A wide range of support materials exists for the heterogeneous immobilization of substrates in solid-phase synthesis, each offering certain advantages in different applications (2, 3). Parameters to be considered are 1) loading (mmol of functional groups per gram support), 2) bead size (measured in "mesh"), 3) swellability in solvents (degree of cross-linking and hydrophilicity), and 4) mechanical stability (mechanical abrasion interferes with filtration workup). A more than 99.99% of the substrate molecules are buried within the gel matrix of a polymeric bead, the resin must be swollen by solvent molecules to allow
Solid-Phase Synthesis of Biomolecules

Figure 1. General flow scheme of solid-phase synthesis. The solid support is represented by the gray ball. In this cartoon, only one functional group is depicted instead of billions on a real bead.

access to reagents. Polystyrene (PS) resins (cross-linked with 1-2% divinylbenzene (DVB), loading up to 2 mmol functional group/g) are the resins most commonly used for oligopeptide and small-molecule solid-phase synthesis. These resins swell in most organic solvents but not in very polar solvents such as alcohols or water. If reactions need to be performed under polar conditions, the use of Tentagel (PS-resin with grafted polyethylene glycol chains), PEGA, Pepsyn, or Argogel is recommended because these substances are distinguished by polar, flexible, polymeric chains. Highly cross-linked macroporous resins (e.g., Argopore) exhibit a solvent-independent permanent pore structure and can be used in any solvent. For oligonucleotide synthesis, controlled pore glass (CPG), which is an inorganic support with well-defined, large pores, is widely used. The low loading and high cost of this material are compensated by the operational advantage of good accessibility in any solvent. Recently, the synthesis of biopolymers on planar surfaces, such as glass (DNA-chips) (4) or cellulose membranes (SPOT-synthesis), (5) has gained increasing attention.

Linker

A linker can be considered as a polymeric (or, more general, immobilized) protecting group. It is a functional group responsible for the covalent attachment of the substrate onto the support that will be cleaved from the product at the end of the synthesis sequence. The following features have to be considered when choosing an appropriate linker: 1) The attachment of substrate should proceed in high yield, 2) the linker should be stable under all reaction conditions considered for the synthetic sequence ("orthogonality"), and 3) the linker should be cleaved under mild conditions in quantitative yield without obscuring the integrity and purity of the product. More than 200 different linkers have been described exhibiting different degrees of orthogonality and releasing different functionalities at the product after cleavage (6, 7). Cleavage of the linker can be initiated by the addition of acid (commonly used in oligopeptide synthesis), base (used in DNA-synthesis), metals, oxidation, reduction, or light (photocleavable linker). Linkers that release their product without any visible functional group are named "traceless linkers." "Safety-catch linkers" require two different chemical reactions in order to undergo cleavage, which offers a unique degree of stability. In Fig. 2 some frequently used linkers and the corresponding cleavage conditions are depicted.

Automation

As solid-phase synthesis basically involves only three types of operations (addition of excess of reagent; shaking the beads in the reagent cocktail; filtration and washing) machines have been designed to perform solid-phase synthesis automatically (8). Automated oligonucleotide and oligopeptide synthesis has become routine.

Synthesis of Biopolymers

The solid-phase synthesis of biopolymers via iterative coupling of monomeric building blocks has found widespread application. It can be adapted easily for the incorporation of non-natural building blocks or labeled monomers, which will be used as probe molecules in chemical biology.
Solid-Phase Synthesis of Biomolecules

Oligopeptides are synthesized on solid phase from the C- to the N-terminus. Depending on the N-terminal protecting group of the amino acid building blocks, two different strategies can be followed. In the "Boc-strategy," the tert-butyloxycarbonyl-protecting group is cleaved by addition under acidic conditions (trifluoroacetic acid), which requires a linker that is resistant to these conditions because it is either cleaved only by very strong acids (HF, trifluoromethanesulfonic acid) or by a different agent (Fig. 3) (1). The "Fmoc-strategy" is based on the fluorenylmethoxycarbonyl-protecting group, which is cleaved by addition of 20% piperidine/DMF. It has the advantage that the important class of acid labile linkers (e.g., Wang-linker or Rink-linker) is compatible with this approach.

In addition, measurement of the UV-absorption of the cleaved protecting group allows the determination of the yield of the previous coupling operation. In practice, oligopeptides with up to 50 amino acids can be synthesized on solid phase. Instrumental for achieving reasonable yields for long oligomers is an almost perfect conversion for each deprotection and coupling step, for which a plethora of modern peptide coupling reagents are available; the coupling reagents also minimize epimerization (9). The synthesis of larger proteins (>60 aa) may be accomplished by preparation of fragments on solid phase that are then assembled using solution-phase ligation reactions (10), such as the native chemical ligation using an N-terminal Cys-residue (11). Table 1 lists representative examples of de novo synthesized proteins using solid-phase peptide synthesis. Recently, the combination of expressed protein ligation and solid-phase peptide synthesis...

Figure 2 Representative linkers frequently used in solid-phase synthesis. The molecular moiety depicted in blue represents the molecule that is released upon cleavage of the linker. DCM = dichloromethane; TFA = trifluoroacetic acid.
Solid-Phase Synthesis of Biomolecules

**Figure 3** Flow scheme of a solid-phase synthesis of an oligopeptide following the Boc-strategy. The first amino acid is attached to the Merrifield linker by alkylative esterification of its cesium-salt. After acidic deprotection of the Boc-group and neutralization, the N-terminus undergoes DCC-mediated coupling with another Boc-protected amino acid building block. This sequence can be iterated, leading to even longer oligopeptides. Finally, the protect is released upon cleavage with HF. DCC = dicyclohexylcarbodiimide.

**Table 1** A selection of notable examples for the total solid-phase synthesis of proteins

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Protein</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merrifield et al.</td>
<td>1969</td>
<td>Ribonuclease A</td>
<td>124 aa</td>
</tr>
<tr>
<td>Sigler et al.</td>
<td>1976</td>
<td>A polipoprotein C1</td>
<td>57 aa</td>
</tr>
<tr>
<td>Fairwell et al.</td>
<td>1983</td>
<td>Human parathyroid hormone</td>
<td>84 aa</td>
</tr>
<tr>
<td>Blake</td>
<td>1986</td>
<td>5-carboxymethyl bovine apocytochrome c</td>
<td>104 aa</td>
</tr>
<tr>
<td>Wlodawer et al.</td>
<td>1989</td>
<td>HIV-Protease</td>
<td>202 aa</td>
</tr>
<tr>
<td>Briand et al.</td>
<td>1989</td>
<td>Ubiquitin</td>
<td>76 aa</td>
</tr>
<tr>
<td>Di Bello et al.</td>
<td>1992</td>
<td>Horse heart cytochrome C</td>
<td>104 aa</td>
</tr>
<tr>
<td>Lu et al.</td>
<td>1998</td>
<td>Bovine pancreatic trypsin inhibitor</td>
<td>58 aa</td>
</tr>
<tr>
<td>Kaiser et al.</td>
<td>1999</td>
<td>Macrophage migration inhibitory factor</td>
<td>115 aa</td>
</tr>
<tr>
<td>Canne et al.</td>
<td>1999</td>
<td>Human group V secretary phospholipase A2</td>
<td>118 aa</td>
</tr>
<tr>
<td>Hackeng et al.</td>
<td>2001</td>
<td>Human matrix G1a protein</td>
<td>84 aa</td>
</tr>
<tr>
<td>Miranda et al.</td>
<td>2001</td>
<td>Human S100A4.12</td>
<td>91 aa</td>
</tr>
<tr>
<td>Low et al.</td>
<td>2001</td>
<td>Cytochrome b562</td>
<td>156 aa</td>
</tr>
<tr>
<td>Becker et al.</td>
<td>2003</td>
<td>Ras + Raf</td>
<td>166 + 81 aa</td>
</tr>
<tr>
<td>Kochendoerfer et al.</td>
<td>2003</td>
<td>Erythropoietin</td>
<td>166 aa</td>
</tr>
<tr>
<td>Kochendoerfer et al.</td>
<td>2004</td>
<td>Vpu</td>
<td>81 aa</td>
</tr>
<tr>
<td>Cabrele et al.</td>
<td>2006</td>
<td>Id2</td>
<td>119 aa</td>
</tr>
</tbody>
</table>

aa = amino acid.
has emerged as a powerful tool for the preparation of proteins with unnatural modifications or labels (12). Solid-phase peptide synthesis has reached a level where the ton-scale synthesis of peptide drugs, such as the HIV drug Enfuvirtide (a 36mer peptide), has been established.

**Synthesis of Oligonucleotides**

The solid-phase synthesis of DNA oligomers has reached a high level of maturity and is extensively used for the synthesis of primers for polymerase chain reaction or the design of new genes (13). In contrast to the cellular process, the synthetic route forms the polymer in 3′-to-5′ direction. The oligomerization is carried out on CPG using a base labile linker. Typically, the nucleobases are introduced as acyl-protected phosphoramidites, which will be oxidized to phosphates after each elongation step (14). Treatment with base leads to global deprotection (nucleobase protecting groups and the cyanoethyl protecting group at phosphate) and cleavage of the oligonucleotide from the solid support. The conversion of each coupling step can be monitored by UV measurement of the dimethoxytrityl-carbocation, which is generated by the deprotection of the 5′-OH before each new coupling cycle (Fig. 4). The introduction of non-natural or modified nucleobases can be easily accomplished by using a corresponding phosphoramidite building block during the synthetic sequence.

The synthesis of RNA-oligomers (e.g., needed for applications in RNAi, aptamers, or as ribozymes) follows the steps described above for DNA but requires the protection of the
Solid-Phase Synthesis of Biomolecules

2'-O-TOM - Method: 2'-O-Trisopropylsilyloxymethyl (TOM)-protecting group

- Cleavage 5'-O-PG: 4% TCA in CH2Cl2
- Cleavage nucleobase-PG: 10 M CH3NH2 in EtOH/H2O
- Cleavage 2'-O-PG: 1 M Bu4NF·3H2O in THF
- Coupling efficiency: 96.4 % per coupling

2'-O-ACE - Method: 2'-O-Bis(2-acetoxyethoxy)methylorthoester (ACE) protecting group

- Cleavage 5'-O-PG: 1 M disodium-2-carbamoyl-2-cyanoethylen-1,1-dithiolate trihydrate in DMF
- Cleavage nucleobase-PG: 40% CH3NH2 in H2O
- Cleavage 2'-O-PG: 1.1 M HF/2.9 M Et3N in DMF
- Coupling efficiency: >99 % per coupling

Figure 5 Most frequently used 2'-OH protecting groups in RNA-oligomer synthesis. DMF = N,N-dimethylformamide, PG = protecting group.

2'-OH group, for which the 2'-O-TOM - and the 2'-O-ACE-protecting groups are most often used (Fig. 5) (15). PNA-oligomers are synthesized using methodology developed for peptide synthesis. The large-scale synthesis of oligonucleotides has been established for the preparation of sufficient quantities for clinical trials of antisense drugs.

Synthesis of Oligosaccharides

The automated solid-phase synthesis of oligosaccharides is expected to have a tremendous impact in glycobiology as many natural proteins are decorated by complex oligosaccharide conjugates that strongly influence the biologic activity of the conjugate. Recently, the first examples of automatically synthesized oligosaccharides have been reported (16). The availability of appropriately protected building blocks and the creation of certain linkage patterns remain as challenges in this field.

Solid-Phase Synthesis of Small Molecules

Protein function can be modulated by small-molecule ligands. Offering complementary advantages to mutation genetics, the search for small molecules as molecular probes for the investigation of biologic systems has gained new interest ("chemical genetics"). In addition, high-throughput screening efforts in hit finding and lead optimization in drug discovery require large collections of small molecules ("molecular libraries"). Solid-phase synthesis has strongly contributed to these efforts as it gives fast access to large compound libraries.

Combinatorial Synthesis

In the early 1990s combinatorial chemistry was developed as a tool for the production of large oligopeptide libraries (17–19). Following a split–mix strategy depicted in Fig. 6, libraries with millions of different compounds can be created. This approach was extended by Ellman and colleagues for the synthesis of small-molecule libraries (20). Diversity-oriented synthesis (DOS) has been invented as a powerful tool for the synthesis of very large structurally diverse compound libraries (21). For the identification of the active compound, several strategies, such as on-bead-screening, deconvolution, or encoding, have been established. The use of "IRORI-Kans" (penetrable little containers filled with beads and labeled either with a radio-frequency tag or an optical bar code on the container lid) allows the synthesis of mg-quantities of compounds in a split–mix format, which offers the additional advantage of having the synthetic history of each kan monitored by a computer-based readout system of the radio-frequency tag or the optical bar code (22).
Solid-Phase Synthesis of Biomolecules

Figure 6: Combinatorial synthesis using a split-mix format. Building blocks are added separately in different vials, but the beads with reaction products are pooled before being distributed again before the next reaction step. Each vial contains all combined products of the previous reaction steps, but each bead contains only a single compound.

Natural Product Libraries

Several of the most useful small-molecule probes are complex natural products isolated from marine, fungal, or plant sources. As natural products have built-in biologic properties efforts in the synthesis of compound libraries increasingly use natural product scaffolds as a starting material because they promise higher hit rates in biologic screens (23). In Fig. 7 a few representative examples of molecular probes identified by screening of combinatorial libraries are depicted.

References

Solid-Phase Synthesis of Biomolecules


Further Reading


See Also

Diversity-Oriented Synthesis of Small Molecules
Natural Products Inhibitors to Study Biological Function
Nucleic Acid Synthesis: Key Reactions of Proteins: Structure, Function and Stability
Solution-Phase Synthesis of Biomolecules
Catalytic Antibodies: Past, Present, and Future

Andrew P. Brogan, Tobin J. Dickerson and Kim D. Brogan, The Scripps Research Institute, La Jolla, California

Catalytic antibodies have emerged as powerful tools for the chemical biologist, enabling the design and realization of specific catalysts for a wide range of chemical reactions. Catalytic antibodies are grounded upon transition state theory and envisioned as programmable mimics of enzyme catalysis. Affinity maturation of the immune response for a small-molecule hapten elicits binding site complementarity within an antibody that facilitates chemical catalysis. The evolution of hapten design strategies for chemical catalysis is presented, including the transition state analog approach, strain-induced hapten design, "bait-and-switch," and "reactive immunization." The range and scope of antibody catalysis is examined by reaction class, highlighting structural and mechanistic investigations to explore the roots of chemical catalysis by these designer biocatalysts. Recently, antibodies, regardless of disposition or origin, have been shown to catalyze the oxidation of water, equipping the antibody with a mechanism for antigen decomposition. These recent developments are presented along with the utilization of this pathway in the oxidative degradation of a commonly abused drug. The achievements in antibody catalysis have enriched our scientific understanding of chemical catalysis, particularly by biological molecules in aqueous systems. However, realizing more operative rate enhancements on the same order as natural enzymes remains as the "holy grail" of this field.

Introduction

The usual paradigm in chemical biology is that chemical means lead to biological ends; however, in the case of catalytic antibodies, the reverse is the case—eliciting catalytic antibodies through biological means provides catalysts for a chemical end. The idea that antibodies could be designed to catalyze a specific chemical transformation was first proposed by Jencks (1) and was built on the foundations of enzyme catalysis originally conceptualized by Pauling (2). In this presumption, the catalytic power of an enzyme is primarily derived from the stabilization of the high energy transition state along a given reaction coordinate (Fig. 1). Therefore, a catalyst could be generated by probing the immune response for an antibody that can bind a small molecule, or hapten, that is a transition state analog of a desired chemical reaction. The first catalytic antibodies reported in 1986 independently by Lerner et al. (3, 4) and Schultz et al. (5) employed this strategy, and since these seminal reports, more than 50 chemical reactions have been catalyzed by antibodies. The technique has been refined over the years, but the general concept remains the same and numerous reviews on catalytic antibodies have appeared in the literature (6–15). Catalytic antibodies offer unique capabilities in a range of scenarios, including stereoselective organic synthesis, therapeutic potential in the treatment of disease, the elimination of toxins, the attenuation of agents used in chemical and biological warfare, and cessation of abused and/or addictive drugs. In this article, a broad overview of antibody catalysis is presented, including modern methods for eliciting catalytic antibodies, the evolution of hapten design, and advances in antibody catalysis.

Antibodies: Structure and Function

A brief introduction to the basic structure and function of an antibody is essential to understanding the catalytic power of a select few of these molecules. Immunoglobulins (Igs) are glycoproteins that can be divided into various classes and subclasses based on structure and function. Most catalytic antibodies are IgG molecules and therefore will be the focus of this review. IgG molecules consist of four polypeptide chains, two identical light chains and two identical heavy chains, that assemble...
Gene undergoing similar recombination of VL and JL genes. This recombination of three translated genes: a VH (variable) gene, a germline diversity (DH) gene, and a JH (joining) gene with the light-chain repertoire to recognize a seemingly limitless number of antigens within the genome, and the source and magnitude of the immune response after hapten hyperimmunization with no attempt to purely specific IgG molecules. Efficiency and cost-effectiveness are hallmarks of polyclonal antibody production; however, accurate characterization of this mixture is difficult. Furthermore, X-ray crystallography to examine structure-function relationships and affinity maturation to optimize the antibody is impossible with the polyclonal antibody method.

Consequently, hybridoma technology was a significant milestone enabling the isolation and production of individual antibodies (18). In this technique, antibody producing cells are isolated from the spleen after hyperimmunization with the hapten-protein conjugate. These B cells are subsequently fused with an immortal cell line, and the resultant hybrids secrete monoclonal antibodies. Monoclonal antibodies are homogenous, can be produced in large quantities, and can be rigorously purified to remove any potential contaminants. Despite greater expense and a more time-consuming process, monoclonal antibody production is the method of choice.

Phage-display technology has many uses for catalytic antibody research (19–23). Generally, phage display involves combinatorial antibody Fab or scFv (single-chain variable fragment) libraries and their expression on phage particles. One advantage to this technology is that the hyperimmunization protocol can be avoided, thereby removing the use of animals. Naïve Fab or scFv are identified in a screen for binding to a desired transition state analog. Although in theory a desired catalyst can be discovered this way, in practice this technique is more successful when an initial hyperimmunization protocol is performed followed by acquisition of the mRNA of the B cells from the spleen (19–21). The corresponding focused combinatorial antibody library is biased toward antibody fragments that recognize the hapten. Phage-display is also used as a tool in affinity maturation of a previously identified catalytic antibody. Mutagenesis of the corresponding Fab or scFv from an existing catalytic antibody can be explored using error-prone polymerase chain reaction (PCR), CDR walking, structure-guided mutagenesis, and DNA shuffling to optimize catalysis of the corresponding chemical reaction (22).

Eliciting Catalytic Antibodies

A merging of chemistry and biology is essential to effectively probe the immune system for catalytic antibodies (Fig. 3). Hapten design is critical in eliciting catalytic antibodies as variations of the central theme that transition state stabilization in the antibody combining site will yield functional catalysts for a desired chemical reaction. The evolution of hapten design will be discussed further in subsequent sections. Once the hapten is selected and synthesized, it is attached to an immunogenic carrier protein, usually via an amide bond, for hyperimmunization. A preliminary screen for antibodies that bind the hapten using an enzyme-linked immunosorbent assay (ELISA) is followed by another screen for catalysis of the reaction for which the hapten was designed. Other screening methods have also been used, including catELISA, which screens for catalysis in the antibody pool rather than hapten binding (16, 17). There are three primary methods for eliciting catalytic antibodies currently used: polyclonal, hybridomas, and phage-display.

Polyclonal antibody production is the most primitive method and has several significant limitations (6). This method inherently yields a complex mixture of antibody molecules from the immune response after hapten hyperimmunization with no attempt to purely specific IgG molecules. Efficiency and cost-effectiveness are hallmarks of polyclonal antibody production; however, accurate characterization of this mixture is difficult. Furthermore, X-ray crystallography to examine structure-function relationships and affinity maturation to optimize the antibody is impossible with the polyclonal antibody method.

Consequently, hybridoma technology was a significant milestone enabling the isolation and production of individual antibodies (18). In this technique, antibody producing cells are isolated from the spleen after hyperimmunization with the hapten-protein conjugate. These B cells are subsequently fused with an immortal cell line, and the resultant hybrids secrete monoclonal antibodies. Monoclonal antibodies are homogenous, can be produced in large quantities, and can be rigorously purified to remove any potential contaminants. Despite greater expense and a more time-consuming process, monoclonal antibody production is the method of choice.

Phage-display technology has many uses for catalytic antibody research (19–23). Generally, phage display involves combinatorial antibody Fab or scFv (single-chain variable fragment) libraries and their expression on phage particles. One advantage to this technology is that the hyperimmunization protocol can be avoided, thereby removing the use of animals. Naïve Fab or scFv are identified in a screen for binding to a desired transition state analog. Although in theory a desired catalyst can be discovered this way, in practice this technique is more successful when an initial hyperimmunization protocol is performed followed by acquisition of the mRNA of the B cells from the spleen (19–21). The corresponding focused combinatorial antibody library is biased toward antibody fragments that recognize the hapten. Phage-display is also used as a tool in affinity maturation of a previously identified catalytic antibody. Mutagenesis of the corresponding Fab or scFv from an existing catalytic antibody can be explored using error-prone polymerase chain reaction (PCR), CDR walking, structure-guided mutagenesis, and DNA shuffling to optimize catalysis of the corresponding chemical reaction (22).

Evolution of Hapten Design

The conformational changes and charge distribution along the reaction coordinate of a chemical transformation are fundamental to hapten design. Catalytic antibodies are designed to mimic the catalytic power of an enzyme, which, in part, stems from the stabilization of the high energy transition state.

---

**Figure 1** Relative energies of substrate (S) and product (P) along a reaction coordinate, revealing the energy differences between uncatalyzed (TS uncatal) and catalyzed (TS cat) transition states.
Catalytic Antibodies: Past, Present, and Future

**Figure 2** Y-shaped structure of immunoglobulin G (IgG). Disulfide linkages and sites of oligosaccharide attachment are shown. CDRs (complementarity determining regions) are the sites of high amino acid variability. Antigen binding site comprises VH (variable heavy) and VL (variable light) domains, which together make the Fab. Fv is composed of VH, VL, CH1, and CL. H is the hinge region. Fc is composed of CH2 and CH3. There are two Fab and Fv per IgG and one Fc region per IgG.

**Figure 3** Flowchart illustrating the key stages in catalytic antibody generation.

**Stable Transition State Analog Hapten Design**

A stable chemical analog that mimics the transition state of a chemical reaction was the first approach used to elicit catalytic antibodies (Fig. 4) (3–5). Acyl transfer reactions are the most studied type of catalytic antibody reaction, and a wealth of knowledge about this reaction has been garnered through antibody acyl transferases (3–5, 23–28). Using ester hydrolysis as a representative example, nucleophilic addition of water to the carbonyl carbon results in a tetrahedral transition state followed by expulsion of the alcohol leaving group. The transition state for this reaction has a delocalized negative charge that is remarkably similar to the chemically stable phosphate ester (29). The phosphorous (V) core, known to be an excellent mimic of the transition state in hydrolytic enzymes (30–33) and...
amino acids HisH35 and TyrH133 serve to stabilize the polarized body that binds the substrate in a ring-strained conformation. Hydrogen bonds from the side chains of the adjacent enzyme, because alkylation at one pyrrole nitrogen distorts the last step in the heme biosynthetic pathway (40). Interestingly, N-alkylporphyrins are known to be potent inhibitors of this enzyme, because alkylation at one pyrrole nitrogen distorts the planarity of the porphyrin macrocycle (41). This finding was used in the design of hapten 1 to catalyze the incorporation of metal ions into mesoporphyrin IX (3) by eliciting an antibody that binds the substrate in a ring-strained conformation.

The lone-pair electrons on the pyrrole nitrogen of the porphyrin ring are more accessible to chelation of metal ions in the ring-strained conformation and leads to metalation of mesoporphyrin IX. Antibody 7G12 catalyzes the incorporation of Zn2+, Cu2+, Co2+, and Ni2+ into mesoporphyrin IX, whereas ferrochelatase uses Fe3+, Zn2+, Co2+, and Ni2+ as substrates in the chelation of protoporphyrin IX. X-ray crystallography of the catalytically active 57G12-Michaelis complex revealed that the porphyrin ring adopts a nonplanar conformation that is essential for catalysis as anticipated from the hapten design (42).

**Bait-and-Switch Hapten Design**

Transition state analog hapten design has been the most universal means to elicit catalytic antibodies; however, limitations with this approach and strain-induced hapten design are realized when fractional bond orders, extended bond lengths, expanded valences, distorted bond angles, and charge distributions cannot be achieved in a stable chemical analog of the transition state suitable for immunization (7, 43, 44). Furthermore, a significant number of haptens designed for a specific chemical transformation have led to antibodies that bind the hapten with exquisite affinity yet no effective catalysis was realized. This is easily explained because the somatic mutations in the IgG molecule favor tighter hapten binding, but not necessarily more effective catalysis. Additionally, the products of a reaction may have significant similarity to the hapten used in hyperimmunization, leading to slow release of the product and/or significant product inhibition. In fact, product inhibition is thought to be the major contributing factor for low efficiency hydrolytic catalytic antibodies (45, 46).

A significant step in the evolution of hapten design was introduced by Janda and Lerner, coined the “bait-and-switch” method (47–49). This novel advancement enables electrophilic/nucleophilic and/or general acid/general base catalysis to be programmed into an antibody combining site. Specifically, a point charge on the hapten in close proximity to, or in direct substitution for, a functional group to be transformed in the respective substrate is used to induce a complementary charge on an amino acid residue in the antibody combining site during hyperimmunization (Fig. 6). The substrate lacks this charge but retains a similar overall structure and the corresponding antibody binds the substrate and acts as a general acid/general base and/or as a nucleophile/electrophile in the desired chemical reaction.

The phosphodiesterase antibody MATT.F-1 is a didactic example of bait-and-switch hapten design, illustrating differences from other hapten design approaches (49). The hydrolysis of a phosphodiester bond, such as those found in RNA and DNA, are catalyzed by ribonucleases (RNases) and deoxyribonuclease-ases (DNases), respectively. RNase A is a thoroughly studied enzyme that has two catalytic histidine residues in the active site (50). The imidazole group of HisH2 acts as a general base by protonating the 2’ oxygen, and the imidazolium group of HisH13 acts as a general acid by protonating the 5’ phosphoryl oxygen in the classic mechanism (51).

The incorporation of a general base and a general acid in the hydrolysis of a phosphodiester was hypothesized to be

<table>
<thead>
<tr>
<th>Catalytic Antibodies: Past, Present, and Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Phosphonate ester as a chemically stable mimic of ester hydrolysis: (b) Transition state analog hapten 1 elicited antibody 48G7 that catalyzes the hydrolysis of ester 2. (c) Key contacts of the 48G7 Fab-I complex.</td>
</tr>
<tr>
<td>(B) Transition state Analog Hapten 1</td>
</tr>
<tr>
<td>(C) Phosphoryl bond of hapten 1 that would assist in forming the transition state of ester 2. Main-chain amida bonds from Tyrl120 and TyrH100 also provide additional hydrogen-bond stabilization forces.</td>
</tr>
</tbody>
</table>

Figure 4 (a) Phosphonate ester as a chemically stable mimic of ester hydrolysis: (b) Transition state analog hapten 1 elicited antibody 48G7 that catalyzes the hydrolysis of ester 2. (c) Key contacts of the 48G7 Fab-I complex.
elicited in an antibody combining site programmed by specific point charges designed into a bait-and-switch hapten (49). Indeed, an antibody against hapten \(6\) successfully catalyzed the hydrolysis of the corresponding substrate \(7\) (Fig. 6). In contrast to the transition state analog hapten \(8\) (52), which elicited the less-proficient phosphodiesterase catalytic antibody 2G12, the precise conformation of the high energy intermediate was sacrificed for charged moieties in specific locations of the hapten. The corresponding counterion charges from amino residues were elicited in the antibody combining site that led to catalytic antibodies with improved catalytic proficiency \((k_{cat}/K_m)/k_{uncat} = 1.6 \times 10^7 \text{M}^{-1} \text{s}^{-1}\) for MATT.F-1 versus \((k_{cat}/K_m)/k_{uncat} = 1.3 \times 10^6 \text{M}^{-1} \text{s}^{-1}\) for 2G12) for phosphodiesterase activity.

**Reactive Immunization Hapten Design**

An essential concept in enzyme catalysis, partly realized through the study of catalytic antibodies, is that the proficiency of an enzyme is not solely due to the stabilization of a high energy transition state. Enzymes are not static entities, but rather they have the dynamic ability to stabilize all possible conformations of a chemical reaction along the reaction coordinate (33, 53–55). Furthermore, many enzymes can form covalent intermediates with the substrate that are essential to catalysis. The previous hapten design methods program specific complementarity in the antibody combining site; however, the hapten is a static snapshot of a dynamic chemical process resulting in catalysts limited to noncovalent interactions that are ultimately less efficient than their enzyme counterparts. Reactive immunization is a hapten design strategy that provides a chance for catalytic antibodies to approach the catalytic efficiency of natural enzymes by using a hapten that undergoes dynamic conformational changes during hyperimmunization and traps chemical reactivity at the B-cell level (Fig. 7) (53, 56–63).

A direct comparison between reactive immunization and transition state analog hapten design was demonstrated by comparing antibody esterase activity elicited against a reactive immunization hapten, phosphonate diester \(9\), and a transition state analog hapten, phosphonate monester \(10\) (Fig. 7) (58, 61). Hapten \(9\) was originally designed for the purpose of resolving a racemic mixture of naproxen esters and contains a modestly reactive diphosphonate ester that is susceptible to nucleophilic attack during hyperimmunization. Antibody 15G2, elicited against hapten \(9\), efficiently catalyzed the hydrolysis of \(11a\) to \(S(+)-naproxen\ 12a\) and phenol \(13\). Another antibody, 5A9, from the reactive immunization panel possessed turnover numbers lower than 15G2, however, kinetically resolved the hydrolysis of racemic substrate \(11\) to the anti-inflammatory agent \(S(+)-naproxen\ 12a\) in 35% yield and 90% ee. Meanwhile, antibody 6G6, raised against hapten \(10\), catalyzed hydrolysis of \(11a\) with comparable turnover numbers to 15G2, but kinetically resolved the hydrolysis of racemic substrate \(11\) to the anti-inflammatory agent \(S(+)-naproxen\ 12a\) in 50% yield and >98% ee.
This direct comparison of hapten design approaches for the same reaction revealed that antibodies generated by each method exhibit quite different catalytic behavior. The transition state analog approach provided catalytic antibodies with good turnover numbers and enantiomeric discrimination; however, it suffered from varying degrees of product inhibition byphenol 13. Comparatively, the reactive immunization approach yielded antibodies that are ultimately better catalysts because, once an efficient catalytic mechanism evolved further complementarity did not develop, leading to broader substrate recognition with reduced product inhibition. Binding site complementarity is the selection criteria rather than chemical reactivity using the transition state analog approach and antibodies developed by this strategy are more substrate specific and yet suffer from product inhibition as a result of exquisite binding affinity.

![Figure 6](image1.png)  
(a) The catalytic mechanism of RNase A, including the postulated transition state. (b) Bait-and-switch hapten 6 elicited antibody MATT-F1 that catalyzes phosphodiester bond hydrolysis of substrate 7. Transition state analog hapten 8 also elicited catalytic antibodies but with slower rates.

![Figure 7](image2.png)  
Figure 7  Reactive immunization hapten 9 elicited antibody 15G2 that catalyzes the hydrolysis of substrate 11. Transition state analog hapten 10 elicited antibody 12C8 to catalyze the same reaction.
Advances in Antibody Catalysis

More than 50 reactions have been catalyzed by antibodies (Table 1). A brief survey of the catalytic antibody landscape is presented below, highlighting creative hapten design strategies, the breadth of reactions catalyzed, and structure-function relationships.

Antibody Cationic Cyclases

Catalytic antibody technology enables the diversification of cyclization products from polyene substrates, because the programmability of antibody catalysis is not limited to the set of naturally occurring polyenes (64–69). Accordingly, hapten 14 elicited the cationic cyclase HA5-19A4 that catalyzes the tandem cationic cyclization of polyene substrate 15 to the bridge-methylated trans-decalins 16a–c (Fig. 8) (67). Cationic cyclization reactions have three components: initiation, propagation, and termination. The zwitterionic N-oxide moiety of hapten 14 mimics the initiation step of this reaction because it is isosteric and isopolar to the first carbocation formed at the beginning of the reaction cascade. The X-ray crystal structure of the Fab fragment of HA5-19A4 complexed with hapten 14 reveals a highly complementary fit in the antibody combining site (70). The hydrophobic pocket is lined with numerous aromatic residues that stabilize the postulated reactive intermediate 17 through cation–π interactions and force it into a chair–chair conformation. Propagation of cyclization proceeds via concerted attack of the C5–C6 π-bond, avoiding accumulation of the unfavorable carbocation at C1, while forming the A ring of 16a–c through the more favorable tertiary carbocation at C5. Subsequent attack on C5 by the C9-10 π-bond forms the B ring of 16a–c. Minor by-products of the reaction include compounds with incomplete closure of the B ring, resulting from either elimination or solvolysis. Termination of the cationic cyclization was programmed into the antibody by including an epoxide group to elicit an antibody residue capable of coordinating a water molecule to facilitate quenching of the terminal carbocation by solvent or to constrain the B ring of the trans-decalin into the half-chair and facilitate proton elimination. Interestingly, termination of the HA5-19A4 catalyzed reaction occurs exclusively by proton elimination because only olefinic products 16a–c are observed.

<table>
<thead>
<tr>
<th>Table 1 Examples of chemical reactions catalyzed by antibodies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ester hydrolysis</td>
</tr>
<tr>
<td>Amide hydrolysis</td>
</tr>
<tr>
<td>Phosphodiester hydrolysis</td>
</tr>
<tr>
<td>Carbamate hydrolysis</td>
</tr>
<tr>
<td>Decarboxylation</td>
</tr>
<tr>
<td>Norrish reaction</td>
</tr>
<tr>
<td>Robinson annulation</td>
</tr>
<tr>
<td>A-oxidation</td>
</tr>
</tbody>
</table>

For further study of the antibody catalyzed reactions see ref. 6–15.

Figure 8: (a) Hapten 14 elicited antibody HA5-19A4 that catalyzes the cationic cyclization of substrate 15. (b) Key contacts of HA5-19A4 Fab with postulated reactive intermediate 17.
Antibody-Catalyzed Disfavored Ring Closure

An astounding aspect of antibody catalysis is the ability of these programmable biocatalysts to preferentially form the less thermodynamically favored product (66, 71–74). The intramolecular cyclization reaction of trans-epoxyalcohol 18 is an archetypal example of antibody catalysis of a disfavored transformation, which preferentially forms the tetrahydrofuran 19 under uncatalyzed conditions due to the overwhelming stereoelectronic constraints predicted by Baldwin’s rules for ring closure (Fig. 9) (75, 76).

The N-oxide and N-methyl ammonium haptens 20 and 21 were designed to mimic the stereoelectronic features of the disfavored endo transition state 22 and function as bait-and-switch haptens by programming specific complementary charges in the antibody combining site (72, 77). Antibody 26D9, elicited against only product. Additionally, hyperimmunization using the N-methyl ammonium 20, efficiently reroutes this transformation and gives tetrahydropyran 23 as the hapten 21 produced antibody 5C8 that also catalyzed the regio- and enantioselective epoxide opening of substrate 18 to yield the disfavored endo product 23 (77). The active sites in both antibodies contain a putative catalytic diad, as determined by X-ray crystallography, confirming bait-and-switch hapten design as a viable approach to catalyze this disfavored ring closure (77). The exact mechanism has not been established; however, the active site of Fab 5C8 reveals plausible general acid–base catalysis occurring by AspH95 acting as a proton donor to the epoxide oxygen assisting the formation of intermediate 22 with HisL89 serving as a base enabling nucleophilic attack by the alcohol group.

Antibody Diels–Alderases

One aspect of antibody catalysis that truly ignites the imagination of the chemical biologist is that these biocatalysts are not limited to reactions that have a natural enzymatic equivalent. The Diels–Alder reaction has immense synthetic utility; however, this chemical transformation is extremely rare in nature. Furthermore, the reaction proceeds via an entropically disfavored, highly organized pericyclic transition state (78). The programmability of a catalytic antibody has enabled the catalysis of the Diels–Alder reaction previously considered beyond the realm of possibility with a protein (79–84).

A significant hurdle in the development of a Diels–Alderase catalytic antibody was minimizing product inhibition, because based on the Curtin–Hammett principle, the transition state is markedly similar to the product of a Diels–Alder reaction. A creative solution to this problem was employed in the development of antibody 1E9 (Fig. 10) (79). The endo-hexachloronorbornene
haptens 24 elicited antibody 1E9 to catalyze the Diels–Alder reaction between diene tetrachlorothiophene dioxide 25 and N-ethylmaleimide 26. Product inhibition was overcome because, after the pericyclic reaction forming 27, SO₂ is liberated spontaneously followed by oxidation to yield the structurally dissimilar aromatic product 28; accordingly, no product inhibition was observed. X-ray crystallographic data of 1E9 Fab revealed that the antibody binding pocket is preorganized to provide significant shape complementarity with hapten 24 through van der Waals contacts, π-stacking with the maleimide functional group, and a hydrogen bond with AsnH₃⁵ (85). A recent study of noncovalent catalyzed Diels–Alder reactions by synthetic, protein, and nucleic acid hosts indicated that antibody 1E9 is the most effective of the noncovalent catalyst systems studied (86). The capabilities of this extraordinary catalytic antibody have been explained by theoretical calculations and the high degree of shape complementarity consistent with the X-ray crystallography data.

Diels–Alderase antibody 39-A11 minimizes product inhibition while generating a more conventional Diels–Alder product (Fig. 11) (80). Bicyclo[2.2.2]octane hapten 29 was designed to mimic the proposed boat-like transition state 30 of the [4π+2π] cycloaddition between diene 31 and dienophile 32. Product inhibition was circumvented by the structural disparity between the product cycloadduct and the pseudo-boat form of the hapten employed for immunization. X-ray crystallographic data of hapten 29 complexed with the 39-A11 Fab indicated that the diene and the dienophile are bound in a reactive conformation that reduces
translational and rotational degrees of freedom (83, 87, 88). The stereoselective capabilities of this antibody are accomplished by two strategically positioned hydrogen bonds (AsnH35a, ThrH50) and π-stacking of the maleimide dienophile with TrpH50, as calculated from quantum mechanical models and docking simulations (89). This unique arrangement allows reorganization of one enantiomeric transition state, facilitating formation the chiral product 33. Interestingly, sequencing and cross-reactivity studies indicate that antibodies 39-A11 and 1E9 are structurally similar and may have the same polyspecific germline origin.

A radical idea to generate Diels–Alderase catalytic antibodies employed a ferroceny moiety in hapten 34 and was designed to catalyze the reaction of diene 35 with dienophile 36 (Fig. 12) (82). Hapten 34 has two pentagonal, delocalized, π-electron ring systems stacked upon each other that were thought to be a loose transition state mimic capable of guiding the diene and dienophile into a reactive ternary complex. Additionally, the hydrophobicity of the hapten will induce a strong immune response and generate antibodies containing hydrophobic microenvironments that sequester the reactants from aqueous solution, increasing reaction rates. The freely rotating ferrocenyl moiety may enable stereoselective catalysis of all possible diastereomers; however, the immune system must be able to bind and stabilize a single conformer of 34 to elicit an effective catalyst. Antibody 13G5 was identified to preferentially catalyze the formation of the disfavored ortho-π-π-stacked product 38. This is finding remarkable considering the flexibility of hapten 34, because antibody 13G5 must preferentially stabilize the π-π-transition state 39 over the **endo**-transition state 40. Quantum mechanical modeling depicts that hapten 34 resembles the van der Waals complex between the reactants more closely than the transition state, yielding antibodies that preferentially recognize the hapten rotamer that mimics 39 (90). Furthermore, the steric restraints imposed by specific hydrogen-bonding interactions revealed in the crystal structure of 13G5 Fab complexed with the inhibitor 41 (an attenuated version of hapten 34) traps 41 in one available conformation that is a loose mimic of the early boat-like transition state for the **endo** Diels–Alder reaction (91).

**Antibody-Catalyzed Oxy-Cope Rearrangements**

The oxy-Cope rearrangement is a thoroughly studied and synthetically useful reaction in organic chemistry that proceeds through a highly organized chair-like pericyclic transition state (Fig. 13) (92, 93). The transition state analog hapten 42 elicited antibody AZ-28, which catalyzed the oxy-Cope rearrangement of substrate 43 to aldehyde 44 (94). Product inhibition was avoided by in situ chemical modification to generate the oxime. Surprisingly, the germline precursor to AZ-28 accelerated this oxy-Cope reaction 164,000-fold faster than the uncatalyzed reaction despite a much lower affinity for hapten 42. An explanation for this unprecedented discovery can be explained by X-ray crystallography of both the apo form and hapten 42-complex (95). The van der Waals and hydrogen bond interactions present in AZ-28 force the hapten into a fixed conformation that is catalytically unfavorable. In contrast, the antibody combining site for the transition state analog does not necessarily result in more efficient catalysis. Antibody-Aldolases

The aldol reaction is a fundamental C–C bond forming reaction that is ubiquitous in both chemical synthesis and nature.
Catalytic Antibodies: Past, Present, and Future

Figure 13
Hapten 42 elicited antibody AZ-28 that catalyzes the oxy-Cope rearrangement of substrate 43. (97). Class I aldolase enzymes possess a reactive lysine residue that forms an enamine intermediate with carbonyl substrates, enabling nucleophilic attack on the corresponding electrophile in the enzyme active site (98, 99). Aldolase catalytic antibodies with a similar reactive lysine residue (59, 60, 100–102) were generated using the reactive immunization hapten 45, which contains a moderately reactive β-1,3-diketone moiety (Fig. 14). Two highly efficient aldolase antibodies, 38C2 and 33F12, were obtained from the catalytic screen (59, 60). The β-1,3-diketone functionality successfully trapped a lysine residue in the antibody combining site forming Schiff base 46 and ultimately the reactive enamine 47, which directly participates in the mechanism of the aldol reaction in the antibody active site. Both 38C2 and 33F12 catalyze the aldol reaction between acetone and aldehyde 48 with enzymic catalytic proficiency $k_{cat}/k_{uncat}$ of nearly $10^9$. The X-ray crystal structure of 33F12 revealed that LysH93 is essential to the catalytic mechanism, which initiates the reaction by forming a stable enamine with the ketone substrate (60). The surrounding hydrophobic residues help to stabilize the unprotonated form of the lysine ε-amino group ($pK_a$ of 10 in bulk water). A significant perturbation of LysH93 must occur to maintain its uncharged status. Antibodies 38C2 and 33F12 are actually better catalysts of the retro-aldol reaction; however, these catalysts are extremely robust and 38C2 participated in major steps in the total synthesis of epothilones A–F (103, 104).

An attempt to improve on these antibodies employed a hybrid approach to hapten design by using a transition state mimic sulfone along with a β-1,3-diketone moiety to trap a reactive site lysine residue (Fig. 15) (62). The hybrid hapten 49 is an excellent mimic of the aldol transition state 50, eliciting two aldolase antibodies, 93F3 and 84G3. In the aldol reaction of 51 with 3-pentanone, antibody 93F3 provided syn-aldol 52 in 90% de and 90% ee, whereas antibody 38C2 only afforded the anti-isomer in 62% de and 59% ee. These second-generation aldolase antibodies 93F3 and 84G3 showed $10^3$-fold increase in proficiency over antibody 38C2.

Figure 14
(a) Reactive immunization hapten 45 elicited antibodies 38C2 and 33F12 that catalyze the aldol reaction of aldehyde 48 with acetone. (b) Hydrophobic environment surrounding LysH93 in 33F12 Fab.
Catalytic Antibodies: Past, Present, and Future

Figure 15  Hapten 49, a hybrid of transition state analog and reactive immunization, elicited antibody 93F3 that catalyzes the aldol reaction of ketone 51 with 3-pentanone.

Antibody Catalyzed Photo-Fries Rearrangement

The photo-Fries rearrangement involves the absorption of light energy by a diphenyl ether substrate resulting in homolytic C–O bond cleavage followed by radical recombination to yield biphenyl products. Multiple products can be formed in this reaction, depending on the electron withdrawing/donating characteristics of the aromatic substituents. Furthermore, the escape of free radicals from the solvent cage leads to additional phenolic products (105). For example, ultraviolet (UV) irradiation of 4-phenoxyaniline 53 forms aromatic products 54–58 (Fig. 16). The electron donating character of the amine substituent leads to preferential cleavage via path a (products 54 and 55) as opposed to path b (products 56–58) that would be favored by an electron withdrawing substituent. An antibody-mediated reaction that suppresses the escape of free radicals and primarily forms biphenyl 54 was generated against haptens 59 (antibody MT2-21 C4) and 60 (antibody MT4-3G2) (106). The rotational freedom of hapten 59 was used to explore the combinatorial power of the immune response to elicit catalysts that stabilize radicals in the combining site and seek either path a or path b. Hapten 60 is rigid and designed to assist the immune repertoire in the selection of a catalyst by limiting the conformations that can be accessed. Antibodies elicited against both haptens catalyzed the photo-Fries reaction revealing the dynamic ability of an antibody to stabilize a high energy surface in the catalysis of a reaction. Additional photochemical catalytic antibodies for the Norrish type II reaction have been developed (107–109), including an enantioselective Yang cyclization (108, 109).

Recent Developments in Catalytic Antibodies

Since the seminal discovery of chemical catalysis by an antibody, numerous complex chemical transformations have been catalyzed by these molecules. However, in biological systems, the primary function of an antibody is to serve as a mediator between recognition of a foreign substance and its destruction. Specifically, the variable region of an antibody has evolved to recognize an antigen and then recruit effector systems such as complement and phagocytic cells to destroy the foreign entity. This paradigm has been challenged by the recent discovery that antibodies have the innate ability to not only recognize foreign substances but to also destroy them (110). Lerner and coworkers have found that all antibodies, regardless of source or antigenic specificity can catalyze the oxidation of water by singlet oxygen (1O2) via a pathway that is postulated to include trioxigen species, such as dihydrogen trioxide (H2O3) and possibly ozone (O3), in the formation of hydrogen peroxide (H2O2) as the ultimate product (110–114). Further examination of this phenomenon indicated that 1O2 could be generated by
Catalytic Antibodies: Past, Present, and Future

either direct UV irradiation of the antibody molecule, by visible light and a triplet oxygen \( (\text{O}_2^\cdot) \) sensitizer like hematoporphyrin IX, or by thermal decomposition of endoperoxides. In each case, the antibody catalyzed formation of \( \text{H}_2\text{O}_2 \) was triggered.

Typically, proteins are not stable under extended exposure to UV irradiation; however, antibodies efficiently form \( \text{H}_2\text{O}_2 \) linearly up to 40 mole equivalents before an observed decrease in rate. Furthermore, \( \text{H}_2\text{O}_2 \) seems to inhibit its own production and >500 equivalents of \( \text{H}_2\text{O}_2 \) can be generated by an antibody when \( \text{H}_2\text{O}_2 \) is removed by catalase. Isotopic labeling experiments suggest that water is the electron source in this oxidation pathway. Functionally, the water oxidation pathway has been postulated to play a role in a range of clinical scenarios, including bacterial killing, inflammation, and the pathogenesis of atherosclerosis (112, 115).

The manipulation of the water oxidation pathway expands the realm of possibilities with catalytic antibodies. A specific example recently reported is the catalytic oxidative degradation of nicotine (53) by antibodies TD1-10E8 and TD1-36H10 (Fig. 17) (116). Catalytic antibodies that degrade cocaine by ester hydrolysis have been previously identified, and the use of catalytic antibodies in the cessation of drug abuse is an active area of research (37, 117–121). Developing an antibody capable of degrading nicotine has been particularly challenging because its chemical structure is not amenable to decomposition by catalytic antibodies using any of the previously outlined hapten design strategies. Many oxidative degradation products of nicotine are known (122) and antibody catalysis could parallel this pathway with a sufficient \( \text{O}_2^\cdot \) source. Riboflavin interacts with immunoglobulins, although this interaction is not completely understood (123) and is a known photosensitizer. The riboflavin-antibody interaction was exploited in the generation of reactive oxygen species to oxidatively decompose nicotine. However, the initial screen using a tight binding (\( K_d < 10\mu M \)) nicotine antibody panel elicited against hapten 54, riboflavin, and either UV or visible light led to no catalysis of nicotine degradation over the appropriate control reactions. Conversely, a weak binding panel of antibodies elicited against the less congruent nicotine hapten 55 (\( K_a > 1 \text{mM} \)) effectively catalyzed the formation of nicotine oxidation products 56 and 57. It is important to note that, in this study, a weak binding nicotine antibody was converted to a catalytic antibody using visible light and riboflavin as a photosensitizer. The inherent utility of the water oxidation pathway is an active research area, and the antibody catalyzed oxidative degradation of nicotine is the first example of manipulating this pathway for a potential therapeutic outcome.

Conclusions

Catalytic antibodies are unparalleled as tailor-made enzyme mimics of a chemical reaction and a paramount advancement in chemical biology. The scope of antibody catalysis continues to increase although enzyme-like rates have yet to be achieved. A snapshot of the reaction coordinate using transition state analog, strain-induced, or bait-and-switch hapten design has led to numerous antibody catalyzed reactions. Enzymes and catalytic antibodies share their primary mode of action by stabilizing a high energy transition state. Enzymes have a superior evolutionary advantage, and additional mechanisms that assist enzyme catalysis, like covalent catalysis, cofactors, proximity effects, and the dynamic ability to complement an infinite number of conformations along a reaction coordinate, play a larger role than originally suspected.

Perhaps the most significant contribution of the catalytic antibody field is the realization that enzyme catalysis is not simply transition state stabilization. Reactive immunization has enabled a mimic of the dynamics involved in enzyme catalysis, and an aldolase antibody that approaches enzymatic rates has been developed using this technique (59, 60). However, only a few types of reactions have been catalyzed using this hapten.
design strategy. The ability to design a de novo enzyme-like catalyst for a multitude of different reactions is an intriguing lure for the chemical biologist. The advent of catalytic antibodies was essential to the current understanding of the relationship between chemical catalysis and molecular structure; however, a better understanding of this relationship is essential for the advancement of catalytic antibody research toward the ultimate goal of readily programmed catalysts with operative rates that are inherently friendly for the environment. The discovery of naturally occurring catalytic antibodies (124) and the realization that nature uses IgG molecules for chemical catalysis provides a basis to strive toward this ultimate goal.

Acknowledgment
This work was supported by The Skaggs Institute for Chemical Biology and the National Institute on Drug Abuse (DA 015700 to K.D.J. and DA 018507 to A.P.B. and DA 018507 to K.D.J.).

References
44. Morgan B, Scholtz JM, Ballinger MD, Zipkin ID, Bartlett PA.
43. Hanson JE, Kaplan AP, Bartlett PA. Phosphonate analogues of
42. Yin J, Andryski SE, Beuscher AE, Stevens RC, Schultz PG.
41. Dailey HA, Fleming JE. Bovine ferrochelatase. Kinetic analy-
39. Cochran AG, Schulte, PG. Antibody-catalyzed porphyrin meta-
33. Baldwin JE. Approach toward antibody catalysis. A stereochemical ap-
19. Hasserodt J, Janda KD, Lerner RA. Formation of bridge-methylated decalins by antibody-catalyzed tandem cationic cy-
17. Balch CM, Hasserodt J, Jones T, Lerner RA, Janda KD, Chu-

Catalytic Antibodies: Past, Present, and Future

Encyclopedia of Life Sciences © 2007 John Wiley & Sons, Ltd. www.els.net 15 - 2001 -
Catalytic Antibodies: Past, Present, and Future


Cross-References
- antigen-antibody interactions
- artificial enzymes, selective reactivities through asymmetric synthesis of biomolecules
- crystallization of proteins: overview of applications in chemical biology
- enzymatic synthesis of biomolecules
- enzyme catalysis, chemistry of
- enzyme catalysis, roles of structural dynamics in enzyme cofactors, chemistry of
- enzyme inhibition, mechanisms of enzyme inhibition, tools to study enzyme kinetics, techniques to study glycodies, major techniques in enzyme kinetics
- immunological structures
- noncovalent interactions in molecule recognition
- organic chemistry in biology
- oxidative metabolism, chemistry of
- phage display
- synthetic chemistry
- vitamins, chemistry of water, chemistry of
Enzymatic Synthesis of Carbohydrate-Containing Biomolecules

Hai Yu and Xi Chen, Department of Chemistry, University of California-Davis, California
doi: 10.1002/9780470153672.wecb153

As important biomolecules in living organisms, carbohydrates have received increasing attention in recent years. Their important roles in biologic events are being continuously unraveled. The development of synthetic methodologies, including both chemical and enzymatic methods, contributes greatly to the advance of the field of glycoscience. The involvement of regio- and stereo-selective enzymes in the synthesis of complex carbohydrate-containing molecules has become an indispensable approach. Many enzymes involved in the biosynthesis and biodegradation of carbohydrates have been characterized and have been applied for the production of carbohydrate-containing biomolecules, including oligosaccharides, polysaccharides, glycoproteins, glycolipids, and glycosylated natural products. A range of strategies for enzymatic synthesis have also been developed, such as protein engineering of glycosidases and glycosyltransferases by site-directed mutagenesis or directed evolution, one-pot multiple-enzyme synthesis, sugar nucleotide regeneration, solid-phase enzymatic synthesis, synthesis using immobilized enzymes, and cell-based synthesis. Enzymatic synthesis will continue to play critical roles in obtaining complex carbohydrate-containing biomolecules. Future efforts should focus on identifying synthetically useful enzymes such as those with flexible or novel substrate specificity and those that can form new bonds. This identification can be achieved by functional genomics and mutagenesis studies. Development of novel enzymatic synthetic methods is also critical to access diverse naturally occurring and non-natural derivatives of carbohydrates and glycoconjugates.

Biomolecules are chemical compounds that naturally occur in living organisms. They primarily contain carbon and hydrogen atoms, some with nitrogen, oxygen, phosphorus, and sulfur. A wide range of biomolecules exists. Major types are carbohydrates, nucleic acids, lipids, and peptides/proteins. Due to their important roles in bioprocesses and their potential use as drugs or drug targets, synthesis of biomolecules has been a subject of enormous interest. Most active biomolecules are chiral compounds. Although new synthetic reagents, catalysts, and strategies have been developed, chemical synthesis of complex biomolecules continues to pose substantial challenges. Therefore, environmentally benign, highly efficient, and highly selective enzymatic synthesis has been and will continue to play indispensable roles in obtaining many biomolecules. This is especially true for the synthesis of complex carbohydrates or glycoconjugates. The enzymatic synthesis of carbohydrate and carbohydrate-containing biomolecules will, therefore, be the subject of discussion here.

Enzymatic Synthesis of Oligosaccharides

Terminal carbohydrate moieties of naturally occurring glycoconjugates are usually the oligosaccharide synthetic targets. They are generally the key determinants recognized by carbohydrate-binding proteins and are directly involved in many important biologic and bioprocesses, including cell adhesion, differentiation, development, and regulation (1). Enzymatic synthesis of these molecules has been extensively studied.
Both glycosidases and glycosyltransferases have been widely applied.

**Glycosidase-catalyzed synthesis of oligosaccharides**

Glycosidases, including endoglycosidases and exoglycosidases, are enzymes that catalyze the cleavage of glycosidic bonds. Under certain conditions, glycosidases can be used for the formation of glycosidic links in which a carbohydrate hydrazyl acts as a more efficient nucleophile than as a water molecule.

Oligosaccharide formation catalyzed by wild-type glycosidases can be accomplished by either a thermodynamically (shifting the equilibrium by using a large excess of acceptors, adding organic solvent, or increasing reaction temperature) or a kinetically (using activated glycosyl donors and exogenous nucleophiles) controlled process (2, 3). Many wild-type glycosidases have been studied. Some are readily available and with low cost. They are relatively stable and have been used in the enzymatic synthesis of oligosaccharides. More recently, with a better understanding of the catalytic mechanism of glycosidases, mutants of glycosidases with high efficiency in catalyzing the formation of glycosidic bonds have been constructed. These mutants, which are mainly of retaining-type glycosidases, are called glycosynthases. They have been increasingly used in synthesis.

**Endoglycosidase-catalyzed synthesis of oligosaccharides**

Endoglycosidases have mainly catalyzed the hydrolysis of internal glycosidic bonds. Many endoglycosidases, however, have transglycosylation activity and have been used for the synthesis of oligosaccharides of varied lengths. For example, the β-glucosidase from *Glucosaminiurn galactaricium* has transglycosylation activity and has been used for the synthesis of a wide range of oligosaccharides, including LacNAc, Gal, and GlcNAc, with yields ranging from 7% to 44% (16). The enzyme showed a preference for the 1,3-α-D-glucosidases from glucosamine (13). Another Bacillus circulans β-galactosidase, however, was specific for the cleavage of the 1,3-β-link (12). It was used to synthesize Gals1.3GalNAc, Gals1.3Gal, and GlcNAc, and their pNP derivatives in 10–46% yields using GalNp as a donor (14). β-Galactosidases from *Penicillium multicolor*, *Aspergillus oryzae*, *Bifidobacterium bifidum*, and *Escherichia coli* have been widely studied and used to synthesize β-linked oligosaccharides, including LacNAc, Galβ1,4Gal. Kragl et al. used 25% of 1,3-di-methylimidazole methyl sulfate (a water-miscible ionic liquid) in water to decrease the hydrolysis of the formed product in a kinetic-controlled manner (13). Another Bacillus circulans β-galactosidase, however, was specific for the cleavage of the 1,3-β-link (12). It was used to synthesize Gals1.3GalNAc, Gals1.3Gal, and GlcNAc, and their pNP derivatives in 10–46% yields using GalNp as a donor (14). β-Galactosidases from *Penicillium multicolor*, *Aspergillus oryzae*, *Bifidobacterium bifidum*, and *Escherichia coli* have been widely studied and used to synthesize β-linked oligosaccharides, including LacNAc, Galβ1,4Gal. Kragl et al. used 25% of 1,3-di-methylimidazole methyl sulfate (a water-miscible ionic liquid) in water to decrease the hydrolysis of the formed product in a kinetic-controlled manner (13). Another Bacillus circulans β-galactosidase, however, was specific for the cleavage of the 1,3-β-link (12). It was used to synthesize Gals1.3GalNAc, Gals1.3Gal, and GlcNAc, and their pNP derivatives in 10–46% yields using GalNp as a donor (14).

**Exoglycosidase-catalyzed synthesis of oligosaccharides**

The native function of exoglycosidases is to cleave a terminal monosaccharide unit, but exoglycosidases can be used to add a monosaccharide in the oligosaccharide synthesis. Many types of exoglycosidases from bacterial, fungal, animal, and plant sources have been reported and have been used in the formation of glycosidic bonds, including sialidases, galactosidases, glucosidases, fucosidases, mannosidases, xylosidases, N-acetylhexosaminidases, and more.
Old World monkeys. Humans naturally produce anti-Gal antibodies that specifically recognize α-Gal epitopes (17). Although in low yields, α-galactosidases from A. oryzae, A. parasiticus, A. flavipes, A. terreus, Talaromyces, and coffee beans have been used to construct Galα1,3Gal (18–20). A highly specific α-galactosidase from Penicillium multicolor was able to catalyze the synthesis of α-Gal disaccharides and trisaccharides from Galα1,3GalP donor in 25–46% yields. The yield of the reaction can be increased by decreasing the concentrations of the substrates (21). Interestingly, a coffee bean and a Thermomyces lanuginosus α-galactosidase could catalyze the galactosylation of a cyclic glucosetetrasaccharide to form an α1,6-linkage (22).

Li et al. used a CLONEZYME thermophilic glycosidase library (Diversa Corporation, San Diego, CA) in the synthesis of Galα1,4Gal, Galα1,6Gal, Galα1,3Gal, and Galα1,3Galα. Some enzymes can use lactose as a donor and transfer two Gal residues in tandem to the hydroxyl group of the acceptor (23).

Glucosidases

Glucosidases have been commonly used as catalysts for the formation of alkyl glucosides, which are surface-active compounds that have applications in the pharmaceutical and the food industry. Alcohols are usually used as both solvent and acceptors. For example, glucoamylase and β-glucosidase have been used to catalyze the synthesis of α- and β-glucosides, respectively. Maximum yields were obtained when the reaction solution consisted 10% (vol/vol) of water in primary alcohol or 15% (vol/vol) of water in diol (24). Kosary et al. used an immobilized β-glucosidase in preparative scale synthesis of alkyl and aryl β-D-glucopyranosides in moderate yields ranging from 12% to 19%. They found that the alcohol solvent component of the reaction mixture can be replaced by 1,2-diacetoxyethane (25).

An α-glucosidase from Xanthomonas campestris was used to synthesize α-arabinosyl (hydroquinone-1-α-glucopyranoside), an important cosmetic ingredient, in 93% yield using maltose as donor and hydroquinone as acceptor (26).

Glucosidases have also been widely used as catalysts for the formation of various oligosaccharides. A thermostable β-glucosidase from Thermus thermophilus was used to transfer the glucose or fucose residue from Glc1PαNP or Fuc1PαNP to a glucose acceptor for the synthesis of Glc1,3(6OEt) and Fuc1,3(6OEt) in an 88% and a 58% yield, respectively (27). β-Glucosidases from Sclerotinia sclerotiorum and Aspergillus niger were both used to synthesize gluco-oligosaccharides, such as gentiose, cellotriose, and cellotetraose, from cellobiose (28). α-Glucosidases from a different source can catalyze the formation of gluco-oligosaccharides with different links. For example, an α-glucosidase from Aspergillus Niger favors the formation of an α1,6-linkage; a Penicillium purpurogenum enzyme yields mostly α1,4-linked maltotriose from maltose (29). An α-glucosidase isolated from fungal keratinases illacinius prefers to catalyze the synthesis of α1,2- and α1,3-linked gluco-oligosaccharides from maltose (29). A thermostable β-glucosidase from Bacillus, strain S-A 1606 shows very broad substrate specificity. Site-directed mutagenesis of this enzyme generates mutants with altered specificity for oligosaccharide formation (30).

Mannosidases

Mannosyl-terminated oligosaccharides are commonly found in the carbohydrate moieties of N-linked glycoproteins of the eukaryotic system. For example, Manα1,2Man and Manα1,3Man are terminal structures of high-mannose type and hybrid type N-glycans. Athanassopoulos et al. (31) reported that incubating mannosylase with a novel α-D-mannosidase from A. phomitis afforded Manα1,6Man and Manα1,8Man an in 21% and 5% yield, respectively.

β-Mannoside links, which are considered one of the most challenging synthetic glycosidic bonds, can be achieved by the catalysis of β-D-mannosidases from Aspergillus oryzae and Helix pomatia (32).

Jack bean α-mannosidase has also been able to catalyze the transglycosylation of rhamnose residue. It has been used in the synthesis of a disaccharide Rhaα1,2ManSEt in 32.1% yield from Rhaα1PαNP and ethyl L-thio-D-thiamorphosyranose. The disaccharide product is a derivative of the common oligosaccharide unit of antigenic lipopolysaccharides from Pseudomonas (33).

Fucosidases

α-L-Fucose is commonly found as the terminal unit in the carbohydrate moiety of many important glycoproteins involved in a variety of biologic events. For example, α1,2-fucosylated LacNAc is the determinant of blood-group type O antigen. Disaccharides Fucα1,3GlcNAc and Fucα1,4GlcNAc are part of Lea and Leb antigens, respectively.

By using α-L-fucosidases from Penicillium multicolor (34), Aspergillus niger (35, 36), Corynebacterium sp. (35), and other sources (37–39), α1,3-linked L-fucosides have been obtained. By using an α-L-fucosidase from Ampullaria niger (35), α1,6-linked disaccharide Fucα1,6GalSEt has been produced in 14% yield from Fucα1PαNP and Galα1PαNP. With Fucα1PαNP as the donor, α1,4-disaccharide Fucα1,4(6OEt)GlcNH2 SEt has been obtained in 50% yield using an α-L-fucosidase from bovine kidney (40) and in 33% yield using an enzyme from bovine testes (41). Compared with α-L-fucosidases, only a few β-D-fucosidases were reported to form β-D-fucosides (42–44).

For example, β-glucosidase from Thal rosewood and almond both catalyse transglycosylation of alcohols to synthesize α-L-fucosyl fucosides with high yields using PNP-Fuc as the donor (43).
Sides to afford sulfated disaccharides with intermediate is involved (displacement mechanism in which a covalent glycosyl-enzyme rect displacement, whereas retaining glycosidases use a double inversion or inversion of stereochemistry, and thus they are classified glycosidase Agrobacterium sp catalytic carboxylate nulceophile mutant of a retaining glycosynthase was reported by Withers et al. on a single synthesis of oligosaccharides but no hydrolytic activity. Since the glycosynthase-catalyzed formation of the glycosidic bond (51). An \( \alpha \)-linked tri saccharide unit of xylo glucan resulted in 15% yield when \( \alpha \)-xylosyl fluoride was used as a glycosyl donor and pNP-\( \beta \)-cellobioside was used as an acceptor.

**Glycosynthases — glycosidase mutants**

Glycosylation or transglycosylation reactions catalyzed by wild-type glycosidases inevitably suffer from low yields and unpredictable regioselectivity because of the hydrolytic nature of the glycosidases and the lack of regioselectivity of many glycosidases. The creation of glycosynthases, a class of glycosidase mutants, has greatly advanced the field of glycosidase-catalyzed synthesis of oligosaccharide.

Glycosynthases are a new class of glycosidase mutants developed based on the mechanistic understanding of glycosidases. Glycosidases hydrolyze glycidosidic links with net reten- tion or inversion of stereocchemistry, and thus they are classified into retaining glycosidases and inverting glycosidases. Inverting glycosidases proceed via a general acid-base-catalyzed direct displacement, whereas retaining glycosidases use a double-displacement mechanism in which a covalent glycosyl-enzyme intermediate is involved (Fig. 3a and b). A glycosynthase is formed by replacing a single active site nucleophile (usually a glutamate or an aspartate residue) of a glycosidase with a neutral alanine or serine residue. An activated glycosyl donor, usually a glycosyl fluoride of the opposite anomeric configuration to that of the natural substrate, is used for the glycosynthase-catalyzed formation of the glycosidic bond (Fig. 3c). These mutated enzymes have increased activity for the synthesis of oligosaccharides but no hydrolytic activity. Since the first glycosynthase was reported by Withers et al. on a single catalytic carboxylate nucleophile mutant of a retaining glycosidase Agrobacterium sp. \( \beta \)-glucosidase (Abg) E358A mutant (52), many glycosynthases have been generated by Withers et al. and others. The first glycosynthase of an endoglycosidase was a \( \beta \)-lactamase mutant of the retaining L.\( \beta \)-1,4-glucanase from Bacillus licheniformis reported by M alet and Planas (53). Currently, about 20 glycosynthases from ten different glycosidase families have been produced. Except for a recently reported mutant of an exo-oligoxylanase (Rex) from Bacillus halodurans (54), which is the first glycosynthase derived from an inverting enzyme, all others are mutants of retaining glycosidases. Despite the development of glycosynthases, many naturally occurring carbohydrate structures have not been obtained by the glycosynthase-catalyzed reaction, because the strategy relies on the availability of gene sequence and on the understanding of the catalytic mechanism of the glycosidases, and this information is not available for all glycosidases known to date.

**Glycosyltransferase-catalyzed synthesis of oligosaccharides**

Leloir-type glycosyltransferases are enzymes that catalyze the transfer of a monosaccharide from activated sugar nucleotide donor to an acceptor. Because of their high efficiency and rego- and stereo-specificities, they offer significant advantages in the formation of glycosidic bonds, and thus, they have been widely used in the synthesis of carbohydrate-containing structures. Glycosyltransferases have been classified based on the carbohydrate residue that they transfer.

**Sialyltransferases**

Chemical synthesis of sialosides is considered one of the most difficult glycosylation reactions because of a hindered tertiary amionic carbon and the lack of a participating auxiliary functional- ity in the carbon next to the amonic carbon in sialic acids (55, 56). Sialyltransferase-catalyzed glycosylation is believed to be the most efficient approach for the production of sialic acid-containing structures.

Sialyltransferases catalyze the transfer of a sialic acid residue from CMP-sialic acid to a galactose, GalNAc, or another sialic acid residue. Sialyltransferases from bacterial and mammalian sources have been extensively studied and used in the enzymatic synthesis of sialosides, sialylglycoconjugates, and enzymatic modification of cell surface.

Rat liver \( \alpha \)2,3SiaT and \( \alpha \)2,6SiaT have been the most widely used mammalian sialyltransferases for the synthesis of sialo- sides. Both enzymes have broad donor and acceptor substrate specificity. They can tolerate a variety of modifications at the Neu5Ac moiety of CMP-Neu5Ac (57–59) and Gal on accep- tors (60–62). Four SiaTs from rat liver or porcine submaxil- lary glands were studied for their abilities to transfer synthetic 9-substituted sialic acid analogs onto N- or O-linked glyco- protein glycans. They all accepted CMP-9-alpha-Neu5Ac as the donor substrate. In contrast, 9-amino-Neu5Ac was only ac- cepted by \( \alpha \)2,6SiaT form rat liver (63).

Wong et al. have studied the acceptor specificity of a recom- binant \( \alpha \)2,3SiaT from Neisseria gonorrhoeae by using several synthetic oligosaccharides, glycolipids, and glycopeptides. Lac- tose, its \( \beta \)-lactosyl derivatives, sulfated oligosaccharides, and the \( \Delta \)54L-1 glycopeptide carrying a sulfotyrosine residue were all excellent acceptors for the enzyme. However, most glycolipids were poor substrates for this bacterial enzyme (55).

Using genes cloned from Neisseria meningitides, a fusion protein has been constructed that has both CMP-Neu5Ac synthetase and \( \alpha \)2,3SiaT activities. This fusion protein could sialylate various oligosaccharide acceptors with Neu5Ac, as well as Neu5Gc and \( \beta \)-propionyl-neuraminic acid, in high yields. A 100 gram-scale of \( \alpha \)2,3-sialylactose was produced using this protein with the regeneration of sugar nucleotide CMP-Neu5Ac.
Enzymatic Synthesis of Carbohydrate-Containing Biomolecules

(a) Retaining glycosidase

(b) Inverting glycosidase

(c) Glycosynthase

(d) Thioglycoligase

(e) Thioglycosynthase

Figure 1 Catalytic mechanisms for wild-type and engineered glycosidases. (a) Retaining glycosidase, to hydrolyze glycosidic links using the double-displacement mechanism; (b) inverting glycosidase, to hydrolyze glycosidic links via a general acid/base-catalyzed direct displacement; (c) glycosynthase, mutation of the catalytic nucleophile allows transfer an activated sugar donor (such as a glycosyl fluoride) to suitable acceptors; (d) thioglycoligase, mutation of the catalytic acid/base allows transfer of an activated \( \beta \)-glycoside to suitable acceptors; (e) thioglycosynthase, mutation of both the catalytic nucleophile and the acid/base residues allows transfer of glycosyl fluoride donor to acceptors.

Very recently, the fusion protein and a recombinant human \( \alpha \)2,6-sialyltransferase (hST6Gal-I) have been used in gram-scale synthesis of sialosides (65).

A bifunctional sialyltransferase Cst-II from \( C. \) jejuni OH4384 carrying both \( \alpha \)2,3- and \( \alpha \)2,8-sialyltransferase activities has been used in the enzymatic synthesis of ganglioside mimic GD3 (66).

Another multifunctional bacterial enzyme that has been reported recently is Pasteurella multocida sialyltransferase (PmST1 or tPm0188Ph) (67, 68). It has four different activities, including an \( \alpha \)2,3,6,8-sialidase, \( \alpha \)2,3-sialidase, and trans-sialidase activities. PmST1 is a highly active sialyltransferase with broad donor and acceptor substrate specificity; therefore it is a powerful tool.
in synthesizing diverse sialosides that contain structurally modified sialic acids.

As more than 50 different sialic acids derivatives have been found in nature, sialyltransferases with flexible donor and acceptor specificity are preferred by chemists. A α2,6SiaT cloned from Photobacterium damsela (Pd2,6ST) can efficiently transfer Neu5Ac, KON, Neu5Gc, and their derivatives with extensive modifications at C-5 and C-9 in the sialic acid residues from their CMP-activated forms to the acceptors (8). Pd2,6ST also has very relaxed acceptor specificity. For example, this bacterial enzyme has been applied for the enzymatic sialylation of glycosyltransferases and glycoconjugates (70). The exhaustively studied glycosyltransferases and has been the most firmly established enzyme responsible for the biosynthesis of a galactan core in the major structural component of mycobacterial cell wall, which has been reported (85). Ga1T transfers Ga1 residues from UDP-galactofuranose (UDP-Galf) to several OST-linked to the neuraminic acid residue in the sequence with 71% yield. More interesting, Pd2,6ST has shown activity in transferring sialic acid to N- and O-linked glycoproteins (69).

Galactosyltransferases

Galactosyltransferases are a family of enzymes that catalyze the addition of a galactose residue from activated sugar nucleotide donor uridine diphosphate galactose (UDP-Gal) to different acceptor substrates in different links.

_Gal1,4Galα1,3Galβ1,4Glcα1_ (1) is one of the most exhaustively studied galactosyltransferases and has been the most commonly used glycosyltransferase in the synthesis of oligosaccharides and glycoprotein units (70). The Galα1,4GalT exhibits poly-functional donor specificity and can transfer modified galactose from their activated UDP form to acceptors to afford corresponding LacNAc analogs (71, 72). The Galα1,4GalT is equally relaxed in respect to its acceptor specificity, and the modified Ga1NAc acceptor can be tolerated by the Galα1,4GalT. As this enzyme has a highly relaxed substrate specificity, it has been exploited in the synthesis of many non-natural LacNAc-based structures. Pdα1,4GalT from bacterial sources such as N. pylori (73) and N. meningitidis (74) have also been cloned, but examples of their application in the synthesis of galactosides have been limited (75).

Carbohydrate structures bearing a Ga1α1,3Galβ1 terminus are α-Galactosyl epitopes. The interaction of these epitopes on the surface of animal cells with anti-α-Galactosyl antibodies in human serum is believed to be the main cause in antibody-mediated hyperacute rejection in xenotransplantation. The unique enzyme responsible for the formation of α-Gal epitopes is α1,3Galactosyltransferase (α1,3GalT). Fang et al. (76) reported that a truncated bovine α1,3GalT (80-368) can be produced as a soluble recombinant enzyme on a large scale with highly specific activity. A variety of α3,4galactosylated oligosaccharides were synthesized using such a recombinant enzyme.

The 2° and 6° donor analogs of globotriose Ga1α1,4Galβ1,4Glcα1,4GlcOAcOR, which is the known receptor for Shiga and Shiga-like toxins, were prepared using UDP-2-deoxy-Gal and UDP-6-deoxy-Gal along with α1,4GalT from N. meningitidis in 11% and 95% yield, respectively (77). The inexpensive galactosyl fluoride has also been used as a donor for an α1,4Ga1T to produce an α1,4-linked galactoside in the presence of a catalytic amount of UDP (78).

Various α1,3GalTα1 have been cloned and expressed from bacterial (46, 79) and mammalian sources (80-82). Core 1, Ga1α1,3GalNAc α1-3mannosyltransferase, is the major constituent of O-glycan core structures in many cells. The core 1 structure is also called T antigen in pathologic studies. Core 1 β1,3galactosyltransferases (core 1 β1,3Galα1T), which are responsible for the synthesis of core 1 disaccharides, have been purified recently (83, 84).

Compared with galactopyranosyltransferases, a few studies on galactofuranosyltransferase (GalT), which is responsible for the biosynthesis of a galactan core in the major structural component of mycobacterial cell wall, have been reported (85). Ga1T transfers Ga1 residues from UDP-galactofuranose (UDP-Galf), instead of UDP-Gal, to the growing galactan chain. Very recently, Lorary et al. (86) reported the high-level expression and purification of a novel bifunctional recombinant galactofuranosyltransferase from Mycobacterium tuberculosis H37Rv, which can produce both α1,5- and α1,6-galactofuranose links in an alternating fashion.

_N-acytylgalactosaminyltransferase

The human blood group A glycosyltransferase (α1,3N-acetylgalactosaminyltransferase) is responsible for the biosynthesis of blood group A antigens. It catalyzes the transfer of a Ga1NAc from UDP-Ga1NAc to the C-3 hydroxyl of Gal residue in (Fucu1,2Ga1α1,6GlcOR) to form blood group A antigen Ga1NAcα1,3Fucα1,2Ga1α1,6GlcOR. α1,3-3Ga1NAcCT from N. meningitidis was reported to transfer a Ga1NAc to 1-thio-D-galactose to produce a trisaccharide, which was further used to synthesize lacto-N-neotetraose (87). Wang et al. (88) had constructed α1,3-N-acetylgalactosaminyltransferase UDP-N-acetylglucosamine C4 epimerase fusion protein and had synthesized a variety of globotetraose and isoglobotetraose derivatives using this fusion protein and UDP-GlcnAc. This construction alters the donor substrate requirement of the reaction from the high-cost UDP-GalNAc to UDP-GlcnAc. Recently, gram-scale synthesis of GD3, GT3, GM2, GD2, GT2, and GM1 gangoside oligosaccharides was reported by Blixt et al. using α1,4-N-acetylgalactosaminyltransferase (α1,4GalNAcCT) from Campylobacter jejuni along with other glycosyltransferases (89).

_N-Acetylgalactosaminyltransferases

N-Acetylgalactosaminyltransferases (Ga1NAcTs) play important roles in the synthesis of 0-glycan core structures and in the branching and subsequent elaboration of N-linked glycans on glycoproteins. GlcNAcTs I-VI (differing in their specificities and glycosidic bonds formed with trimannose core-containing N-glycans) have been studied on their substrate specificity and their application. For example, Core 2 Ga1NAc (β1,6Ga1NAcCT) has been used in the synthesis of 0-linked core 2-type 6° blood type (90). Norberg et al. (91) had investigated the donor and acceptor substrate specificity of β1,3Ga1NAcCT from N. meningitidis. This enzyme can use both UDP-GlcnAc and UDP-GalNAc as donors. It is also capable of tolerating deoxy derivatives at any position other than C-3 of the Gal residue in Ga1/2LacNAcOAcOR acceptor.

Fucosyltransferases

Several mammalian fucosyltransferases (FucTs) have been characterized, and their application to the synthesis of fucose-
noting oligosaccharides has been well studied. Recently Drumliard et al. (92) developed an efficient bioengineering method for large-scale production of fucosyl α1,2-linked oligosaccharides from lactose. 2-Fucosyl lactose and lacto-N-neofucopentaose-1 (LNF1), an H-2 antigen oligosaccharide were produced by expressing Helicobacter pylori α1,2-fucosyltransferase in metabolically engineered E. coli cells. Trisaccharide 2’-fucosylactose could be synthesized industrially on a multiton scale for nutraceutic applications. The pentasaccharide LNnF-1, which contains the H-2 antigen structure, could be used as a precursor for the synthesis of other antigens of the ABH histo-blood-group system.

Mannosyltransferases

The β1,4-mannosidic linkage is common to most asparagine-linked oligosaccharides (N-glycans) in glycoproteins as part of their pentasaccharide core structure. The formation of the glycosidic link in Manα1,4GlcNAc is particularly challenging for chemical synthesis. Glycosidases have been used, but they require a large excess of the expensive chitobiose acceptor and produce low yields. A highly active recombinant β1,4-mannosyltransferase was reported (93), and this enzyme was used to produce the core triasaccharide of N-glycans with a yield of 80% using UDP-mannose as donor and the synthetic chitobioyl phosphatidyl as acceptor. A recombinant α1,2-mannosyltransferase was recently reported to use GDP-S-thio-Man as a sugar nucleotide donor to produce the corresponding S-mannosides (94). S-Glycans are expected to have potentially stronger affinity for receptors compared with their natural counterparts, and they may also find other applications in hydrodase-resistant vaccines.

Enzymatic formation of S-glycosidic bonds

Natural occurring carbohydrates contain monosaccharides linked through O-glycosidic bonds. Thioglycosides, in which the glycosidic oxygen atom has been replaced by the sulfur atom, can be tolerated by most biological systems and are less susceptible to acid/base or enzyme-mediated hydrolysis. They are invaluable in the studies of glycosidases and are gaining interest as targets for pharmaceutical industry. Many efforts have been focused on the synthesis of thioglycosides by conventional chemical synthetic methods. Only recently, mutants of glycosidases and glycosyltransferase have been constructed and applied to form S-linked oligosaccharides using thiolated acceptors.

Glycosidase mutants (thioglycoligase and thioglycosynthase) in the synthesis of S-linked oligosaccharides

A glycosidase mutant, in which the acid/base carboxyl residue was replaced by a catalytically inactive residue, was recently developed by Wilthers et al. as a novel strategy for the synthesis of S-linked oligosaccharides (95). This new class of glycosidase mutants is named “thioglycoligase” (Fig. 3a). Two different alanine acid/base mutants of retaining β-glycosidases, a β-glucosidase from Agrobacterium sp. Abg E171A and a β-mannosidase from Cellulomonas fimi Man2A, were used for the formation of thioglycosidic bonds. The readily available dinitrophenyl glycoside donor DNP-Glc or DNP-Man was incubated with the mutant enzyme and the acceptor to provide corresponding thioglycosaccharides in good yields. With this new methodology, Stick and Stubbs (96) have synthesized various types of S-linked disaccharides using glycosidase mutant Abg E171A. Incubation of the glucose donor DNP-Glc and various thiol acceptors with Abg 171A provided corresponding β1,4-, β1,3- and β1,6-S-linked disaccharides, but it did not work for the formation of β1,2-thio-linkage. More recently, the generation of α-thioglycosides has been successfully achieved for the first time using thioglycoligases derived from glycoside hydrolase family 315. solütARIOculos α-glycosidase and E. coli α-mannosidase (97).

The thioglycoligase technology was further extended with the advent of thioglycosynthases (Fig. 3a), double-mutants of retaining glycosidases which lack both catalytic nucleophile and catalytic acid/base residues and can efficiently catalyze the thioglycosidic bond formation using glycosyl fluoride donors and thioglycoside acceptors (98). Reaction of the synthetic α-glycosyl fluoride with the acceptors in the presence of the double mutant Abg E171A, E558D at neutral pH afforded two β1,4-S-linked disaccharides with 51% and 45% yields, respectively.

Glycosyltransferase mutants in the synthesis of S-linked glycodies

Rich et al. (99) have reported a new enzymatic method for the synthesis of thioglycosides using glycosyltransferase. Incubation of a mixture of thiol acceptor and UDP-Gal with unit quantities of a recombinant bovine α1,3-galactosyltransferase in the presence of DTT afforded the thio-linked tetrasaccharide in 92% yield instead of the expected trisaccharide product because of the second glycosyl transfer that occurred after the initial transfer of a galactosyl residue to the thiolated acceptor. The desired trisaccharide, which is an analog of the Clostridium difficil toxin A binding ligand, was obtained in near-quantitative yield after treatment with an α-galactosidase from green coffee beans. Furthermore, their initial results confirmed that a β1,3-N-acetylglucosaminyltransferase from N. meningitides can catalyze the transfer of UDP-GlcNAc to a thiol donor to afford a thio-linked trisaccharide. This is the first example of using glycosyltransferases in synthesizing thioglycosides. Such thioglycosides could serve as important immunogens and components of conjugate vaccines.

Enzymatic Synthesis of Polysaccharides

Polysaccharides, along with nucleic acids and proteins, belong to three major important classes of naturally occurring biopolymers. Polysaccharides are important biomacromolecules with unique physical properties that lead to advanced biomaterials and biomedical applications (100). For example, glycosaminoglycans (GAGs) are linear polysaccharides containing repeating disaccharide units of a hexosamine and a uronic acid.
Heparin, heparan, and chondroitin are three prevalent glycosaminoglycans. Vertebrates use glycosaminoglycans in structural, recognition, adhesion, and signaling roles. Chemical synthesis of naturally occurring polysaccharides is considered to be impractical. Most polysaccharides, especially those from bacteria origins, are obtained by purification from natural sources or from cell culture, enzymatic approaches have been increasingly applied to obtain some structures.

Heparin/Heparan Sulfate

Heparin and heparan sulfate are linear polysaccharides with a repeating disaccharide unit of 1,4-linked uronic acid (D-glucuronic or L-iduronic acid) and D-glucosamine residues. Both uronic acid and glucosamine can contain sulfo groups at different positions including 2-O-sulfo substitution of the uronic acid residue and 2-N-, 3-O-, and 6-O-sulfo substitution in the glucosamine residue. Heparin and heparan sulfate are structurally related glycosaminoglycans that participate in numerous important biologic processes, such as blood coagulation, viral and bacterial entry and infection, angiogenesis, and cancer development.

The enzymatic synthesis of heparin and heparan sulfate is currently under extensive investigation using glycosyltransferases of E. coli (101, 102) and heparanase synthase of P. multocida (103, 104). Rosenberg et al. reported an approach to rapidly assemble anti-coagulant III-binding classic and nonclassic anti-coagulant heparan polysaccharides using the enzymes involved in the heparin sulfate biosynthesis (105). Recently Chen et al. (106) have described a method for the enzymatic sulfation of multigram amounts of heparin sulfate with specific functions using immobilized sulfotransferases combined with a 3-phosphoadenosine 5′-phosphosulfate regeneration system. Because the recombinant sulfotransferases are expressed in bacteria and the method uses a low cost sulfo donor, it can be readily used to synthesize large quantities of anti-coagulant heparin drug or other biologically active heparan sulfates.

Hyaluronan

Hyaluronan (hyaluronic acid, HA) is a highly anionic unbranched linear polymer containing a (1→4)-linked uronic acid (β-D-gluconic or α-L-iduronic acid) repeating unit, which plays important roles in modulating cell adhesion, signaling, and motility. Several enzymes responsible for HA synthesis, namely Hyaluronidases A and B Synthases, have been cloned from bacteria and mammals (107, 108). HA synthase from Streptococcus equisimilis was employed for milligram-scale synthesis of HA (109). In this reaction system, UDP-sugars (UDP-GlcA and UDP-GlcNAc) were effectively regenerated by the catalyses of several enzymes, synthetic HA was produced in 90% yield. In addition, mutated HA synthase from Type A Pasteurella multocida (PmHA S1-703 aa) was recently used for the stepwise synthesis of HA, which has a monodisperse molecular mass of up to 20 sugar units (110). On the other hand, using hyaluronidase (HAase) to synthesize HA has also been reported (111, 112). HA oligomers of up to 22 sugar units were synthesized by enzymatic reconstruction of HA chains using the transglycosylation reaction of Bovine Testicular HAase (111).

Chondroitin

Chondroitin (Ch) and chondroitin sulfate (ChS) are naturally occurring heteropolysaccharides belonging to the family of GAGs. Ch is a nonsulfated derivative of ChS, which consists of a G(A)1,3GlcNAc2 disaccharide repeating unit connected through β-N-acetylglicosamido bonds. ChS exists predominantly as polysaccharide side chains of proteoglycans in extracellular matrixes where it plays important roles in the bioactivities of living systems. Ch is widely used as a therapeutic material for the prevention or alleviation of symptoms of diseases.

The enzymatic polymerization to provide synthetic Ch and its derivatives catalyzed by hyaluronidase has been reported by Kobayashi et al. (113). Synthetic oxazoline monomers were recognized and catalyzed by ovine testicular HAase (OTH) to produce Ch and the polymerization behaviors greatly depend on the reaction conditions. The Mn value of synthetic Ch reached 4600, which corresponds to that of naturally occurring Ch.

Polysialic acids

Polysialic acids (PSA) are linear homopolymers of N-acetyllactosaminic acid and N-glycolylneuraminic acid joined by α2,8- and α2,9-ketosidic linkages. Polysialic acids play many important biologic roles. For example, polysialylation of mammalian neural cell adhesion molecules affects cell-cell adhesive interactions during embryogenesis (114). Polysialic acid is also a component of many bacterial capsular polysaccharides that are important virulence factors. Sialyltransferases catalyze the addition of sialic acid to form diverse carbohydrate molecules. The substrate specificity of α2,8 and α2,8/α2,9-polySiaT toward glycolipids and sialylated polysaccharides has been characterized (115–117). For example, the neu5 gene product from E. coli K92 that exhibits α2,8/α2,9-polySia transferase activity both in vitro and in vivo has been described for the first time by Wong et al. (116).

Cellulose

Cellulose is a linear polysaccharide of β-D-glucan linked dehydration β-glucose repeating units. Cellulose is the most abundant polysaccharide on earth. It is a major component in higher plant cell walls and has been used as raw material in paper, fibers, and lumber industries.

Enzymatic synthesis of cellulose has been achieved by cellulase. For example, incubation of β-cellobiose fluoride with a cellulase from Trichoderma viride can produce cellulose in 54% yield with DP around 22 after 12h. In addition, change of the reaction conditions (substrate concentration or organic solvent concentration) enabled the selective synthesis of the water-soluble cellobiosepolysaccharides (118).

Amylose

Amylose is a linear polymer of glucose mainly linked with an α-(1,4)-glycosidic bond. Amylosepolysaccharides have been effectively prepared by polycondensation of α-D-maltosyl fluoride using an α-amylase from Aspergillus oryzae as the catalyst in a mixed solvent of methanol-phosphate buffer (119).
Amylose was also prepared via in vitro polymerization of D-glucosyl phosphate catalyzed by a potato phosphorylase (120). A large excess amount of Glc-1-P is required in this equilibrium-controlled reaction.

Xylan

Xylan, a xylose polymer having a β(1,4)-glycosidic linkage in the main chain, is the important component of hemicellulose in plant cell walls. Kobayashi et al. (121) reported that synthetic xylan that consists exclusively of xylose units can be prepared by the crude cellulase (containing xylosidase polymerization of β-xylosyloxy fluoride as a monomer in a mixed solvent of acetonitrile/acetate buffer. The reaction proceeded to produce β(1,4)-linked synthetic xylan with a degree of polymerization of 23.

Chitin

Chitin is a β(1,4)-linked polymer of N-acetyl-D-glucosaminopyranose. It is used as structural material in nature, such as the main component in the cell walls of fungi, the exoskeletons of insects and other arthropods, as well as in some animals. It attracts much interest in several scientific and application areas as a multifunctional substrate.

Enzymatic Synthesis of Glycoproteins

Many biologically important proteins are carbohydrate-containing glycoproteins. Two major types of glycoproteins are O-linked glycoproteins (a sugar moiety is linked to the β-hydroxy group of either a serine or a threonine residue in the polypeptide) and N-linked glycoproteins (an oligosaccharide is linked to the amide-side chain of asparagine residue of the polypeptide). For O-linked glycosylation, three different ways can occur: 1) A single GlcNAc residue may be reversibly added to proteins in the cytoplasm or nucleus, 2) initiation of synthesis of proteoglycans may occur by the addition of xylose in the Golgi, and 3) mucin-type O-linked glycosylation is initiated in the ER. As glycoproteins usually present a mixed population of glycan structures, the isolation of homogeneous glycoproteins from natural sources in significant quantity is impossible. The chemical method for the de novo synthesis of large glycoproteins is currently unavailable. Although challenging, enzymatic synthesis of structurally defined glycoproteins is feasible.

O-GlcNAc-modified glycoproteins

Many nuclear and cytosolic proteins are ρ-O-GlcNAc-modified, which is the covalent attachment of ρ-GlcNAc to Ser/Thr residues in the proteins. Nearly 80 proteins bearing the O-GlcNAc group have been identified to date (124). Understanding the functional roles of O-GlcNAc requires the development of new strategies for the detection and study of O-GlcNAc-modified proteins. Khodde et al. (125) reported the direct, high-throughput analysis of O-GlcNAc-proteins using a chemoenzymatic approach. An engineered galactosyltransferase was exploited to selectively label O-GlcNAc-proteins with a ketone-based tagged Gal, which permitted the enrichment of low-abundance O-GlcNAc species and the localization of the modification site. Another chemical strategy was developed by Bertozi et al. for identifying O-GlcNAc-modified proteins from living cells. O-GlcNAcase is capable of transferring azidoacetylgalactosamine (GlcNAz) from UDP-GlcNAz to known protein substrates, such as recombinant nuclear pore protein p62. These O-GlcNAz-modified proteins can be covalently derivatized with various biochemical probes at the site of protein glycosylation using the Staudinger ligation reaction (126). Such a strategy for in vivo modification of target proteins can provide a rapid means for the identification of sites of O-GlcNAc modification on purified recombinant protein. However, so far no reports exist on enzymatic synthesis of O-GlcNAc-linked glycoproteins by directly transferring GlcNAc to proteins. Zhang et al. (127) developed a novel strategy for the generation of homogenous glycoprotein by selective incorporation of glycosylated amino acids (GlcNAc-c-c1-0-) into proteins. An orthogonal M. jannaschii RNA synthetase (TyRS) mutant was identified that selectively incorporated GlcNAc-c-c1(1-0-) into myoglobin. With the evolved TyRS synthetase, milligram quantities of homogenous glycoprotein was obtained upon coexpression of the synthetase, suppressor tRNA, and TAG-mutated myoglobin genes in E. coli in medium that contained the glycosylated amino acid. This approach has advantages of high selectivity, efficiency, and high yield without the generation of other glycoforms or unmodified forms of myoglobin.

O-GalNAc-modified glycoproteins

Mucin-type O-linked glycoproteins that are involved in biologic interactions in higher eukaryotes are ρ-O-GalNAc modified. Strategies for the detection and study of O-GalNAc-modified proteins have been developed (128-131). Bertozi et al. have developed a metabolic labeling approach toward proteomic analysis of mucin-type O-linked glycosylation, a unique azido chemical tag introduced permits the identification of O-GalNAc-modified proteins from complex cell lysates (128).

O-GalNAc-linked glycopeptides have been synthesized using glycoenzymes. Aijikara et al. (132) recently used the transglycosylation activity of an endo-β-N-acetylgalactosaminidase from Streptomyces sp. for the synthesis of O-linked glycopeptides. Using Galp1,3GalNAc-Cl as the glycosyl donor, this enzyme can transfer the disaccharide to a serine in a hexapeptide and produced a Galp1,3GalNAc-linked hexapeptide in 11% yield.
proteins prepared by conventional chemical methods (137). Greater structural homogeneity as compared with PEGylated enables the manufacturing of long-acting protein drugs with sialic acid-PEG to oligosaccharides, G-CSF, IFN-α, and GM-CSF. Selective addition of sialic acid conjugated with PEG to the introduced GalNAc expressed in specific serine and threonine residues in non-glycosylated proteins (134). The process involves enzymatic GalNAc glycosylation at specific sites and threonine residues in non-glycosylated proteins expressed in Escherichia coli, followed by the enzymatic transfer of sialic acid conjugated with PEG to the introduced GalNAc residues. The strategy was applied to three therapeutic polypeptides, GM-CSF, IFN-α, and GM-CSF. Selective addition of sialic acid-PEG to O-linked GalNAc on a protein provides a novel, highly site-selective mechanism for PEGylation, which enables the large-scale manufacturing of long-acting protein drugs with greater structural homogeneity as compared with PEGylated proteins prepared by conventional chemical methods (137).

N-linked glycoproteins

N-linked glycoproteins are found in a wide range of organisms ranging from archaea to mammals and other eukaryotes (138). The major method for enzymatic synthesis of N-linked glycoproteins is to elaborate the existing sugar moieties on the proteins. Glycosidases are used to trim the existing glycan structures, other sugars can then be put back on by glycosyltransferases (this process is called glycoprotein remodeling). For example, the synthesis of homogeneous unnatural glycoform of ribonuclease B (RNase-B, which presents a mixture of high-mannose glycoforms) was achieved by endo H degradation to GlcNAc2-Man9GlcNAc2 (139) followed by elaboration with a galactosyltransferase, a fucosyltransferase, and a sialyltransferase to form the desired glycoform (140).

Some endo-2-3 sialidase/glycosidase (ENGase), such as Endo A from Arthrobacter protophormiae (141) and Endo M from Mucor hiemalis (141), have transglycosylation activities. They can transfer a large intact oligosaccharide to a GlcNAc-peptide acceptor in a single step. This methodology has also been successfully employed in constructing glycoproteins. Incubation of the tetra- or hexasaccharide oxazoline with homogeneous GlcNAc-Rnase B in the presence of Endo-A afforded the glycoproteins in 82% and 96% yield, respectively (143).

Hamilton et al. (149) reported a new approach to produce complex human N-glycoproteins using yeast-based expression systems. They observed that Man9GlcNAc-Rnase B expression in endogenous yeast glycosylation pathways, five active eukaryotic proteins were properly expressed and localized in the yeast Pichia pastoris. Targeted localization of the enzymes enabled the generation of human glycoproteins with homogeneous N-Glycan structures. This is a big step toward producing therapeutic glycoproteins. Glycoengineered yeast Pichia pastoris was also employed to produce human antibodies with specific human N-glycan structures (150) and human glycoproteins with complex terminally sialylated N-glycans (151).

Enzymatic Synthesis of Glycolipids

Glycolipid synthesis (GSL) is a field that has experienced significant advancements in recent years. Enzymatic synthesis of GSLs using recombinant glycosyltransferases has become a powerful tool for the production of complex glycosphingolipids, mimics of native GSLs, and GSL analogs with modified structures. This technique allows for the precise and controlled synthesis of GSLs, which are important in various fields such as pharmacology, biology, and medicine.

Glycosphingolipids (GSL) are amphiphilic molecules that are synthesized by sequential actions of glycosyltransferases. They consist of a glycosphingolipid backbone linked to a carbohydrate moiety. The synthesis of GSLs is typically initiated by the transfer of a sialic acid-containing glycosphingolipid, called ganglioside, to a glycosphingolipid acceptor. The phosphorylation of GSLs can be achieved by the catalysis of a series of glycosyltransferases to provide targeted glycopeptides. Glycosyltransferases are also often used in the elaboration of the existing glycan structures on glycopeptides and glycoproteins. For example, a sulfated N-terminal fragment of PSGL-1 (P-selectin glycoprotein ligand 1, which is an important cell-surface glycoprotein on vascular endothelial cells) has been prepared using this method (134).

Modification of protein drugs by covalent attachment of polyethylene glycol (PEG) can prolong the half-life and enhance the pharmacodynamics of therapeutic proteins. Defrees et al. (136) recently developed a novel approach for site-directed PEGylation using glycosyltransferases to attach PEG to O-glycans. The process involves enzymatic GalNAc glycosylation at specific serine and threonine residues in non-glycosylated proteins expressed in Escherichia coli, followed by the enzymatic transfer of sialic acid conjugated with PEG to the introduced GalNAc residues. The strategy was applied to three therapeutic polypeptides, GM-CSF, IFN-α, and GM-CSF. Selective addition of sialic acid-PEG to O-linked GalNAc on a protein provides a novel, highly site-selective mechanism for PEGylation, which enables the large-scale manufacturing of long-acting protein drugs with greater structural homogeneity as compared with PEGylated proteins prepared by conventional chemical methods (137).

N-linked glycoproteins

N-linked glycoproteins are found in a wide range of organisms ranging from archaea to mammals and other eukaryotes (138). The major method for enzymatic synthesis of N-linked glycoproteins is to elaborate the existing sugar moieties on the proteins. Glycosidases are used to trim the existing glycan structures, other sugars can then be put back on by glycosyltransferases (this process is called glycoprotein remodeling). For example, the synthesis of homogeneous unnatural glycoform of ribonuclease B (RNase-B, which presents a mixture of high-mannose glycoforms) was achieved by endo H degradation to GlcNAc2-Man9GlcNAc2 (139) followed by elaboration with a galactosyltransferase, a fucosyltransferase, and a sialyltransferase to form the desired glycoform (140).

Some endo-2-3 sialidase/glycosidase (ENGase), such as Endo A from Arthrobacter protophormiae (141) and Endo M from Mucor hiemalis (141), have transglycosylation activities. They can transfer a large intact oligosaccharide to a GlcNAc-peptide acceptor in a single step. This methodology has also been successfully employed in constructing glycoproteins. Incubation of the tetra- or hexasaccharide oxazoline with homogeneous GlcNAc-Rnase B in the presence of Endo-A afforded the glycoproteins in 82% and 96% yield, respectively (143).

Hamilton et al. (149) reported a new approach to produce complex human N-glycoproteins using a yeast-based expression system. They observed that Man9GlcNAc-Rnase B expression in endogenous yeast glycosylation pathways, five active eukaryotic proteins were properly expressed and localized in the yeast Pichia pastoris. Targeted localization of the enzymes enabled the generation of human glycoproteins with homogeneous N-Glycan structures. This is a big step toward producing therapeutic glycoproteins. Glycoengineered yeast Pichia pastoris was also employed to produce human antibodies with specific human N-glycan structures (150) and human glycoproteins with complex terminally sialylated N-glycans (151).

Enzymatic Synthesis of Glycolipids

Glycolipid synthesis (GSL) is a field that has experienced significant advancements in recent years. Enzymatic synthesis of GSLs using recombinant glycosyltransferases has become a powerful tool for the production of complex glycosphingolipids, mimics of native GSLs, and GSL analogs with modified structures. This technique allows for the precise and controlled synthesis of GSLs, which are important in various fields such as pharmacology, biology, and medicine.

Glycosphingolipids (GSL) are amphiphilic molecules that are synthesized by sequential actions of glycosyltransferases. Regarding the structural variation of GSLs, the expression of enzymes acting on lactosylceramide (LacCer) is the rate-determining step. Several genes that code for enzymes responsible for the synthesis of the carbohydrate moiety of glycosphingolipids have been recently identified.

Sialic acid-containing glycosphingolipids, called gangliosides, have various important biologic functions. In vertebrates, almost all gangliosides are synthesized from a common precursor, ganglioside 

1,4GlcNAc could serve as a substrate for ENGase-catalyzed transglycosylation. Wang et al. then explored sugar oxazolines as donor substrates for Endo A and Endo M in the N-glycopeptide synthesis using large peptides and partially deglycosylated Rnase-B as model systems (146–148). They found that Endo-A and Endo-M both can effectively catalyze the reaction between the synthetic oligosaccharide oxazoline and the GlcNAc-heptapeptide to form the glycopeptide in 82% and 78% yields, respectively. The newly formed glycosidic bond was indicated to be a 1,4-type by NMR analysis, which confirms that the ENGase-catalyzed transglycosylation using sugar oxazoline as the glycosyl donor proceeded in a stereo- and regiospecific manner to form the desired glycopeptide (147). The ENGase-catalyzed transglycosylation method has also been successfully employed in constructing glycoproteins. Incubation of the tetra- or hexasaccharide oxazoline with homogeneous GlcNAc-Rnase B in the presence of Endo-A afforded the glycoproteins in 82% and 96% yield, respectively (148).

Hamilton et al. (149) reported a new approach to produce complex human N-glycoproteins using a yeast-based expression system. They observed that Man9GlcNAc-Rnase B expression in endogenous yeast glycosylation pathways, five active eukaryotic proteins were properly expressed and localized in the yeast Pichia pastoris. Targeted localization of the enzymes enabled the generation of human glycoproteins with homogeneous N-Glycan structures. This is a big step toward producing therapeutic glycoproteins. Glycoengineered yeast Pichia pastoris was also employed to produce human antibodies with specific human N-glycan structures (150) and human glycoproteins with complex terminally sialylated N-glycans (151).
Glycosylation of Natural Products

Many biologically active natural products are glycospides. It is becoming more and more obvious that glycosylation affects bioactivity or selectivity of natural products, such as anticancer drugs and antibiotics (160, 161). Altering glycans has become a focus within natural product chemistry and pharmaceutical sciences. Recent developments in molecular glycobiology make it possible to develop new and more effective glycodrugs. The glycosides vancomycin and teicoplanin are clinically important antibiotics. As the carbohydrate portions of these molecules affect biologic activity, developing efficient strategies to make carbohydrate derivatives is of interest in searching for vancomycin and teicoplanin analogs. Glycosyltransferases from Amycolatopsis orientalis were used by the Walsh group to produce variant sugar forms on both vancomycin and teicoplanin classes of glycopeptide antibiotics using nucleotide diphospho-sugar (NDP-sugars) as the glycosyl donors (162). Although GaTs are generally perceived as unidirectional catalysts, very recently, Thomson et al. (163) reported that four glycosyltrans-ferses from two distinct natural product biosynthetic pathways (calicheamicin and vancomycin) readily catalyzed reversible re-actions, allowing sugars and aglycons to be exchanged with ease. Using these new reactions, more than 70 differentially glycosylated calicheamicin and vancomycin variants were pro-duced.

The glycosyltransferase DesVII can catalyze the attach-ment of TDP-D-desosamine onto 12- and 14-membered macro-lactone rings to make methymycin/methemymycin and nar-bormycin/pikromycin, respectively, in Streptomyces venezuelae. The purified DesVII is active only in the presence of another protein, DesVIII, at high pH. Y-C17 (10-deoxy-D-methylmycin) was synthesized in preparative scale using TDP-D-desosamine as the donor and 10-deoxy-D-methylmycin as aglycon acceptor (164).

Steroidal glycosides constitute a structurally and biologically diverse class of molecules such as cardenolides or saponins. They have received considerable recent attention because of their physiologic and pharmacologic activities. Synthesis of steroidal glycosides via enzymatic systems is still rare as most enzymes are not available. β-Galactosidosidase from Aspergillus oryzae was used in the synthesis of various cardiac glycosides which are unstable under basic or acidic condition during chemi-synthetic synthesis (165). A peripllogenin β-D-glucoside was prepared in 37% yield using the biotransformation of digitoxigenin by hybrid cells (166). Thiem et al. have re-port ed the enzymatic synthesis of the β-glucosidories of estradiol and ethynylestradiol on a preparative scale by incubating bovine liver UDP-glucuronol transferase with corresponding phenolic aglycone substrates (167).

Peptidoglycan is a polymer of carbohydrate chains connected by peptide crosslinks. It is the major component in bacterial cell wall. The enzymes that synthesize peptidoglycan layers have received special attention because many known antibiotics function by blocking peptidoglycan synthesis (168). Among these enzymes, bacterial transglycosylases (GTases) are located on the external surface of the bacterial membrane where they polymer-ize lipid II, a disaccharide anchored to the membrane by a 55 carbon undecaprenyl chain (169, 170). Terrak et al. re-port ed that the penicillin-binding protein (PBP) 1b of E. coli catalyzes the conversion of C55H89 lipid-transported disaccha-ride pentapeptide units into polymeric peptidoglycan (171).
bifunctional enzyme catalyzes both glycosylation for the formation of the carbohydrate backbone of murein and transpeptidation for the formation of the interstrand peptide links (172).

Strategies in Enzymatic Synthesis of Carbohydrates and Glycoconjugates

Protein engineering of glycosidases and glycosyltransferases

Other than protein crystal structure-based construction of glycosidase mutants, such as glycosylases, thioglycosylases, and thioglycoligases discussed above, for the synthesis of carbohydrate-containing structures, protein crystal structure-based rational design of glycosyltransferase with altered substrate specificity has also been performed. In addition, directed evolution has emerged as a powerful tool in generating mutants with designed function.

Crystal structure-based rational design of glycosyltransferase mutants

Many X-ray crystal structures have been reported for glycosyltransferases, which makes the structure-based redesign of glycosyltransferases feasible. For example, crystallographic analysis of the Y289L mutation of β1,4GalT predicted that such a mutation should provide space for a C-2 N-acetyl group (173). Using the GalT tolerating the C-2 donor sugar substrates, Khidekel et al. (174) reported that Y289L mutant β1,4GalT could efficiently transfer a Gal analog with a ketone functionality at C-2 from donor to the O-GlcNAc glycosylated protein CREB. The ketone functionality on the Gal was used as a tag to identify O-GlcNAc posttranslational modified proteins.

Directed evolution

Wild-type enzymes are powerful tools in the synthesis of carbohydrates. Enzymatic synthesis using wild-type enzymes, however, is limited by the enzyme instability and the restriction on the substrates that can be recognized by the wild-type enzymes. Protein crystal structure-based rational design and mutants generated by site-directed mutagenesis provide some solution for the problems. Some properties of the mutants, however, cannot be obtained by rational design because of the limited information available about the structure-function relationship of proteins. Directed evolution has emerged as a promising approach to obtain enzymes with broader or altered substrate specificity. The key for a successful directed evolution process involves an efficient screening system to identify randomly generated mutants with desired properties. This has been a challenge for applying directed evolution approaches for glycosyltransferases. A recent report by Mayer et al. describes the development of a novel agar plate-based coupled-enzyme screen to select an improved glycosynthase form a library of mutants (175). Withers et al. developed a novel fluorescence-based high-throughput screening methodology for the directed evolution of sialyltransferases. Using this methodology, a library of >10^6 SaT mutants was screened and a new sialyltransferase variant was discovered, which had more than 400-fold higher catalytic activities than the parent enzyme (176).

Feng et al. (177) reported the converting a β-glycosidase of Thermus thermophilus to a β-transglycosidase by directed evolution. Mutants possessing high transglycosase activity are identified by using a simple screening procedure. Using these mutants, self-condensation of nitrophenyl glycosides can reach nearly quantitative yield, whereas transglycosylation of maltose and cellobiose can reach 60% and 75%, respectively.

One-pot multiple-enzyme synthesis

Glycosyltransferases have been used in combination with other enzymes such as sulfotransferases, proteases, lipases, and acetyltransferases to synthesize complex oligosaccharides. Most of these enzymatic reactions can be successfully achieved under similar conditions, which makes it possible to carry out a multiple-enzyme reaction in one-pot system to produce oligosaccharide and their derivatives. One-pot multiple-enzyme reaction can simplify the product purification process without the necessity of isolation of intermediates, thus avoiding the compound loss during the multiple purification steps. More important, it avoids the use of high cost sugar nucleotides and their analogs by using less-expensive starting materials.

Yu et al. (8, 67) recently reported a highly efficient and convenient one-pot three-enzyme chemoenzymatic approach for the synthesis of libraries of α2,6-linked and α2,3-linked sialosides containing naturally occurring as well as non-natural sialic acids. In this method, sialic acid modifications can be chemically introduced at the very beginning, onto the six carbon sugar precursors (ManNAc or mannose) of sialic acids. These ManNAc or mannose analogs can then be directly converted to naturally occurring and non-natural sialosides in one-pot using three enzymes, including a sialic acid aldolase, a CMP-sialic acid synthetase, and a sialyltransferase, without the isolation of intermediates. (Fig. 2) Such process takes advantage of the relaxed substrate specificity of all the enzymes involved in the synthesis.

Sugar nucleotide regeneration

The application of glycosyltransferases in the glycosidic link formation is limited because of the high cost of sugar nucleotides. This limitation has led to the development of sugar nucleotide recycling systems in enzymatic glycosylations. These regeneration systems require the use of only catalytic quantities of the sugar nucleotides, which can be regenerated continuously from inexpensive precursors, making the large-scale enzymatic synthesis of complex carbohydrates economically viable.

Sugar nucleotide regeneration systems have been developed by mimicking their biosynthetic pathways. For example, Wang et al. reported the incorporation of a UDP-GalNAc-4-epimerase in a UDP-GalNAc regeneration system, which in combination with a GalNAcT, was used in the synthesis of globotetraose and its derivatives (Fig. 3) (88).
Enzymatic Synthesis of Carbohydrate-Containing Biomolecules

Figure 2: One-pot three-enzyme chemoenzymatic synthesis of sialosides containing sialic acid modifications. In this strategy, mannose or ManNAc derivatives are chemically or enzymatically synthesized. These compounds are then used by a recombinant E. coli K-12 sialic acid aldolase to obtain sialic acids and their derivatives followed by an N. meningitidis CMP-sialic acid synthetase for the formation of CMP-sialic acids. From which, sialic acids can be transferred to lactose, LactNAc, galactose, GalNAc, or their derivatives by a multifunctional P. multocida sialyltransferase (PmST1) or a P. damsela α2,6-sialyltransferase (Pd2,6ST) to form α2,3- or α2,6-linked sialosides in one pot without the isolation of intermediates.

Figure 3: Enzymatic synthesis of globotetraose with in situ UDP-GlcNAc regeneration. The system contains UDP-GlcNAc pyrophosphorylase (GlmU), pyruvate kinase (PykF) and inorganic pyrophosphatase (PPase) from E. coli K12; GlcNAc-phosphate mutase (Agm1) from S. cerevisiae; and GlcNAc kinase (GlcNAcK) from C. albicans.

Other examples of multi-enzyme systems with sugar nucleotide regeneration for large-scale economic synthesis of many oligosaccharides have also reported (76, 178). Trisaccharide Galα1,3Galβ1,4GlcNAcβO(CH2)8CO2Me was enzymatically synthesized by combining four enzymes (sucrose synthase, UDP-Glc 4-epimerase, β1,4-GalT, and α1,3-GalT) in one pot using acceptor GlcNAcβ1-O(CH2)8CO2Me with in situ UDP-Gal regeneration (Fig. 4). This is an efficient and convenient method for the synthesis of Galα1,3Galβ1,4GlcNAc epitope, which plays an important role in various biologic and immunologic processes.

Solid-phase enzymatic synthesis

Solid-phase synthesis of oligosaccharides is a practical and convenient method as it simplifies the product purification and makes combinatorial process feasible (179). Combining the highly efficient enzyme-catalyzed glycosylation with solid-phase techniques offers a particularly convenient approach for the synthesis of oligosaccharides and glycoconjugates.
solid-phase enzymatic synthesis has been applied for sLe\(^\alpha\) tetrasaccharide (180) and a biomedically important tetrasaccharide Neu5Ac\(\alpha_2,3\)Gal\(\beta_1,4\)GlcNAc\(\beta_1,3\)Gal (an inhibitor of the attachment of H. pylori to mucous cells) (181). Other polymer supported syntheses of oligosaccharides have also been reported. Nishimura et al. reported a new approach for highly efficient chemoenzymatic synthesis of glycopeptide using a combination of solid-phase and water-soluble supports (Fig. 5) (182). This approach was efficient and versatile for the synthesis of a glycopeptide library.

**Immobilized enzymes in the synthesis—superbeads**

The use of immobilized enzymes in bioprocesses offers greater productivity because the same enzyme molecules can be used multiple times over a long period of time. Wang et al. reported a novel approach that transfers in vitro multiple enzyme sugar nucleotide regeneration systems onto solid beads (the superbeads) that can be used and reused as common synthetic reagents for production of glycoconjugates. First the multiple
enzymes involved in the biosynthetic pathway for the regeneration of UDP-Gal were expressed as N-terminal His-tagged enzymes; these enzymes were then co-immobilized on Ni-NTA beads by taking the advantage of the affinity of the histidine tag and nickel-nitrilotriacetic acid (NTA) resin. The sugar nucleotide regeneration superbeads can then be conveniently combined with glycosyltransferases for the synthesis of specific oligosaccharides. Using the UDP-Gal regeneration beads (containing galactokinase, galactose-1-phosphate uridylyltransferase, glucose-1-phosphate uridylyltransferase, and pyruvate kinase), which are required for the regeneration of UDP-Gal from UTP, with a truncated bovine α1,3-galactosyltransferase, the triasaccharide Galα1,3LacOBn was synthesized in 72% yield in gram-scale (183). Thse beads were reused three times during a three-week period and still retained 90% enzyme activity. Also the nickel beads can be recharged for more use after removing the deactivated enzymes. The superbeads can also be used in combination with multiple immobilized glycosyltransferases to generate target oligosaccharides. For example, using 2 equiv of galactose as the starting sugar and GlcNAc(1→3)Galα1→3Galβ1→4GlcNAc as the glycosyl acceptor, a combination of α1,3GalT and β1,4GalT immobilized on the UDP-Gal superbeads was used to produce a pentasaccharide with an overall yield of 76% (183). Combined with galactosyltransferase, the superbeads become a very useful reagent to synthesize a variety of oligosaccharides and their derivatives.

Another approach by Chen et al. (186) used a single product-producing E. coli strain (superbug) containing all the genes required for the synthesis of complex oligosaccharides, including fucosylated oligosaccharides, sialylated oligosaccharides, and lipooligosaccharides. Another approach by Chen et al. (186) used a single product-producing E. coli strain (superbug) containing all the genes required for the synthesis of complex oligosaccharides, including fucosylated oligosaccharides, sialylated oligosaccharides, and lipooligosaccharides. This superbug technology has been used in the synthesis of αGal epitope, P1 trisaccharide antigen, globotriose, and their derivatives. The superbug technology is cost effective because only catalytic amounts of the high energy phosphates are required. The use of a single bacterial strain instead of multiple strains avoids transport of reaction intermediates among strains and prevents the complication of maintaining the growth of different strains.

**Conclusion And Perspective**

Significant progress has been made over the past two decades for the application of enzymes in the synthesis of complex carbohydrate-containing biomolecules. Challenges are still exist. Taking advantage of the increasingly available genomic data, advancement in enzymatic synthesis of biomolecules relies on the identification and characterization of enzymes with better stability and novel or more flexible substrate specificity, a better understanding of the enzyme mechanism, genetic manipulation for tailor-made enzymes, and new methodology development.

**References**

16

Enzymatic Synthesis of Carbohydrate-Containing Biomolecules


84. Jahn M, Chen H, Kulingar J, Marles J, Warren RA, Withers SG. Thioglycosides: double mutant glycosidases that
99. Rich JR, Szpakenko A, Palcic MM, Bundle DR. Glycosyltrans- 
sferase-catalyzed synthesis of polysialoglycans. Angew. Chem. 
100. De Luca C, Manfred L, Martini I, Crescenzi F, Shen GJ, 
Hoshi H, Nakagawa H, Nishiguchi S, Iwata K, Niikura K, 
O'Regan M, Wong CH. Enzymic synthesis of hyaluronic acid 
and its derivatives catalyzed by hyaluronidase. J. Am. Chem. Soc. 2004:126:4179- 
4187.
101. Rich JR, Szpacenko A, Palcic MM, Bundle DR. Glycosyltrans- 
sferase-catalyzed synthesis of polysialoglycans. Angew. Chem. 
102. DeAngelis PL, Oatman LC, Gay DF. Rapid chemoenzymatic 
synthesis of classical and non-classical an-
103. Hoshi H, Nakagawa H, Nishiguchi S, Iwata K, Niikura K, 
O'Regan M, Wong CH. Enzymic synthesis of hyaluronic acid 
and its derivatives catalyzed by hyaluronidase. J. Am. Chem. Soc. 2004:126:4179- 
4187.
104. Kobayashi S, Shimada J, Kashiwa K, Shoda S. Enzymic poly-
merization of α-d-maltopyranosyl fluoride using α-amylose as catalyst: 
a new approach for synthesis of maltosylglycosides. Macromolecules 1992:25:3237- 
3241.
105. Kuberan B, Beeler DL, Lech M, Wou ZL, Rosenberg RD. 
Characterization of the glycosylation enzyme from the 
Escherichia coli K5 capsular type 1 cluster and identification and 
characterization of the glucuronic acid site. J. Biol. Chem. 1998:273:12572- 
12577.
106. Griffiths G, Cook NJ, Gottfridson E, Lind T, Lisholt K. Identification of KFA, 
a protein essential for the biosynthesis of the Escherichia coli 
K5 capsular polysaccharide, is an α-d-1,5-D-glucan:α-glycosyl- 
transferase. The formation of a membrane-associated K5 biosyn-
thetic complex requires KFA, KFB, and KFC. J. Biol. Chem. 2000:275:27311- 
27315.
107. DeAngelis PL, White CL. Identification and molecular cloning of a 
heparosan synthase from Pasteurella multocida type D. J. 
108. Kobayashi S, Fujikawa S, Ohmae M. Enzymatic synthesis of 
monodisperse hyaluronan oligosaccharides with im-
109. Cunningham BA, Hemperly JJ, Murray BA, Prediger EA, Brack-
enburg R, Eidelberg GM. Neural cell adhesion molecules struc-
ture, immunoglobulin-like domains, cell surface modulation, and 
110. DeAngelis PL, Jing W, Drake RR, Achyuthan AM. Identification 
of the glycosyltransferase enzyme from the 
Escherichia coli K5 capsular gene cluster and identification and 
characterization of the glucuronic acid site. J. Biol. Chem. 1998:273:12572- 
12577.
111. Saitoh H, Takagaki K, Majima M, Nakamura T, Matsuki A, 
Kobayashi S, Shimada J, Kashiwa K, Shoda S. Novel method for 
poly saccharide synthesis using an enzyme: the first in vitro 
synthesis of cellulose via a nonenzymatic path using cellulase as 
for polysaccharide synthesis using an enzyme: the first in vitro 
synthesis of cellulose via a nonenzymatic path using cellulase as 
for polysaccharide synthesis using an enzyme: the first in vitro 
synthesis of cellulose via a nonenzymatic path using cellulase as 
Enzymatic Synthesis of Carbohydrate-Containing Biomolecules


Enzymatic Synthesis of Carbohydrate-Containing Biomolecules


173. Kanamitsu B, Qasba PK. Structure-based design of β1,4-galactosyltransferase I (βGαCT) with equally efficient N-acetylglactosaminyltransferase activity: point mutation broadens α-gal-T1 donor specificity. J. Biol. Chem. 2002;277:20833–20839.


Further Reading


- Kanamitsu B, Qasba PK. Structure-based design of β1,4-galactosyltransferase I (βGαCT) with equally efficient N-acetylglactosaminyltransferase activity: point mutation broadens α-gal-T1 donor specificity. J. Biol. Chem. 2002;277:20833–20839.


Enzymatic Synthesis of Carbohydrate-Containing Biomolecules


See Also

Enzyme Catalysis, Chemical Strategies for Glycan Biosynthesis in Mammals
Glycosyltransferases, Chemistry of Glycolipids, Synthesis of Glycopeptides and Glycoproteins, Synthesis of
The biologic significance of glycans (i.e., carbohydrates and saccharides) and the need to obtain structurally defined material for study have resulted in a proliferation of strategies for the synthesis of this extraordinary class of biomolecules. Each strategy carries with it both advantages and disadvantages. The two major approaches, chemical and enzymatic synthesis, will be discussed with an emphasis on the former.

Linear synthesis

The glycosylation reaction is fundamental to glycan synthesis (see the article "Key Reactions in Glycan Synthesis"). The iterative addition of monosaccharide units in a synthetic sequence constitutes a linear glycosylation strategy. With respect to overall synthetic efficiency, this strategy is not optimal, particularly for large targets (5). The classic Koenigs-Knorr reaction, which was discovered in 1901, signaled the advent of chemical oligosaccharide synthesis (6). The preparation and use of glycosyl bromides and chlorides as glycosyl donors in combination with appropriately protected glycosyl acceptors and heavy metal activators/promoters (typically silver or mercury salts) remains to date a formidable methodology for glycan construction despite toxicity and high cost. Because of the general instability of these intermediates, coupled with the harsh conditions necessary for their preparation, glycosyl bromides have been limited largely to a linear strategy, as opposed to a more efficient convergent (block) approach.

Convergent synthesis

The development of novel glycosylation methodologies in the 1970s and 1980s featuring more stable glycosyl donors such as orthoesters (7), fluorides (8), trichloroacetimidates (9), thio-glycosides (10), N-pentenyl glycosides (NPGs) (11), glycals (12), and sulfoxides (13), among others, enabled access to larger oligosaccharides by virtue of a convergent approach. Mild preparation methods of these donors coupled with chemical stability (for purification and storage) greatly increased the overall efficiency of glycan synthesis. Routine conversion of oligosaccharide intermediates (blocks) into glycosyl donors and subsequent coupling with various acceptors has enabled chemists to access more complex structures in an efficient manner; for example, Schmidt et al. (14) have efficiently applied the block synthesis strategy in the preparation of various glycans.

Chemical Synthesis of Glycans

Glycans encountered in nature in the form of glycoconjugates (glycoproteins and glycolipids) mediate a remarkable variety of biologic events, including inflammation, fertilization, cell growth and development, tumor growth and metastasis, host-pathogen interactions, and the storage and transfer of information (see the article "Glycans in Information Storage and Transfer") (2, 3). To study the structure and function of these biomolecules, they must be obtained in pure form. As isolation from natural sources often yields impure material by virtue of microheterogeneity (4), recourse is made to either chemical or enzymatic synthesis (or a combination of the two referred to as chemoenzymatic synthesis).
Selective, two-stage and latent-active strategies

The necessity to recruit and use protecting groups on both the glycosyl donor and acceptor is synthetically cumbersome (see the article “Glycan Synthesis, Protection and Deprotection”). A distinct advantage of manipulations of oligosaccharide intermediates before coupling lowers overall yield. In an ideal synthesis, the number of protecting group operations should be held to a minimum. Toward this goal, various strategies have been introduced that take advantage of selective activating conditions. A generalized example of this method is the selective activation of a glycosyl donor (halide, trichloroacetimidate) in the presence of an acceptor bearing a leaving group (thioglycoside or NPG). After the coupling event, the disaccharide is reacted with another acceptor in the presence of a suitable activator. This strategy cuts down the number of protecting group operations in the synthesis. Thioglycosides and N-pentenyl glycosides are unique donors in that they are stable to traditional protecting group manipulation and can be activated selectively in the glycosylation event under appropriate conditions. A recent example of this strategy was Barchi’s preparation of a tumor-associated T antigen building block (15).

In 1984, the Nicolaou group introduced a two-stage glycosylation strategy (16), which is also represented in Fig. 1. In this approach, thioglycoside intermediates are converted into glycosyl fluorides with N-bromosuccinimide (NBS) and (diethylamino)sulfox trifluoride (DAST) and are coupled subsequently (Stage 2) with a thioglycosyl acceptor. This process can be repeated, offering quick access to large structures. The Nicolaou group (17) has leveraged this methodology to prepare a host of complex glycans. A more powerful two-stage glycosylation approach is K. Ahn’s glycosyl sulfoxide methodology wherein a phenyl thioglycosides is used as a stable anomer protecting group, which is oxidized to the anomic sulfone with m-chloroperbenzoic acid (MCPBA). Subsequent triflation activates the anomic center for glycosylation. The K. Ahn group (18) has demonstrated the utility of this method in the synthesis of several blood group antigens.

Roy (19) introduced a latent-active glycosylation strategy in 1992 to describe the use of a stable (latent) anonic group that can be converted into a reactive (active) anonic group. In this context, a 4-nitrophenyl thioglycoside bearing a free hydroxyl (acceptor) can be coupled selectively with a 4-N-arylphenyl thioglycoside donor. The nitro group can then be reduced and acetylated, and this procedure can be repeated. By modulating the electronics of the thioarane, selective glycosylation can be accomplished. A further example of this strategy is the use of O-allyl glycosides (latent) by Boons (20), which are converted into O-vinyl glycosides (active) and activated subsequently with trimethylsilyl triflate (TMSCl) in the presence of a suitable acceptor (Fig. 1).

Orthogonal strategy

An orthogonal glycosylation strategy was outlined in 1994 by O’Gara (21) that featured the use of two glycosyl donors (phenyl thioglycosides and fluorides) meeting the following two criteria: (1) Either anomic group should be activated selectively in the presence of the other; and (2) both anomic groups should remain compatible with subsequent protecting group manipulation (Fig. 1). This strategy was utilized in the highly efficient synthesis of an extended blood group B determinant (22).

In 1997, Boons (23) reported a highly convergent and efficient synthesis of Group B Type III Streptococcus hexacarhide using three unique donor sets (ethyl thioglycosides, cyanoethylidene, and NPGs) that required no protecting group steps beyond those used in building-block synthesis.

Chemoslective glycosylation

Fraser-Reid (24) introduced a chemoselective strategy (armed–disarmed glycosylation) in 1988 after making the observation that NBS-promoted hydrolysis of NPGs bearing electron-donating protecting groups (alkyl ethers such as benzyl) on the C-2 hydroxyl proceeded much faster relative to those bearing electron-withdrawing protecting groups (esters such as acetyl). This strategy led to the experiment in which two monosaccharides bearing identical leaving groups (4-pentenylxoy) are coupled wherein a “disarmed” monosaccharide bearing a free hydroxyl bears an ester on the C-2 position and an “armed” donor possesses an alkyl ether substituent on the C-2 position. In the event, no self-coupled product was formed (Fig. 2). This strategy differs from the orthogonal strategy in that the same anomic group is used. Chemoslectivity is achieved by modulating the electronics of the incipient oxocarbenium ion. By proximity, the C-2 ester of the coupled product into an alkyl (benzyl) ether. More specifically, the less-reactive promoter iodonium dicollidine perchlorate (IDCP) can be used in the first coupling, whereas the more powerful N-isodosuccinimide (NIS)/triflic acid (TfOH) or NIS/trithylsilyl triflate (TESOTf) combination is sufficient to glycosylate disarmed donors (25).

In addition to changing the electronics of the C-2 protecting group, Fraser-Reid et al. have introduced the use of cyclic acetal protecting groups (1,3-dioxanes and 1,3-dioxolanes) to “disarm” glycosides. This process works by placing torsional strain on the pyranose scaffold, as the oxocarbenium ion is best stabilized by a d/hedral angle of 0° (C6D6–C5H5). Conformational restraints placed by cyclic protecting groups on the saccharide preclude achieving the desired angle (26). Ley et al. have expanded on this torsional control element with the introduction of dipropionate (27) and cyclohexane-1,2-dicarboxylic acids (28). By using different promoters in the sequencing of these glycosylations (mild promoters for early coupling, stronger promoters for later coupling), the number of synthetic operations performed on advanced intermediates is reduced (Fig. 2). Friese and Danielczyk (29)
Key Strategies for Glycan Synthesis

Selective Activation

![Selective Activation Diagram]

Example 1: Y = 4-pentenyloxy or SR, X = Br or -OC(NH)CCl₃
Example 2: Y = SR, X = SeR (R = alkyl or aryl)

Two-Stage Activation

![Two-Stage Activation Diagram]

Stage 1: DAST NBS
Stage 2: SnCl₂ AgClO₄

Latent-Active Glycosylation

![Latent-Active Glycosylation Diagram]

Orthogonal Glycosylation

![Orthogonal Glycosylation Diagram]

Figure 1 Various strategies in chemical glycan synthesis.

has applied the chemoselective strategy to glycals for the preparation of 2-deoxyglycosides.

Chemoselective glycosylations can also be realized by changing the steric environment about the anomeric center (aglycon). Boons and colleagues (30) have shown that ethyl thiglycoside donors can be activated selectively in the presence of sterically encumbered dicyclohexylmethyl thiglycoside acceptors with IDCP. This methodology has been exploited in the synthesis of various glycan targets. The resulting coupled product can be activated again with the more powerful NIS/TfOH system in the presence of a suitable acceptor (Fig. 2).

One-pot glycosylation

The assembly of complex glycans via either a linear or a block approach is often subject to tedious purification (chromatography) steps that lower the overall chemical yield. It would be desirable to use an approach that strives to lower the number of purifications, particularly as the synthesis advances. Toward this end, many groups have developed and refined one-pot...
Key Strategies for Glycan Synthesis

Chemoselective Glycosylation

One-pot glycosylations (OPGs) fall into three different categories. Each category is based on the particular strategy being employed, which includes reactivity-based (chemoselective), selective (including orthogonal donor sets), and iterative one-pot glycosylations. Historically, Kahne and coworkers in 1989 disclosed the first OPG strategy for the synthesis of a cyclamycin trisaccharide, which was based on 1) the relative reaction rates (chemoselectivity) of electronically distinct phenyl-sulfoxides and 2) the use of a TMS ether as a latent acceptor in the second glycosylation. The desired trisaccharide was isolated in 25% overall yield. In 1994, Ley and coworkers demonstrated that various glycosyl donors could be sequenced in a one-pot strategy; namely, glycosyl bromides, fluorides, or trichloroacetimidates were first coupled with a thioglycosyl acceptor. The addition of a second, reducing end acceptor and a thiophilic promoter (NIS) resulted in trisaccharide formation. Yields as high as 84% for the overall process were realized, showcasing the efficiency of this strategy as well as the promoter compatibility, which can be a problem. The method has been extended to branched structures as well. Takahashi has advanced the orthogonal OPG strategy in an impressive synthesis of the heptasaccharide phytoalexin elicitor (HPE) in 24% overall yield.

Ley and coworkers were the first to quantify the reactivity of various ethyl thioglycosides. In 1999, Wong and coworkers disclosed a programmable one-pot glycosylation strategy in which a database of relative reactivity values (RRVs) was established for numerous monosaccharides and disaccharides bearing the p-methylthiophenyl (STol) anomeric group. With this information in hand, they could optimize a one-pot synthetic sequence in which a series of donors bearing free hydroxyls (with the exception of the first unit, which is fully protected) would be sequentially reacted in the order of decreasing RRV terminating with a “reducing end” cap. Disaccharides (or larger oligosaccharides) can be sequenced to offer access to branched oligosaccharides. A software program, OptiMer, was developed to select the appropriate building blocks from the database to perform an optimal one-pot oligosaccharide synthesis. Coupling of STol glycosides can be affected with a variety of thiophilic reagents. This software program has enabled the synthesis of a variety of important glycans, including poly-N-acetylactosamines, Fucose...
Key Strategies for Glycan Synthesis

**Selective (including Orthogonal) One-Pot Glycosylation**

- X = Br, OC(NH)CCl₃, F
- X = SR, OH

**Reactivity-Based (Chemoselective) One-Pot Glycosylation**

GM₁ (small-cell lung cancer epitope) (39), the Lewis Y hexasaccharide (colon cancer epitope) (40), and Globo H (breast cancer epitope) (41) in addition to combinatorial carbohydrate libraries (42). As of 2005, 600 thioglycoside building blocks have been entered into the database with RRVs ranging from 1 to >10⁶ (43). Wong and colleagues have introduced a novel activator, N-(phenylthio)caprolactam/Tf₂O, for programmable one-pot synthesis (44).

Yu and co-workers synthesized several saponins in a one-pot fashion using monosaccharide and disaccharide trichloroacetimidates and thioethyl glycosides (45). In 2000, Mukaiyama and colleagues established that glycosyl fluorides or phenylcarbonates could be used in concert with thioglycosides for the synthesis of several trisaccharides (46). This methodology was used in the synthesis of HPE using fluorides and thioethyl glycosides (47).

**Iterative One-Pot Glycosylation**

The iterative OPG strategy combines favorable elements of both chemoselective and orthogonal approaches in that it uses similar activating conditions for each coupling step yet does not rely heavily on differentiating building blocks with protecting groups to modulate and optimize reactivity (armed–disarmed concept), respectively. Danishefsky et al.'s glycal assembly method (48), Gin’s chemoselective dehydrative glycosylation (49), and Yamago et al.’s bromine-activated selenoglycosides (50) all reflect the utility of the iterative glycosylation strategy in glycan synthesis.

Crich and colleagues have demonstrated elegantly that the stereoselective synthesis of β-mannosides with glycosyl sulfoxides proceeds via anomeric α-triflate intermediates (51). Crich and Sun (52) observed that premixing the mannosyl sulfoxide with the promotor system (triflic anhydride and a hindered pyridine base) before the addition of the acceptor, as opposed to direct triflation of both donor and acceptor, led to higher β:α ratios. The authors theorized that the increased selectivity originated from an S₄N₂ displacement reaction of the acceptor and that the α-triflate derived from the sulfoxide. Crich et al. (53) also showed that when thioglycosides and aromatic bromides were subject to triflation, the same α-triflate intermediate was observed via low-temperature NMR studies. In summary, Crich and colleagues had demonstrated the viability of the donor pre-activation concept. These findings account for the feasibility of the iterative OPG strategy, which is generalized in Fig. 3.

Yu and co-workers synthesized several saponins in a one-pot fashion using monosaccharide and disaccharide trichloroacetimidates and thioethyl glycosides (45). In 2000, Mukaiyama and colleagues established that glycosyl fluorides or phenylcarbonates could be used in concert with thioglycosides for the synthesis of several trisaccharides (46). This methodology was used in the synthesis of HPE using fluorides and thioethyl glycosides (47).

**Reactivity-Based (Chemoselective) One-Pot Glycosylation**

GM₁ (small-cell lung cancer epitope) (39), the Lewis Y hexasaccharide (colon cancer epitope) (40), and Globo H (breast cancer epitope) (41) in addition to combinatorial carbohydrate libraries (42). As of 2005, 600 thioglycoside building blocks have been entered into the database with RRVs ranging from 1 to >10⁶ (43). Wong and colleagues have introduced a novel activator, N-(phenylthio)caprolactam/Tf₂O, for programmable one-pot synthesis (44).

Yu and co-workers synthesized several saponins in a one-pot fashion using monosaccharide and disaccharide trichloroacetimidates and thioethyl glycosides (45). In 2000, Mukaiyama and colleagues established that glycosyl fluorides or phenylcarbonates could be used in concert with thioglycosides for the synthesis of several trisaccharides (46). This methodology was used in the synthesis of HPE using fluorides and thioethyl glycosides (47).

**Iterative One-Pot Glycosylation**

The iterative OPG strategy combines favorable elements of both chemoselective and orthogonal approaches in that it uses similar activating conditions for each coupling step yet does not rely heavily on differentiating building blocks with protecting groups to modulate and optimize reactivity (armed–disarmed concept), respectively. Danishefsky et al.'s glycal assembly method (48), Gin’s chemoselective dehydrative glycosylation (49), and Yamago et al.’s bromine-activated selenoglycosides (50) all reflect the utility of the iterative glycosylation strategy in glycan synthesis.

Crich and colleagues have demonstrated elegantly that the stereoselective synthesis of β-mannosides with glycosyl sulfoxides proceeds via anomeric α-triflate intermediates (51). Crich and Sun (52) observed that premixing the mannosyl sulfoxide with the promotor system (triflic anhydride and a hindered pyridine base) before the addition of the acceptor, as opposed to direct triflation of both donor and acceptor, led to higher β:α ratios. The authors theorized that the increased selectivity originated from an S₄N₂ displacement reaction of the acceptor and that the α-triflate derived from the sulfoxide. Crich et al. (53) also showed that when thioglycosides and aromatic bromides were subject to triflation, the same α-triflate intermediate was observed via low-temperature NMR studies. In summary, Crich and colleagues had demonstrated the viability of the donor pre-activation concept. These findings account for the feasibility of the iterative OPG strategy, which is generalized in Fig. 3.

In iterative OPG, the glycosylating species (donor) is treated with the promoter system in the absence of the acceptor. A priori, the activated species must be sufficiently stable so as not to decompose before glycosylation and yet reactive enough to undergo reaction with the acceptor. In the event, a glycosyl donor is first activated with the promotor, followed by the addition of an acceptor bearing a stable anomeric group (alkylthio moiety). The glycosylation occurs irrespective of the protecting group ensemble of the acceptor. This process removes the burden of electronically tuning each glycoside via protecting group manipulation; moreover, protecting groups can be selected so as to streamline deprotection steps (endgame). The process can be repeated (iterated) and terminated with a final “cap” acceptor (Fig. 3).

In 2003, the van Boom/van der Marel group cleverly used 1-hydroxy and thioglycosides in a sequential OPG strategy in
which the powerful Ph₃SOTf₂O/TTBP (2,4,6-tri-tert-
butylpyrimidine) (54) promoter system was leveraged to syn-
thesize an α-Gal epitope trisaccharide in 80% yield (stepwise
synthesis was accomplished in 69% overall yield) (55). Initial
activation of the 1-hydroxy donor with the above promoter sys-
tem followed by reaction with a thioglycoside acceptor yielded
a disaccharide intermediate. As a by-product of the reaction is
Ph₂O, more Tf₂O was introduced to activate the thioglycoside
in the presence of an acceptor to furnish the target trisaccharides
(56, 57).

In 2004, Huang et al. outlined a general, iterative OPG stra-
tegy based on “pre-activation” of the glycosyl donor for the syn-
thesis of several trisaccharides and tetrasaccharides using 2,701
glycosides and a P-TolSO/AgOTf/TMS-AW300 (in situ prepara-
tion of p-TolSOII) promoter system. This approach obviates the
need to tune electronically or torsionally each donor, which is
critical in reactivity-based OPG (58). This strategy was applied
in a four-component, one-pot synthesis of α-Gal Pentasaccha-
ride, which is a glycan epitope responsible for the shortcomings
of p-Gal to human xenotransplantation. Yields for the one-pot se-
quence ranged from 39% to 41% (59). Recently, Huang et al.
(60) introduced a novel α-thiophenylamine, benzene sulfimyl
morpholine (BSM), for iterative OPG processes. The novel reagent
was used in the synthesis of several trisaccharides.

Solid-phase synthesis

The advent of automated oligopeptide and oligonucleotide syn-
thesizers revolutionized the fields of protein and nucleic acid
chemistry, respectively. These technologies have greatly en-
abled the study and understanding of structure and function by
making routine access to these structures straightforward.

At the heart of these methodologies lies a solid-phase strat-
egy (61) for which Merrifield was awarded the 1984 Nobel
Prize in Chemistry. Solid-phase methodologies possess sev-
eral advantages: purification is made easy (filtration as op-
posed to chromatography), excess reagents can be employed
to drive reactions to completion via mass action as purifica-
tion is rendered trivial, and the process lends itself well to
automation. Critical to a successful solid-phase oligosaccha-
ride synthesis are the proper selection of the following items:
1) polymer support such as insoluble polyurethane or solu-
ble polyethylene glycol (PEG), 2) orthogonal protecting groups
sets, 3) linker (which can be regarded as a resin-bound pro-
tecting group), and finally 4) building blocks. In addition,
it is desirable to have “on-resin” analytical techniques avail-
able so as to monitor the progress of the synthesis. Toward
that goal, various techniques, including high-resolution magic
angle-spinning NMR (62), gated-decoupling 13C NMR spec-
troscopy using 13C-enriched protecting groups (63), fluorinated
protecting groups and 19F NMR spectroscopy (64), FT-IR (65),
MALDI-TOF MS (66), and colimetric assays have been de-
vloped (67).

Solid-phase strategies for glycan synthesis can be broken
down into three types: donor bound, acceptor bound, and bidi-
rectional (Fig. 4). In the donor bound mode, an excess acceptor
is used to drive the reaction to completion (or to maximize
yields). As the donor in a typical glycosylation reaction is the
“reactive intermediate,” unproductive side reactions correspond
to a direct loss in overall yield as the material is bound to the
solid support. The acceptor bound approach, which is the most
popular of the three, uses an excess of donor in solution. This
parallels Merrifield’s strategy for peptide synthesis in which
the reactive species is in solution (and in excess); hence, any
undesired side reactions along with unreacted donor are sim-
ply washed away. The bidirectional strategy represents a hybrid
of the two approaches and is well suited for the synthesis of
branched glycans (68).

With the success of solid-phase methods for peptide synthe-
sis, early studies of solid-phase oligosaccharide syntheses were
launched in the 1970s with the pioneering work of F Rechert
and Schurig (69) who prepared several disaccharides and triac-
tehrides using Merrifield’s resin and glycosyl bromides. Van
Boom and colleagues (70) reported on the solid-phase synthesis
of a heptaglycosylfuranoside in 1987 using Merrifield’s resin and
glycosyl chlorides in an impressive 23% overall yield. Both ap-
proaches used an acceptor bound strategy. The field underwent
a renaissance in the 1990s as powerful new glycosylation meth-
ods and protecting group strategies for oligosaccharide synthesis
emerged (71).

Danzhafsky and coworkers (72) have used the glycal as-
sembly method in a donor bound approach for the solid-phase
synthesis of various glycans, including among others the Lewis
B hexasaccharide, a blood-group determinant that has been iden-
tified as a mediator in the binding of pathogen Helicobacter
pylori to human gastric epithelium and is implicated in the
onset of peptic ulcers (73). A attachment of a glycal monosac-
charide to a polyethylene resin via a silyl linker was followed by
epoxidation with dimethyldioxirane (DMDO) and reaction with
a glycal acceptor in the presence of ZnCl₂. Protection (ester-
ification) of the newly formed C2 hydroxyl (or coupling with
another glycosyl donor, which is a unique advantage of the gly-
cal assembly method) and subsequent iteration of this process
leads to a resin-bound, fully protected glycan. Cleavage or re-
moval of the material from the resin and final protecting group
removal steps furnishes the desired oligosaccharide. To address
the issue of slower reaction rates on the solid-support as com-
pared with solution-phase, Krepinski et al. (74) used soluble
polyethylene glycol (PEG) polymer in an acceptor-bound ap-
proach with glycosyl bromides. The PEG strategy was used by
van Boom et al. (75) to prepare the heptasaccharide phytoalexin
elicitor (HPE).

As the solid-phase paradigm began to attract interest, a host
of research groups began transferring their respective glycosy-
lization technologies onto the solid-support, with each using
the acceptor-bound strategy. K. Ahmed et al. (76) prepared a triac-
tehride on Merrifield’s resin with glycosyl sulfonates and a
thiophenyl linker, and they demonstrated that sulfonides could
be used to install stereoselectively α-fucosidic linkages. The
power of this technology was demonstrated in the combinational
synthesis of a library of 1300 disaccharides and triacchorides
(77). Rademann and Schmidt have shown that trichloroacetim-
ides, which is one of the most powerful and popular donors
to date, have been effective in the solid-phase paradigm (78).
Schmidt et al. have introduced a variety of linker systems (79)
and have used different supports over the past decade (80).
Recently, they disclosed the synthesis of a library of N-glycans

Key Strategies for Glycan Synthesis

Fig. 4
Key Strategies for Glycan Synthesis

Solid-Phase Strategy:

a) Donor Bound:

b) Acceptor Bound:

c) Bidirectional

Figure 4 Donor-bound, acceptor-bound, and bidirectional solid-phase strategies.

(81). Nicolaou et al. (82) used phenolic polystyrene as a solid support to synthesize several oligosaccharides. Thioglycosides were employed as donors as well as a photolabile linker for the synthesis of HPE and a protected dodecasaccharide related to the phytoalexin elicitor family with a block strategy. Fraser-Reid et al. (83) translated the NPG method to the solid support (polystyrene) functionalized with a photolabile linker. A trisaccharide was prepared in which the chloroacetyl (ClAc) was used as the temporary protecting group between couplings. Seeberger et al. (84) have leveraged glycals for the rapid preparation and utility of glycosyl phosphates, which is a powerful class of glycosyl donors. A series of glycans has been prepared with this method (85). In 1999, Seeberger et al. (86) reported on the development and application of a 4(3)-1,8-octenediol linker, which was used in the preparation of several linear oligosaccharides. Cleavage from Merrifield’s resin with Grubbs et al.’s first-generation ruthenium catalyst (87) under an atmosphere of ethylene performed a cross metathesis reaction to yield an NPG, which can be used for additional glycosylation in solution or can be modified to access neoglycoconjugates (88).

The development of an automated oligosaccharide synthesizer would be highly beneficial to glycobiologists and would serve to drive glycomics, as the synthesis of glycans has been largely restricted to specialized groups. It would allow for routine assembly of desired targets and would accelerate the pace of discovery in glycoscience. Toward this end, Seeberger and coworkers (89) used their solid-phase methodology and disclosed the first automated oligosaccharide synthesizer in 2001. A modified ABI peptide synthesizer was used to prepare a protected heptasaccharide (HPE) and a protected dodecasaccharide using monosaccharide and disaccharide phosphate donors. The synthesis of the latter was accomplished in 16 hours after purification by high-performance liquid chromatography. The synthesizer has also been used to prepare glycans related to the branched Leishmania cap tetrasaccharide (90), a synthetic anti-toxin malaria vaccine (91), a core N-linked pentasaccharide that is common to all N-linked glycoproteins (92), the Lewis X pentasaccharide, the Lewis Y hexasaccharide, and the Le\(^\text{a}\)-Le\(^\text{b}\) nonasaccharide (93). Glycosyl phosphates and trichloroacetimidates have emerged as privileged donors in this regime.
Key Strategies for Glycan Synthesis

Enzymatic Synthesis

Nature has evolved carbohydrate-processing enzymes for the efficient assembly of glycans. Glycosyltransferases and glycosidases perform highly regioselective and stereoselective glycosylations, which thus obviates the need for tedious protecting group manipulation and controls the stereospecific installation of glycosidic linkages (two key issues in chemical glycan synthesis) (94). Moreover, the reactions are carried out under mild and green (nontoxic, environmentally friendly) conditions. As such, a growing trend has occurred in the synthetic carbohydrate community to leverage these enzymes in order to streamline glycan synthesis (95). These reactions are particularly useful for difficult chemical glycosylations (e.g., installation of sialic acid residues).

Glycosyltransferases

The biosynthesis of oligosaccharides is mediated by glycosyltransferases, which transfer either monosaccharide nucleotide monophosphates (e.g., CMP-Neu5Ac) or monosaccharide nucleotide diphosphates (e.g., Glc-UDP) to acceptors with either retention or inversion at the donor’s anomeric center (Fig. 5). These enzymes have evolved to be highly regiospecific and stereospecific. The high cost and availability of both enzyme and substrate are major drawbacks of the enzymatic approach. In addition, nucleotide diphosphate generated during the enzymatic glycosylation inhibits the enzyme. The issue of substrate cost has been addressed with elegant nucleotide donor recycling strategies introduced by Wong et al. (96), which have been translated to other glycosyltransferase systems (97). Feedback inhibition has been tackled by including a phosphatase (98). Wong et al.’s (99) synthesis of sialyl Lewis X, which is a glycan involved in the inflammation cascade, is a testament to the power of this strategy. The enzymatic glycosylation of sialic acid has found widespread use in glycan synthesis, as the chemical installation of this monosaccharide is difficult. Many glycans have been prepared enzymatically, including one-pot and solid-phase enzymatic approaches (100).

Glycosidases

During glycoprotein synthesis, glycosidases are involved in processing the glycans via hydrolysis of glycosidic linkages. These enzymes have been used for in vitro glycan synthesis under the appropriate conditions and make use of readily available, inexpensive donors such as nitrophenyl glycosides (Fig. 5). In addition, glycosidases are more stable than glycosyltransferases and more compatible with organic solvents. Although highly stereospecific, glycosidase are not regiospecific and hence result in a lower yield of desired oligosaccharide. Nevertheless, these highly useful biocatalysts have been employed in the synthesis of various glycans (101). Finally, a considerable amount of effort has gone into the engineering of novel glycosyltransferases and glycosidases that feature desirable characteristics (e.g., thermostability) as well as in preparing novel structures (e.g., thioglycosides). These strategies will certainly strengthen the ability to prepare more efficiently complex structures enzymatically (102).

Other strategies

To address issues of large-scale oligosaccharide synthesis, Koizumi et al. (103) at Kyowa Hakko (Kogyo Co. Ltd.) disclosed the utility of multiple metabolically engineered microorganisms containing all of the necessary genes for nucleotide generation and glycan synthesis; moreover, inexpensive orotic acid was used as a UTP precursor. The strategy resulted in a highly efficient synthesis of the globotriose epitope. Wang and coworkers (104) have also addressed the sugar nucleotide issue with the development of “superbeads,” in which enzymes involved in the glyconjugate biosynthetic pathway are expressed in recombinant Escherichia coli strains, isolated, and immobilized on an agarose resin. Additional glycosyltransferases could be used in solution as well. In 2002, Wang et al. (105)

![Figure 5](image-url)  
Glycosyltransferases and glycosidases in enzymatic glycan synthesis.
engineered an E. coli strain, which they termed "superbug," containing all of the biosynthetic genes necessary for the synthesis of α-Gal epitopes. This process obviates issues associated with isolating the individual enzymes (as in the superhead approach). The "superbug" technology was used in the large-scale synthesis of globotriose trisaccharide Gb3 (106).

Finally, the hybrid chemical and enzymatic approach (chemoenzymatic strategy) has also been popular for the synthesis of many glycans (107). The chemical approach with its drawbacks still allows for greater flexibility in the synthetic scheme (e.g., carbohydrate-based drugs and unnatural glycans). Moreover, relatively straightforward components of the glycan can be assembled readily with chemical synthesis, which leaves the difficult glycosylations (sterically and/or stereochemically demanding) to be carried out enzymatically. A highly illustrative example of this strategy is Unverzagt's chemoenzymatic synthesis of dodecasaccharide (108). Recently, Chen and co-workers (109) reported on an efficient one-pot, three-enzyme delivery of the target dodecasaccharide (108). Recently, Chen and co-workers (109) reported on an efficient one-pot, three-enzyme chemoenzymatic synthesis of various sialic acid-containing trisaccharides.

Future Outlook

The biologic significance of glycans and their potential as therapeutic agents has energized the field of glycochemistry. The prognosis for the development and implementation of novel strategies for the synthesis of glycans is very good as no general solution has been uncovered. Discoveries made in the chemical and enzymatic/biologic arenas will continue to "raise the bar" as the need to procure glycans for study persists.

References

Key Strategies for Glycan Synthesis


Key Strategies for Glycan Synthesis


Key Strategies for Glycan Synthesis


Further Reading


See Also

Glycans in Information Storage and Transfer
Glycan Synthesis, Key Reactions of
Glycan Synthesis, Protection and Deprotection Steps of
Glycomics, Major Techniques in
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

Robert Bittman, Queens College of The City University of New York, Flushing, New York
doi: 10.1002/9780470048672.wecb576

Sphingolipids are a highly diverse class of lipids, which encompasses thousands of different structural combinations of the polar head group, amino group or amide chain, and sphingoid long-chain base. Initially regarded as inert structural components of eukaryotic cell membranes and plasma lipoproteins, sphingolipids now are recognized as key participants in the life of virtually all cell types. They influence a myriad of biological processes, including cell growth, cell death, and susceptibility to inflammation and infections. The number of known sphingolipid-dependent cell signaling systems is increasing rapidly as research efforts in this field progress. Sphingolipids stabilize transient microdomains in membranes known as “rafts.” Signaling proteins and receptors partition into membrane rafts, and some toxins, pathogen, and viruses bind to sphingolipids that are concentrated in rafts and subsequently undergo endocytosis (see also Extracellular Lipid Signals and Lipid Rafts). Synthetic, non-natural analogs of sphingolipids hold promise as drug candidates for treatment of various disease states via their ability to perturb cell signaling, membrane trafficking, and binding of infectious agents. The repertoire of synthetic sphingolipids includes analogs that target enzymes in the sphingolipid biosynthetic pathway and those that act as antiproliferative agents, immunomodulators, and regulators of lipid raft formation (and resultant receptor function). This review is intended to bring attention to selected synthetic analogs of sphingosine, ceramide, and glycosphingolipids that may lead to therapeutic intervention in certain pathophysiologic conditions.

Chemical Structures of Sphingoid Long-Chain Bases and Their Glycosylated and Phosphorylated Derivatives

Sphingolipids contain a sphingoid base backbone, which often is referred to as the long-chain base. The long-chain base is an...
Sphingolipids with modifications at C-1 and C-2

Many examples of sphingolipids exist in which the C2-amino and/or terminal hydroxy group is modified. N,N-Dimethyl-D-erythro-sphingosine is a naturally occurring sphingolipid that, like D-erythro-sphingosine, inhibits the ubiquitous signal transducer protein kinase C (PKC) (11). At high concentrations (10µM), N,N-dimethylsphingosine also inhibits sphingosine kinases, the enzyme that is responsible for the formation of sphingosine 1-phosphate (S1P), but at low concentrations (<1 µM) it activates sphingosine kinase in the cytosol via a PKC-dependent mechanism (12).

The C-1 hydroxy group of sphingosine is replaced by a hydrogen in 1-deoxy-D-erythro-sphingosine (25,3R)-2-amino-3-decanol. Because no functional group is present at the terminal carbon, chemical/biochemical modification at that site is precluded. This naturally occurring sphingosine analog, which is called spisulosine or ES-285, was isolated from the clam Macromeris polyacantha and was found to induce an atypical form of apoptosis in murine leukemia cells (13). A cyclodextrin-based formulation of spisulosine is being used in Phase I clinical trials as an anticancer agent (14).

Sphingolipids with a sulfate group esterified to the C-1 position were isolated from marine invertebrates and shown to possess various bioactivities, such as the inhibition of neuraminidase and telomerase activity (15).}

Ceramides are N-acyl-D-erythro-sphingosines (compound 1, Fig. 2). Thus, ceramide is a 2-amido-1,3-diol. The fatty amide group in ceramide is structurally heterogeneous, differing in the length and extent of saturation and hydroxylation of the alkyl chain. In some species the fatty amide bears an α-hydroxy group. Typical fatty acids that are prevalent in the fatty amide group in ceramide is structurally heterogeneous, differing in the length and extent of saturation and hydroxylation of the aliphatic chain.

Sphingomyelin is a naturally occurring sphingolipid that, like D-erythro-sphingosine, inhibits the ubiquitous signal transducer protein kinase C (PKC) (11). At high concentrations (10µM), N,N-dimethylsphingosine also inhibits sphingosine kinases, the enzyme that is responsible for the formation of sphingosine 1-phosphate (S1P), but at low concentrations (<1 µM) it activates sphingosine kinase in the cytosol via a PKC-dependent mechanism (12).

The C-1 hydroxy group of sphingosine is replaced by a hydrogen in 1-deoxy-D-erythro-sphingosine (25,3R)-2-amino-3-decanol. Because no functional group is present at the terminal carbon, chemical/biochemical modification at that site is precluded. This naturally occurring sphingosine analog, which is called spisulosine or ES-285, was isolated from the clam Macromeris polyacantha and was found to induce an atypical form of apoptosis in murine leukemia cells (13). A cyclodextrin-based formulation of spisulosine is being used in Phase I clinical trials as an anticancer agent (14).

Sphingolipids with a sulfate group esterified to the C-1 position were isolated from marine invertebrates and shown to possess various bioactivities, such as the inhibition of neuraminidase and telomerase activity (15).
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

Figure 1 The structural feature common to all sphingolipids is a long-chain sphingoid base. Sphingoid bases are 2-amino-1,3-diols that bear an aliphatic chain. The chain length varies from about 14 to 24 carbon atoms, and the extent of unsaturation, hydroxylation, and methyl-branching in the chain also varies. 1, (2S,3S,4R)-sphingosine; 2, (2S,3R)-dihydrosphingosine; 3, (2S,3S,4S)-phytosphingosine; 4, 4E,6E-sphingadiene.

Figure 2 The structures of ceramide: 1, a schematic structure of ceramide that illustrates the sites at which structural modifications may be introduced by chemical synthesis, and the structure of a representative ceramide found in human skin (2). Structural changes in the ceramide molecule have been introduced at many sites. Alterations include the configuration at C-2 and C-3, the lengths of the fatty amide chain and of the aliphatic chain attached to C-5, the positions of the unsaturation and the secondary hydroxy group, the replacement of the hydroxyl groups with other atoms (hydrogen or fluorine) or functionalities (methoxy, methylthio, and keto), the incorporation of aromatic, heteroaromatic, and other rings in place of the alkenyl side chain of the sphingoid base, and the replacement of the carbonyl group of the carboxamide group. In addition, the 2-amino-1,3-diol functionalities have been incorporated into cyclic structures.

Other examples of sphingolipids with modifications at C-1 are glycosylated and the following phosphorylated derivatives.

Phosphorylated sphingolipids

Sphingomyelin (ceramide 1-phosphocholine, compound 1, Fig. 3) is a phospholipid in which the terminal hydroxyl group of ceramide is esterified to phosphocholine. Sphingomyelin and the glycerocephospholipid phosphatidylycholine have the same polar head group; together, the two choline-containing phospholipids account for >50% of the total phospholipid content of many mammalian membranes. Sphingomyelin is one of the most abundant phospholipids in the plasma membranes of mammalian cells and plasma lipoproteins, and its metabolism...
provides a source of sphingolipid second messengers. Sphingomyelin is a typical example of the stability of sphingomyelin in the formation of ceramide and phosphocholine (Fig. 3). Hydrolysis by sphingomyelinase catalyzes the phosphodiester bond, and the amine group transfers the phosphocholine to the lipid mediator lysophosphatidic acid. Autotaxin (ATX), an enzyme that catalyzes the hydrolysis of lysophosphatidic acid, is subject to product inhibition by two important lysophospholipid signaling molecules, lysophosphatidic acid and sphingosine-1-phosphate. Therefore, this enzyme is an attractive drug target that may be inhibited by synthetic lysosphingo- and/or glycosphingolipids.

In mammalian cells, the predominant long-chain base in sphingomyelin is D-erythro-C18-sphingosine. Dihydrosphingosine is a minor long-chain base of sphingomyelin of most mammalian membranes; for example, the dihydrosphingosine content of egg (chicken) and milk (bovine) sphingomyelin is 1% and 15-20%, respectively (23). An exception, however, is the adult human eye lens membrane, in which dihydro sphingosine comprises about 50% of the total phospholipid and about 80% of the total sphingomyelin content (24). The absence of unsaturation in the long-chain base results in modified physical properties. The double bond of sphingomyelin is essential in maintaining the transbilayer asymmetry of a glycosphingolipid in phospholipid vesicles (25) and in inhibiting protein-mediated glycosphingolipid transfer between membranes (26), as demonstrated by replacing sphingomyelin with dihydro sphingomyelin in bilayer membranes.

The C-3-hydroxy group and the C-4,C-5-double bond of sphingomyelin and other sphingolipids are localized in the lipid-water interfacial region. These groups play a role in organizing interfacial water via hydrogen bonding with the surrounding water molecules and neighboring lipids. Intramolecular hydrogen bonding (with the participation of strongly bound water molecules) between the C-3-hydroxy group of the long-chain base and the bridging oxygen of the phosphate results in restricted polar head group mobility (27). The trans-double bond of the sphingoid base plays a role in the extent of hydration of the interfacial region of sphingomyelin-containing bilayers. Dihydrosphingomyelin has been proposed to promote stronger intermolecular hydrogen bonds with neighboring dihydrosphingomyelin, cholesterol, and other lipid molecules than does sphingomyelin (28), whereas sphingomyelin may form more stable intermolecular hydrogen bonds (27). Dihydrosphingomyelin has a higher affinity for cholesterol than sphingomyelin and protects cholesterol and unsaturated phosphatidylcholines from oxidation better than sphingomyelin (29).

A heterogeneous mixture of fatty amide chains is found in natural sphingomyelins and other sphingolipids, varying with the source and age (30). The fatty acyl chain content of milk sphingomyelin, for example, is 33% C23:0, and 20% each of C16:0, C22:0, and C24:0. About 50% of the N-acyl chain composition of bovine brain sphingomyelin is C18:0, with about 20% C24:1 and 8% C22:0. Egg sphingomyelin is an exception, being relatively homogeneous with respect to N-acyl content—largely C16:0 (80–85%) and C18:0 (12%). An abundance of long, saturated amide chains exists in sphingomyelin compared with the fatty acyl chains of glycerophospholipids, which provides a marked disparity in the intramolecular chain lengths of the sphingoid chain and the N-acetyl chain. A long, saturated N-acetyl chain in sphingomyelin is conducive to tight packing between sphingomyelin and cholesterol, which may result in the formation of ordered microdomains in bilayer membranes. Synthetic sphingomyelin bearing a N-oleyl chain is not incorporated into cholesterol-rich domains (31), probably because of its high miscibility with phosphatidylcholine (32).

The lengths of both the long-chain base and fatty amide chain tend to be shorter in the sphingolipids isolated from invertebrate sources than from mammalian sources. For example, the major long-chain base of the sphingolipids of Drosophila (33), the moth Manduca sexta (6), and honey bees (34) is C14-D-erythro-sphingosine (the 14-carbon homolog of the long-chain base, with smaller amounts of the C15-C17 sphingoid bases), and the N-acetyl chains of their sphingolipids are also shorter (35). Incidentally, the fatty acyl chains of their glycosphingolipids are also shorter than those of glycosphospholipids in mammalian cell membranes; thus the temperature of the main gel to liquid-crystalline phase transition is lower in invertebrate membranes than in mammalian membranes.

Invertebrates also contain ceramide phosphoethanolamine, an analog of sphingomyelin in which the choline is replaced by ethanolamine, as the major sphingolipid of Drosophila and other invertebrates. Ceramide phosphoethanolamine seems to have an antioxidant function in Drosophila (33) similar to sphingomyelin in mammalian cells (36, 37). Little information is available regarding the other functions of ceramide phosphoethanolamine because most physical studies of sphingophospholipids have been focused on sphingomyelin. Sphingomyelin isolated from natural sources partitions efficiently into highly ordered lateral domains of lipid bilayers enriched in cholesterol and interacts preferentially with cholesterol (38), but ceramide phosphoethanolamine seems to form less tightly packed bilayers with cholesterol (39).

Sphingophospholipids are another type of sphingophospholipid in which the polar head group contains a carbon-phosphorus bond. Ceramide 2-aminoethylphosphonate is a typical example (compound 2, Fig. 3). In some species, the head group also bears an N-acryloxycarboxylic acid or an N-methyl group. Sphingophosphonolipids are minor constituents of mammalian cell membranes but are abundant in the bacterium Bacteriovorax stolpii, in which the long-chain fatty acids are mostly C17 iso-branched phytoceramide and dihydrosphingosine backbone and an inositol phosphate head group. Inositol phospholipids are formed in a reaction catalyzed by InsPcCer, compound 3, Fig. 3, which has a phytoceramide or dihydrosphingosine backbone and an inositol phosphate head group. InsPcCer is subject to product inhibition by two important lysophospholipid signaling molecules, lysophosphatidic acid and sphingosine-1-phosphate. Therefore, this enzyme is an attractive drug target that may be inhibited by synthetic lysosphingoi-and/or glycosphingolipids.
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

Figure 3: The structures of sphingomyelin (1), ceramide aminoethylphosphonate (2), and inositol phosphorylceramide (InsPCer) (3).

PCer synthase, which transfers inositol phosphate from phosphatidylinositol to the C-1 hydroxy group of dihydroceramide or phytoceramide and therefore liberates diacylglycerol (42). Glycosylated derivatives of InsPCer constitute the lipid moiety of most glycosylphosphatidylinositol-(GPI)-anchored membrane proteins of eukaryotic cell membranes (43, 44). As no mammalian analog of InsPCer synthase exists, this enzyme is a potential therapeutic target for the development of selective antifungal and antiprotozoal agents that interfere with sphingolipid biosynthesis in fungal and protozoan cells but not in mammalian cells (45, 46). Microbial natural products that inhibit InsPCer synthase include the cyclic peptide aureobasidin A, an amphipathic lipid known as khafrefungin, and a macrolide called galbonolide A or rustmicin (46, 47). The structures of the latter two compounds are shown in Fig. 4.

The phosphorylated metabolites of sphingomyelin are sphingosylphosphocholine (lyso-sphingomyelin, the N-deacylated derivative of sphingomyelin), sphingosine 1-phosphate, and ceramide 1-phosphate. These sphingolipid metabolites are signaling molecules that regulate diverse cellular functions (48–53).

Glycosylated sphingolipids

The terminal hydroxy group of sphingosine is linked to at least one carbohydrate moiety in glycosphingolipids. In mammalian cells, the sugar moiety is in a β-glycosidic link to the sphingoid base, as in β-glucosylceramide (Fig. 5). Cerebrosides are one subgroup of glycosphingolipids in which a galactose or glucose is linked to ceramide (compound 1, Fig. 5). Sulfatides are a subgroup of galactocerebrosides in which a sulfate group is esterified to the C-3 position of the sugar; the sulfatide group of 3′-phosphodiestersilane 5′-phosphosulfate is transferred to the galactocerebroside in a reaction catalyzed by a sulfotransferase. β-Galactosylceramide and its 3′-sulfate ester bind to the gp-120 envelope protein of HIV-1; this interaction allows HIV-1 to infect CD4-negative cells (54). Globosides have at least two saccharide units linked by the anomeric hydroxy group of one sugar to the terminal hydroxy group of the sphingoid base. Plants and fungi contain inositol-linked phosphophytoceramides (InsPCer) such as compound 3 (Fig. 3), a major glycosphosphatidylcholine of yeast, and mannosylated and dimannosylated derivatives. Gangliosides (compound 2, Fig. 5) are ubiquitous components of vertebrate cells and are abundant in nervous tissues. They are localized in the outer leaflet of plasma membranes and form clusters in lipid microdomains, where they participate in various recognition and cell signaling processes (55). The oligosaccharide head group of gangliosides contains at least one N-acetylneuraminic acid residue. This monosaccharide, which is a sialic acid and is abbreviated as NANA, NeuAc, or Neu5Ac, is recognized (together with uncharged saccharide units) by many carbohydrate-related proteins (56). Thus gangliosides are sialylated glycosphingolipids. The predominant N-acyl chain of the ceramide backbone in gangliosides is stearoyl (C18:0). Figure 5 shows the structures of the gangliosides GM1 and GM3. In ganglioside nomenclature, the subscript letter “M” refers to “mono,” which is the number of NANA residues in GM1. The difference between 5 and the number after the “M” indicates the number of uncharged sugar residues present in the ganglioside: four in Ga2, and two in Ga3. Ga4 is the principal...
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

Galbonolide A (rustemicin), a 14-membered macrolide inhibitor of IPC synthase

Khafrefungin, an IPC synthase inhibitor with a C22 aliphatic chain esterified to polar head group (an aldonic acid)

Figure 4 The structures of natural inhibitors of InsP4Cer synthase: galbonolide A and khafrefungin.

1: Glycosphingolipids

β-glucosylceramide

a sugar head group (such as glucose) is linked via a β-glycosidic bond to C-1 of sphingosine

2: Gangliosides

an oligosaccharide head group with one or more sialic acid residues

Gangliosides affect a multitude of cellular processes, including binding specifically to viruses and many bacterial toxins. A well-known example is the binding of cholera toxin, the enterotoxin produced by *Vibrio cholerae*, to GM1 on the surface of human intestinal epithelial cells (57). The development of artificial receptors for the toxin may form the basis for the design of a potential anticholera drug (58). Another example of a specific interaction between a protein and a ganglioside is the binding of GM1 to α-synuclein, a presynaptic protein with a propensity to form fibrils and aggregates that have been implicated in Parkinson's disease and other neurodegenerative diseases (59). Ganglioside GM3 binds to insulin receptors; in fact, the accumulation of GM3 in membrane microdomains (rafts) of adipocytes results in the development of insulin resistance, apparently because of the dissociation of the insulin receptor from its complex.
with caveolin-1 in rafts (68). Because \( G_{\alpha 3} \) is a negative regulator of insulin signaling, the inhibition of \( G_{\alpha 3} \) biosynthesis may offer a novel therapeutic approach for insulin-resistant type 2 diabetes.

**Role of Chemical Structure in the Biological Activity or Biophysical Behavior of Ceramide**

The presence of two chiral carbon atoms in sphingosine results in the possible formation of four stereoisomers (Fig. 6). Only the D-erythro stereoisomer occurs naturally, and this stereoisomer is the only form that is used efficiently as the substrate for sphingolipid-using enzymes such as ceramidase (63, 62) and protein phosphatases (63). The "D" nomenclature is based on the Fischer projection and its relationship to the Fischer projection of D-glyceraldehyde (Fig. 6). Many studies have been carried out to assess how the various bioactivities of the unnatural stereoisomers of ceramides differ from that of D-erythro-ceramide. Short-chain ceramides (having a C2 to C8 \(-\)acyl chain) were used in many of these studies because they are water soluble and cell permeable. In general, the D-erythro form was found to be the most effective with regard to bioactivity. For example, only the D-erythro isomer of C8-acyl-ceramide was capable of promoting the fusion of an alpha virus with target liposomes; the other three isomers were inactive (64). However, in some studies the D-threo isomer had equal or higher activity than the D-erythro isomer (65), and in still other studies both the D-threo and D-threo isomers had higher activity than did the D-erythro or D-erythro isomer (66). In other studies, a lack of stereospecificity was found, as all four stereoisomers exhibited approximately equal activity (67-69).

Short-chain analogs of D-erythro-ceramide trigger apoptosis, thereby mimicking the activity of endogenous ceramide. However, biophysical studies indicate that short-chain ceramides perturb the physical properties of model membrane bilayers differently than their long-chain counterparts. For example, they did not form the ceramide-rich domains that are observed characteristically with C16:0- and C18:0-ceramide (70). A study of the ability of ceramide analogs to displace sterols from rafts indicated that C12:0- and C16:0-ceramides are faithful mimics of natural ceramides in model bilayers (71). The presence of the trans-double bond between C-4 and C-5 contributes to the close packing of D-erythro-ceramide molecules at the lipid-water interface (72). The unsaturation is required for most bioactivities of ceramide that have been studied, as dihydroceramide generally exhibits much lower activity than the natural ceramide. How- ever, biophysical studies indicate that short-chain ceramides can mimic the activity of endogenous ceramide. In the anabolic pathway of ceramide biosynthesis, N-acylation of a sphingoid long-chain base takes place in the endoplasmic reticulum and Golgi apparatus. In the catabolic pathway of ceramide biosynthesis, degradation of sphingomyelin and glycosphingolipids takes place at the plasma membrane and lysosomes, respectively. The enzymes in both the anabolic and catabolic pathways of sphingolipid metabolism are important determinants of cell survival, growth, and differentiation. Because sphingolipids have been implicated in a wide range of human diseases and neurological syndromes, sphingolipid-using enzymes represent potential targets for therapy of a variety of diseases characterized by an accumulation of sphingolipids or an improper balance among growth-supporting (such as sphingosine 1-phosphate, ceramide 1-phosphate, and glycosylceramides) and growth-inhibiting sphingolipids (such as ceramide and sphingosine).

A defect in a single lysosomal enzyme or cofactor results in substrate accumulation in one or more organs. This imbalance between the formation and breakdown of glycosphingolipids is associated with neurodegeneration and high mortality (see also Lipid Homeostasis, Chemistry of). Partial reduction in the rate of biosynthesis of the glycolipid that accumulates in a lipid storage disease is known as substrate reduction therapy, which is an alternative therapeutic approach to enzyme replacement. An inhibitor of an enzyme in an anabolic glycosphingolipid pathway is used to reduce a key glycosphingolipid partially to a level at which the residual catalytic activity is capable of preventing excessive substrate accumulation. The imino sugar N-butyldexosamine (NB-DNJ), which is marketed as miglustat or Zavesca, is a synthetic sphingolipid deriva- tive that selectively modulates cellular glycosphingolipid levels. It inhibits the ceramide-specific glycosyltransferase that catalyzes the first committed step in glycosphingolipid synthesis and shows promising results in the management of some an- imal models of glycosphingolipid storage diseases because it offsets the accumulation of glucosylceramide and more complex glycosphingolipids (82).

Dihydroceramide exhibits different physical and biochemical effects compared with the corresponding ceramide bearing the same \(-\)acyl chain. For example, channel formation in mitochondrial membranes (78) and protein phosphatase activity in pancreatic beta cells are inhibited by dihydroceramide but are activated by ceramide (63, 79). Dihydroceramide is less effective than ceramide in inducing apoptosis (68).

The terminal hydroxy group of ceramide has also been modified. 1-O-Methoxyceramide did not inhibit mitochondrial ceramide (69), and 1-fluroceramide was a weak inhibitor of glucosylceramide formation in cultured murine neuronal cells (80). The 1-methylthio analog of dihydroceramide activated sphingomyine kinase and disrupted neuron axonal growth in cell cultures (81).

**Biological Background**

In the anabolic pathway of ceramide biosynthesis, N-acylation of a sphingoid long-chain base takes place in the endoplasmic reticulum and Golgi apparatus. In the catabolic pathway of ceramide biosynthesis, degradation of sphingomyelin and glycosphingolipids takes place at the plasma membrane and lysosomes, respectively. The enzymes in both the anabolic and catabolic pathways of sphingolipid metabolism are important determinants of cell survival, growth, and differentiation. Because sphingolipids have been implicated in a wide range of human diseases and neurological syndromes, sphingolipid-using enzymes represent potential targets for therapy of a variety of diseases characterized by an accumulation of sphingolipids or an improper balance among growth-supporting (such as sphingosine 1-phosphate, ceramide 1-phosphate, and glycosylceramides) and growth-inhibiting sphingolipids (such as ceramide and sphingosine).
myriad of cell signaling events, a great deal of attention has been devoted to the study of ceramide metabolism and function in the past two decades. The direct targets of ceramides that have been identified are the protease cathepsin D, ceramide-activated protein phosphatases, ceramide-activated protein kinases, the kinase suppressor of Ras (KSR), isoforms of protein kinase C (PKC), c-Jun N-terminal kinase (JNK), Akt, and PTEN (83, 84). The phosphorylation state of proteins with pro- and antiapoptotic activities is modulated by intracellular ceramide levels.

**Regulation of the endogenous levels of ceramide with naturally occurring inhibitors of sphingolipid-using enzymes**

The structures of some naturally occurring compounds that act on sphingolipid-metabolizing enzymes are shown in Fig. 7. These compounds include 1) sphingols, which inhibit serine palmitoyl-CoA transferase, 2) fumonisin and australifungin, which inhibit sphingomyelinase, 3) scyphostatin, which inhibits neutral-pH optimum isomerase of sphingomyelinase, 4) 2-acetyl-4-tetrahydroxybutylimidazole (THI), which inhibits S1P lyase, and 5) aureobasidin A, kahrefungin, and gaibondide A, which inhibit lysIP cer synthase (Fig. 4).

**Overview of sphingolipid biosynthesis**

The de novo synthesis of ceramide takes place in the endoplasmic reticulum and begins with the decarboxylative condensation of L-serine with palmitoyl-CoA, catalyzed by serine palmitoyl-CoA transferase (SPTase) (Fig. 8). As the name of the enzyme implies, palmitoyl-coenzyme A is recognized preferentially by mammalian SPTase, and the palmitoyl chain is transferred to the amino group of L-serine with the formation of 3-ketosphinganine. In this reaction, carbon dioxide is released from serine, with the other two carbon atoms of serine providing the source of C-1 and C-2 of the long-chain base. Therefore, the 18-carbon long-chain base predominates in the sphingolipids of mammalian cells. Although most prokaryotic cells do not contain sphingolipids, some bacteria such as the Sphingomonas species synthesize glycosphingolipids. The SPTase of these bacteria is a cytoplasmic homodimer, in contrast to the membrane-bond heterodimeric form found in eukaryotes (85); as mentioned above their sphingoid long-chain bases tend to have fewer than 18 carbons, which implies that their SPTase prefers an acyl-CoA with fewer than 16 carbons in the acyl chain.

Inhibition of the first step in sphingolipid biosynthesis with sphingols, L-cycloserine, or p-chloro-L-alanine results in the depletion of all sphingolipids. Sphingolipids such as myriocin (also known as ISP-1, Fig. 9) and sphingofungin (Fig. 9) have been isolated from fungal cultures and often are used as biochemical tools to examine the roles of sphingolipids in vitro. Growth inhibition of cultured cells by myriocin and sphingofungin B was rescued by exogenous sphingolipids (86).

Myriocin treatment of apoE1 knockout mice decreased the sphingomyelin content of lipoproteins and lowered the levels of plasma cholesterol and triglycerides, which indicates that sphingoid biosynthesis may be a therapeutic target for treating dyslipidemia and atherosclerosis (87). Myriocin is a competitive inhibitor of SPTase, which forms an aldimine adduct with the active site of the enzyme. However, cycloserine and p-chloro-L-alanine inhibit SPTase by a different mechanism; they react with the coenzyme of SPTase, pyridoxal 5′-phosphate, inhibiting the enzyme irreversibly and thereby downregulating the biosynthesis of sphinganine (88). As shown in Fig. 9, myriocin and sphingofungin B (both of which inhibit SPTase in the nanomolar range) bear a structural resemblance to sphinganine and its biosynthetic intermediates and may also mimic the transition state of the SPTase-catalyzed reaction. The configurations at C-2, C-3, C-4, and C-5 of sphingofungin B are important for its inhibitory potency on SPTase; however, the 14-hydroxy epimer of sphingofungin B is as potent as the natural 14S stereoisomer, which indicates that the configuration at C-14 is not critical (89).

**Inhibition of other enzymes in the sphingolipid metabolic pathway by natural products**

In the second step of the sphingolipid biosynthetic pathway, the carbonyl group of 3-ketosphinganine is reduced with NADPH to form sphinganine (Fig. 9). Biological activities of sphinganine may be due to its ability to downregulate the biosynthesis of sphinganine (88). As shown in Fig. 9, sphingofungin B was rescued by exogenous sphingolipids (86). Growth inhibition of cultured cells by myriocin and sphingofungin B (both of which inhibit SPTase) was rescued by exogenous sphingolipids (86). Myriocin is a competitive inhibitor of SPTase, which forms an aldimine adduct with the active site of the enzyme. However, cycloserine and p-chloro-L-alanine inhibit SPTase by a different mechanism; they react with the coenzyme of SPTase, pyridoxal 5′-phosphate, inhibiting the enzyme irreversibly and thereby downregulating the biosynthesis of sphinganine (88). As shown in Fig. 9, myriocin and sphingofungin B (both of which inhibit SPTase in the nanomolar range) bear a structural resemblance to sphinganine and its biosynthetic intermediates and may also mimic the transition state of the SPTase-catalyzed reaction.
form D-erythro-sphinganine. No naturally occurring inhibitors of 3-ketosphinganine reductase have been reported yet.

The N-acylation of sphinganine with a fatty acyl-CoA is catalyzed by acyl-CoA-dependent ceramide synthases (also called dihydroceramide synthases; at least six genes for this enzyme exist in mammalian cells), which produces dihydroceramide. The fungal natural products fumonisin and australifungin (Fig. 7) are specific inhibitors of N-acylsphinganine transferase activity in mammalian and fungal cells, respectively. The structure of fumonisin resembles the structures of both the sphingosine base and fatty acyl-CoA cosubstrates of dihydroceramide synthase. Therefore, the aminopentol portion of fumonisin may compete with the long-chain base for binding to the enzyme, and the polar region that contains the two TCA substituents may block binding of the fatty acyl-CoA. Fumonisin-induced disruption of sphingolipid metabolism resulted in neurotoxicity, birth defects, cancer, and renal and liver failure (90, 91). γ-Tocopherol, the predominant form of dietary vitamin E, also raises the levels of dihydroceramide and dihydrophosphonosine in human prostate cancer cell lines (92). Although the mechanism by which these sphingolipids are elevated has not been established, this result is another example of how an interruption in the de novo sphingolipid pathway can culminate in cell death.

Dihydroceramide desaturase catalyzes the last enzyme in the de novo biosynthetic pathway, which is the introduction of the C-4,C-5-trans-double bond into the long-chain base. No naturally occurring inhibitors of this enzyme have been reported yet. In plants and fungi, dihydroceramide is hydroxylated at C-4 instead of undergoing dehydrogenation. The hydroxylated product, phytoceramide, is the precursor of complex sphingolipids such as Ins PCer and various glycosyl-InsPCer derivatives such as mannose-IPC, mannose diinositolphosphorylceramide, dimannose inositolphosphorylceramide, and galactose-dimannose inositolphosphorylceramide.

**Metabolism of ceramide**

Ceramide formed in mammalian systems is metabolized to sphingomyelin, glucosylceroglipids, and ceramide 1-phosphate by the actions of sphingomyelin synthase, glucosylceramide synthase, and ceramide kinase on ceramide, respectively. Ceramide synthesized in the endoplasmic reticulum is transported to the Golgi apparatus for synthesis of complex sphingolipids (93). Sphingomyelin is synthesized from ceramide by transfer of the
Figure 8: Outline of the sphingolipid biosynthetic pathway. Natural and synthetic inhibitors of sphingolipid-metabolizing enzymes are shown adjacent to the inverted T bars.

Figure 9: Structures of naturally occurring inhibitors of SPTase.

phosphocholine head group of phosphatidylcholine. As mentioned above, sphingomyelin synthase catalyzes this reaction, and diacylglycerol (an activator of PKC) is released. InsPCer synthase carries out the analogous process in yeast and plants, transferring inositol phosphate to phytoceramide or dihydroceramide. Cerebrosides and gangliosides are synthesized from ceramide and nucleotide sugars in the presence of glycosyl transferases.

Hydrolysis of the amide bond of ceramide by ceramidases generates sphingosine, which has broad antibacterial and antifungal activity in addition to its PKC inhibitory activity (94). Phosphorylation of sphingosine at the terminal hydroxy group by sphingosine kinases affords S1P. S1P lyase, a pyridoxal 5'-phosphate dependent enzyme that has a potential role in controlling cell fate and stress responses (96, 97), cleaves S1P between C-2 and C-3 to produce an α,β-unsaturated fatty aldehyde and phosphoethanolamine. Pyridoxal 5'-phosphate forms a Schiff base with the amino group of S1P. S1P lyase plays a role in determining the balance between the intracellular levels of S1P and ceramide.

S1P is dephosphorylated by S1P phosphatases. The S1P phosphatase-1 isozyme decreases the transport of ceramide from the endoplasmic reticulum to the Golgi, as determined by the visualization of a fluorescent analog of ceramide that accumulates in the ER (98). S1P phosphatase type-2 is upregulated by inflammatory stimuli (99).

Interrelationships Among Sphingolipid Metabolites

Ceramide also is formed by sphingomyelinase (SMase)-catalyzed hydrolysis of sphingomyelin. Acidic and neutral SMases are activated by stress stimuli and external agents such as the Fas ligand (FasL), tumor necrosis factor-α (TNF-α), growth factors, and chemotherapeutic agents, which produces a transient rise in intracellular ceramide levels by hydrolysis of the phosphocholine moiety of sphingomyelin. SMases play important roles in cell signaling pathways that regulate cell growth, differentiation, cell cycle arrest, and apoptosis. A secretory form of acidic SMase is important in sphingomyelin catabolism, generating ceramide from sphingomyelin in plasma membrane domains. The resulting coalescence of ceramide-rich domains leads to the initiation of apoptosis, internalization of pathogens, and secretion of cytokines (100, 101).

The fungal metabolite scyphostatin (Fig. 7) is a potent, reversible inhibitor of membrane-bound Mg2+-dependent neutral sphingomyelinase (nSMase), thereby interfering with the generation of ceramide (102, 103). The activity of nSMases is sensitive to the cellular redox state, for example, the ratio of glutathione in the reduced versus the oxidized form (104, 105). The farnesyltransferase inhibitor manumycin A, a polyenamide produced by the Streptomyces species, also inhibits nSMase irreversibly (106).

The lysosomal acid sphingomyelinase (aSMase) isoform is inhibited by L-carnitine, a cofactor of acyl-coenzyme A transport in mitochondria (107, 108), and by phosphatidylinositol polyphosphates, which occur in plant, yeast, and mammalian cells (109).

A caramel food colorant and component of coffee, 2-acetyl-4-tetrahydroxybutylimidazole (THI, Fig. 7), inhibits S1P lyase, which causes S1P to accumulate in the lymph nodes (110).
by glycosylceramide synthases; it is converted to ceramide 1-phosphate by ceramide kinase and to sphingosine by ceramidases (Fig. 10a). The interplay between the kinases and phosphatases may control the concentrations of these sphingolipid mediators. Figure 10b shows the interconversions among the sphingolipid metabolites, together with the natural and synthetic inhibitors of enzymes in the sphingolipid pathway.

Elevating the ceramide content of cells

Stimulating ceramide biosynthesis and inhibiting the conversion of endogenous ceramide to other sphingolipid metabolites is a strategy that may be applied to block key steps in cancer progression such as cell growth and cell survival. The intracellular level of ceramide can be elevated by 1) activating neutral...
and acid SMases, 2) stimulating the de novo synthesis of ceramide, and 3) blocking the metabolic conversion of ceramide to other sphingolipids (for example, by inhibiting enzymes that use ceramide as a substrate, such as sphingomyelin synthase, glycosylceramide synthases, ceramidases, ceramide kinase, and O-acyltransferases).

Lowering the ceramide content of cells
Inhibitors of the enzymes that participate in the de novo biosynthesis of ceramide may be used to treat pathologies associated with elevated intracellular ceramide levels, such as chemotherapy-induced cell death and insulin resistance in muscle.

Designing Sphingolipid Analogs to Modulate Sphingolipid Metabolism
Synthetic analogs of sphingolipids offer advantages over natural sphingolipids

1. The extensive hydrophobicity of some natural compounds can be reduced by synthesizing compounds with shorter hydrocarbon chains or with more polar groups, thereby enhancing the aqueous solubility and cellular uptake (which may be directed to different intracellular compartments). Many cell-permeable analogs of ceramide mimic the effects of TNF-α, chemotherapeutic agents, oxidants, and ionizing radiation, which activate SMases to produce ceramide. Moreover, the subcellular localization of the analog may be dependent on the lipophilicity and/or net charge of the analog. N-Hexanoyl- and N-octanoyl-D-erythro-ceramide are examples of cell-permeable analogs of natural ceramide that induce cell cycle arrest or apoptosis in A549 cells (111, 112).

2. As discussed above, altering the stereochemistry of the sphingosine backbone has produced, in many instances, stereoisomers with altered bioactivity compared with the natural stereoisomer, for example, in activating the catalytic subunit of serine-threonine protein phosphatase 2A (63) or in inhibiting mitochondrial ceramidase (69).

3. The chemical and metabolic stability of labile natural compounds can be enhanced by incorporating stabilizing groups such as a phosphonate in place of a phosphodiester bond or a C-glycosidic bond, thereby creating new agents that resist phosphatase (113–115) and glycosidase action in cells (116).

4. The toxicity and bioavailability can be varied by conjugation to form new derivatives such as glucuronides that are metabolized differently than the parent compound. An example is a glucuronide derivative of the aminophanolamide of all-trans retinoic acid, fenretinamide (N-(4-hydroxyphenyl)-retinamide, 4-MPR) (117).

5. Structure-activity relationships may be established if several analogs are available. This step is a necessary step in drug design that leads to optimizing the properties of the therapeutic agent and developing a lead compound.

Many non-natural sphingolipid analogs have been synthesized with the aim of achieving a potential therapeutic advantage with respect to one of the following bioactivities: 1) to manipulate the activity of enzymes in the sphingolipid biosynthetic pathway and thus alter the balance between the pro-survival and antiproliferative sphingolipid metabolites, 2) to alter the stability of lipid nets and cause aberrant localization of signaling molecules, 3) to produce immunosuppression via alteration of the number of circulating lymphocytes, or 4) to activate natural killer T cells to produce a desired array of cytokines.

Elevating Intracellular Ceramide Levels with Unnatural Sphingolipid Analogs
Irradiation (UV and γ rays), chemotherapy, pathogenic infections, and many other external stress stimuli activate SMases and elevate the endogenous levels of ceramide in tumor cells, which promotes apoptosis. However, tumor cells may counter these treatments by activation of ATP-dependent efflux pumps and by converting endogenous ceramide to other sphingolipid metabolites, which thereby evades apoptosis.

Use of unnatural sphingolipid analogs to elevate endogenous ceramide levels in tumor cells by interfering with ceramide trafficking
After the de novo synthesis of ceramide in the endoplasmic reticulum, additional biosynthesis of sphingolipids continues in the Golgi apparatus and plasma membrane. Sphingomyelin synthase uses ceramide as a substrate for sphingomyelin production in the lumen of the Golgi. A cytosolic ceramide transfer protein, CERT, transports ceramide in a nonvesicular manner from the endoplasmic reticulum to the Golgi complex (118). Ceramide transport is required for the synthesis of sphingomyelin and presumably other sphingolipids. A Drosophila mutant that lacked the functional CERT gene had a depressed content of ceramide phosphoethanolamine (the sphingomyelin analog in Drosophila) and ceramide, which resulted in altered membrane trafficking.
permeability behavior and an enhanced susceptibility to oxidation of cellular components (13). A synthetic analog that resembles ceramide with respect to structure and stereochemistry, (1R,3R)-N-(3-hydroxy-1-hydroxymethyl-3-phenylpropyl)dodecamide (1R,3R)-HPA-12 (Fig. 11), inhibited sphingomyelin biosynthesis by blocking ceramide trafficking (119).

Use of synthetic sphingolipid analogs to modulate endogenous ceramide levels

1. Inhibition of neutral SMases. Mg²⁺-dependent nSMases in plasma membranes hydrolyze sphingomyelin to yield ceramide and phosphocholine. As ceramide is a key signaling molecule in the apoptotic and inflammation response to stress signals, inhibitors of SMase isoforms may provide lead compounds for the treatment of inflammation, ischemia, neurodegenerative diseases, and infarction. The following examples are representative of efforts to develop synthetic sphingophospholipid analogs with antiapoptotic activity.

   - 3-O-Methyl- and 3-O-ethyl-sphingomyelin (Fig. 11) inhibited nSMase without markedly affecting the activity of the acid isoform (120). Inhibition of nSMase also was observed with hydrolytically stabilized tert-butylcarbamate and urea derivatives of sphingomyelin, which also prevented apoptotic neuronal cell death in an ischemic model (121, 122). S1P analogs with a difluoromethylene phosphonate link instead of the phosphate group (SMA-3 and SMA-7, Fig. 12) inhibited nSMase in pheochromocytoma PC-12 cells and inhibited cerebral infarct in mice (123). Phosphonocho line analogs of sphingomyelin, in which an oxygen atom in the phosphate ester head group was replaced by a carbon, nitrogen, or sulfur atom, were also inhibitors of nSMase (124) as was a lactone derivative of ceramide (125) (Fig. 11).

   - Inhibition of SM synthase. Tricyclodecan-9-yl xanthogeneate (D609, Fig. 12) is a known inhibitor of phospholipase C (which hydrolyzes phosphatidylcholine) (127). It inhibited SM synthase and induced apoptosis in U937 human monocyte leukemia cells (128) and rat PC12 cells (129).

   - Inhibition of glucosylceramide (GlcCer) synthase. This enzyme incorporates a glucosyl residue from UDP-glucose into ceramide to form β-glucosylceramide, which is a precursor of complex glycosphingolipids that participate in many physiologic and pathophysiologic processes. Therefore, this enzyme is a potential drug target. NB-DNJ (Fig. 12) is a competitive inhibitor of GlcCer synthase with respect to ceramide and is a noncompetitive inhibitor with respect to UDP-glucose; molecular modeling studies (indicated that NB-DNJ) is structurally similar to ceramide but not to glucose (130). NB-DNJ reduced the level of GlcCer that accumulates in Gaucher disease and also acted as a chemical chaperone for the acid β-glucosidase that is defective in this disease. Other competitive inhibitors of GlcCer synthase are ceramide analogs with a cyclic amino head group such as morpholine or pyrrolidine instead of the primary hydroxy group at C-1. They also have a phenyl group in place of the aliphatic chain of the sphingosine backbone. A lead compound that was developed to inhibit GlcCer synthase is (1R,2R)-1-phenyl-2-amino decanoyl-3-morpholino-1-propanol (D-threo-PDMP, Fig. 12), which has an...
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

N-decanoyl amide chain (131). This ceramide analog leads to the accumulation of ceramide in neuroblastoma and other cell types and also possesses antitumor activity (132). The inhibitory potency on GlcCer synthase was enhanced by the elongation of the N-acyl group from C10 to C16, by the introduction of electron-rich aromatic substituents, by the replacement of the morpholino with a pyrrolidino head group, and by the addition of water-soluble links. In addition to inhibiting the activity of GlcCer synthase, D-threo-PDMP exerts several other effects, including the alteration of the structure and function of membrane domains in late endosomes, the inhibition of LDL degradation, and the enhanced uptake of paclitaxel (133). Administration of high doses of L-PDMP increases GlcCer synthesis in the brain (134).

4. Inhibition of ceramidases. Ceramidases specifically hydrolyze the amide bond of ceramide to form sphingosine and a fatty acid without acting on the amide bond of other sphingolipids. Enzymatic forms of ceramidase with acidic, neutral, and alkaline activities have been studied. The enzyme with an acidic pH optimum is localized primarily in lysosomes and thus is important in catabolism. It is activated by the glycoprotein sphingolipid activator protein D (SAP-D) (135, 136).

N-Oleoyl ethanolamine (Fig. 12), the first known inhibitor of this enzyme, has a low potency (IC₅₀ > 0.5 mM) and a low specificity (137, 138). Point mutations in the gene encoding acid ceramidase lead to the lysosomal storage disease known as Farber disease. As acid ceramidase is overexpressed in several human cancers, currently a great deal of interest exists in developing new inhibitors of this enzyme to induce an accumulation of endogenous ceramide and a suppression of cell growth (139). Ceramidase also catalyzes the reverse of the hydrolysis reaction, for example, the condensation of sphingosine with a fatty acid without the participation of cofactors (140). N-(2-Oxo)-acylsphingosines inhibited acid ceramidase with a low selectivity and low potency (138). D-erythro-2-(N-Myristoylamino)-2-phenyl-1-propanol (D-e-MAPP, Fig. 12) inhibited alkaline ceramidase, which raises the intracellular levels of ceramide and arrests cell growth (141). A ceramide analog in which a 4-nitrophenyl group replaces the aliphatic chain (denoted as B13, Fig. 12) is a potent inhibitor of acid ceramidase; it induced apoptosis in prostate and colon cancer cell lines (142, 143). A urea derivative of ceramide also inhibited mitochondrial ceramidase (69). Analogs of D-e-MAPP and B13 affect the levels of ceramide, sphingosine, and S1P in MCF7 cells (144). Examples of other potent and specific inhibitors of acid ceramidase are LCL-204 (Fig. 12) and LCL-102, which are lysosomotropic analogs of B13.
5. Inhibition of ceramide synthase. A short-chain cyclopropenyl-containing ceramide analog called GT-11 (145) (also called CB-cyclopropenylceramide) (146) inhibits ceramide synthase.

6. Inhibition of dihydroceramide desaturase. The last enzyme in the de novo biosynthesis of ceramide is inhibited by GT-11, but at high concentrations this compound also inhibited S1P lyase and SPTase (147, 148). Fenretinide (4-HPR, Fig. 12) also inhibited dihydroceramide desaturase (147, 148). This synthetic retinoid and its analogs have apoptogenic activity, which elevates intracellular ceramide levels and induces cell death in a variety of cell types in vitro and in vivo by multiple mechanisms (117, 149, 150).

7. Inhibition of S1P lyase. S1P lyase is inhibited by sulfhydryl reagents, FTY 720 (151), and by racemic 2-vinyldihydrosphingosine 1-phosphate (152) (Fig. 12). When S1P lyase is overexpressed, ceramide levels are increased and apoptosis is induced (153).

Other Structural Analogs of Ceramide with High Antiproliferative Activity

The bioactivity of many synthetic ceramide analogs that bear modifications in the long-chain base or in the fatty amide chain has been studied.

Modifications in the sphingosine backbone
(25,3R,4E,6E)-N-Octanoylamidococadecadiene-1,3-diol, a ceramide analog with an additional double bond (between C-6 and C-7), induced accumulation of endogenous ceramide in multidrug-resistant breast cancer cells and induced apoptosis by the mitochondrial pathway without inhibiting the growth of normal epithelial cells (76).

Modifications in the length of the N-acyl chain
In vitro and in vivo studies have shown that short-chain ceramides induce growth arrest (154). A fluorescent analog of C6-ceramide partitioned into caveolin-enriched microdomains of rat aorta vascular smooth muscle cells and led to growth arrest via activation of PKC-ζ, which is recruited to lipid microdomains and subsequently reduces the activity of Akt (155).

Additional polar groups
Ceramide analogs with an additional hydroxyl group in the long-chain base, as in 6-hydroxyceramide (76) and phytoceramide (156), are more effective than ceramide in arresting proliferation of some tumor cell lines. The introduction of a uracil or thiouracil group at C-1 of ceramide provided an effective apoptotic agent (157).

Reduction of the carboxamido group of ceramide to a methylene group
An analog called ceramine induced apoptosis in leukemic cells, which indicates that the carbonyl group of the amide group of ceramide is not required for in vitro cytotoxicity (66, 158). Ceramide analogs with an arylsulfonamido group had higher cytotoxicity activity than the corresponding alkylsulfonamide analogs (159).

N-acylated 2-amino-1,3-diols
Reaction of amino diols with a fatty acid afforded a series of simple ceramide analogs with pro-apoptotic activity in human cancer cell lines (160). An example is N-oleylserinol, which induced apoptosis in cells expressing the pro-apoptotic protein PAR-4 and was used to purify transformed cells from embryonic stem cells before implantation into mouse brain (161).

Targeting of ceramide analogs to specific cell organelles
Short- and long-chain ceramides form channels in the outer mitochondrial membrane, which enables proteins in the intermembrane space to be released (162). An aromatic "ceramidoid" in which the N-acyl chain terminates in an N-alkylpyridinium group bears a net positive charge (163) (Fig. 13). This ceramide analog preferentially targets mitochondria, where it accumulates and induces apoptosis by triggering the release of cytochrome c into the cytosol, activating the apoptotic cascade, and blocking the growth of cell carcinomas in vitro and in vivo (164). LCL-204 (also called AD2646) accumulates in lysosomal membranes, inhibiting acid ceramidase, inducing the release of cathepsins, and triggering apoptosis of prostate cancer cells (165) and head and neck squamous cell cancer cells (166). LCL-204 also reduced resistance to Fas.

Ceramide analogs that inhibit protein kinase C isozymes and induce downstream loss of extracellular signal-regulated kinase (ERK1/ERK2)
Inhibition of the PKC isozyme superfamily elicits apoptosis in tumor cells. Phenethyl isothiocyanate (PEITC) conjugates of sphingosine and sphinganine (Fig. 13) exert potent antiproliferative effects in human leukemia HL-60 cells by the inhibition of conventional PKC and novel PKC activity and ERK1/ERK2 activity (167). The activity of these derivatives surpassed that of safingol (L-threo-sphinganine), a non-natural analog of the natural
Sphingolipid Regulation of Signaling via Control of Raft Formation and Stability

Sphingomyelin and glycosylated sphingolipids (GSLs), together with sterols and glycerophospholipids, are key lipid building blocks of rafts. As rafts are considered to represent sites for the initiation of many receptor-mediated signaling events including the uptake of pathogens, sphingolipids play pivotal roles in a variety of dynamic cellular events such as membrane trafficking and the activity of membrane proteins.

Disruption of plasma membrane microdomains by sphingolipid analogs that cannot pack tightly with neighboring lipids

Fluorescent analogs of D-erythro-(or 25,3S)-lactosylceramide and D-erythro-sphingomyelin undergo endocytosis from the plasma membrane via caveolae, whereas non-natural analogs such as L-threo-(or 25,3S)-lactosylceramide and L-threo-sphingomyelin follow a predominantly clathrin-dependent route of endocytosis (16). These observations suggest that the stereochemistry at the C-3 position of the sphingosine backbone plays a role in the internalization pathway of the sphingolipid. Studies with a fluorescent, excimer-forming derivative of lactosylceramide (BODIPY-LacCer, Fig. 14) indicated that the D-erythro stereoisomer forms clusters in membranes, from which caveolar endocytosis is initiated. The corresponding L-threo stereoisomer was excluded from these domains; molecular modeling indicated that it does not pack as tightly with neighboring lipids as the natural stereoisomer.

A non-natural glycosphingolipid inhibited caveolar uptake, viral binding and infectivity, and β1-integrin signaling

The addition of D-erythro-N-octanoyl-lactosylceramide (CD-D-e-LacCer, Fig. 14) to human skin fibroblasts at low temperature promoted the formation of plasma membrane microdomains as shown by confocal fluorescence microscopic studies. The addition of CD-D-e-LacCer to cells also initiated the clustering of β1-integrins within these domains and the activation of β1-integrins. On warming to 37°C, β1-integrins were internalized rapidly via caveolar endocytosis in cells treated with CD-D-e-LacCer, whereas little β1-integrin underwent endocytosis in untreated fibroblasts. The incubation of cells with CD-D-e-LacCer followed by a brief warm-up also caused src activation and a reorganization of the actin cytoskeleton (16).

Conversely, addition of C8-L-threo-LacCer (Fig. 14) inhibited the formation of microdomains in the plasma membrane, which resulted in the inhibition of both endocytosis and β1-integrin signaling (170). Thus, non-natural sphingolipids that can block the formation of microdomains in the plasma membrane may offer a novel means for interfering with the cell entry mechanism that is employed by various pathogens and viruses.

Synthetic Immunomodulatory Sphingolipids: Analogs of the Immunosuppressant FTY720 and the Immunostimulator α-Galactosylceramide

FTY720 interferes with S1P signaling and modulates lymphocyte function

Sphingosine 1-phosphate (S1P) mediates numerous biological processes; therefore, potential new drug candidates may be specific agonists and antagonists of S1P receptors. Chemical manipulation of myriocin led to the development of a new immunosuppressive sphingosine analog known as FTY720 (2-amino-2-(4-octylphenylethyl)-1,3-propanediol), which is phosphorylated in vivo to form the (S)-phosphate (FTY720-phosphate, Fig. 15) (173). This synthetic analog of S1P is a potent agonist of the S1P-type 1 receptor but does not activate the S1P-type 2 receptors on the surface of thymocytes and lymphocytes. FTY720 possesses more potent immunosuppressive activity than myriocin without inhibiting sphingolipid biosynthesis and host immune defense responses to many infectious agents. FTY720 inhibits lymphocyte trafficking in vivo, which promotes the sequestration of lymphocytes into lymph nodes and impairs the S1P-type 1 receptor-mediated migration of lymphocytes between secondary lymphoid tissues and the blood. Thus, the cells become unresponsive to S1P and external signals that direct these cells to sites of inflammation (172). In addition, FTY720 has other potent suppressive effects on T cells unrelated to migration, including the inhibition of the S1P-evoked generation of cytokines that promote autoimmune inflammation such as IL-17 (173). (S)-FTY720-phosphate and other FTY720 analogs that activate S1P receptors and stimulate various signaling pathways seem to be useful for the treatment of a variety of pathologic conditions, including angiogenesis, inflammation, respiratory distress syndrome, and autoimmune diseases.

Promotion of myelination by FTY720 and (S)-FTY720-phosphate

Although FTY720 failed to improve efficacy for preventing renal allograft rejection in Phase III clinical studies (174), a different purpose has been found for its potential use. Phase III clinical trials are underway to examine the use of FTY720 for...
Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

Potential treatment of systemic lupus erythematosus and autoimmune demyelinating diseases such as multiple sclerosis (175). (S)-FTY720-phosphate stimulated, via induction of ERK1/2 and Akt phosphorylation, the survival of progenitor cells that give rise to myelin-producing oligodendrocytes (176). Therefore, in addition to its immunosuppressive function, FTY720 shows promise as a therapeutic agent in the treatment of multiple sclerosis via the protection of oligodendrocytes and the replenishment of lost oligodendrocytes, thus promoting remyelination.

Other potential clinical applications of FTY720

The combination of FTY720 with a tyrosine kinase inhibitor induced apoptosis in melanoma cells (177). (R)-AAL, an ether analog of FTY720 (which also has a methyl group in place of a hydroxymethyl group) (Fig. 15), was phosphorylated in vascular cells and blocked vascular endothelial growth factor (VEGF)-induced vascular permeability in vivo (178). (S)-FTY720-phosphate also may regulate calcium ion channels in smooth muscle cells in an S1P-receptor independent manner (179). FTY720 also enhances pulmonary endothelial cell barrier integrity by a mechanism that seems to be different than that used by S1P, which also augments endothelial cell vascular barrier integrity (180).

Glycosphingolipid-Based Immunotherapies: Presentation of Glycosphingolipids to a Subset of T Cells

A naturally occurring α-anomer of galactosylceramide known as KRN7000 (Fig. 16), was isolated from the marine sponge Agelas mauritianus in 1992 and was subsequently identified as an antitumor agent and a potent immunostimulant of invariant natural killer T (iNKT) cells. These cells are an innate subset of T lymphocytes that express a semi-invariant T-cell receptor (TCR). They rapidly produce immunoregulatory cytokines on recognition of glycolipid ligands bound to the CD1d glycoprotein on the surface of antigen-presenting cells, which results in the activation of dendritic cells, NK cells, B cells, and T cells. The function of CD1d is related to the nonclassical major histocompatibility complex (MHC) class I molecules. CD1d presents glycolipid antigens to the TCR of NKT cells (181). A synthetic analog of KRN7000 known as α-galactosylceramide (α-GalCer, Fig. 16) is the most extensively studied lipid antigen that binds to CD1d. Although α-GalCer is a highly potent glycolipid antigen for cells that express CD1d, it is not a natural product of mammalian cells. The identity of the natural mammalian glycolipid antigen that binds to CD1d for presentation to iNKT cells has remained elusive despite intense study. X-ray crystallographic studies showed that the phytosphingosine backbone and fatty amide chain of α-GalCer reside in the two CD1d clefts, with hydrogen bonding between residues in CD1d and the 2’, 3’, and 4’-hydroxy groups of galactose and the 3-hydroxy group of phytosphingosine (182). The 4’- and 6’-positions of the sugar head group are exposed for recognition of the αGalCer/CD1d complex by the TCR. CD1d is recycled through the endosomal and/or lysosomal pathway, and this trafficking is essential for the loading of CD1d with glycolipid antigens (183).

Manipulation of the Th1 versus Th2 polarization with α-GalCer analogs

Although the therapeutic potential of α-GalCer as an immunomodulator for treating autoimmune and infectious diseases, cancer, hepatitis B, and malaria is well established, several limitations to its therapeutic use have been recognized. These include the production of both T-helper 1 (Th1) and T-helper 2 (Th2) cytokines, long-term iNKT cell unresponsiveness in mice on repeated administration, and a very low solubility of the glycolipid in aqueous media. Therefore, many α-GalCer analogs with modifications in the two lipid chains or in the galactose head group have been synthesized and studied. The balance between the levels of Th1 and Th2 cytokines secreted may be crucial to attaining therapeutic efficacy of the immunostimulant because Th1 and Th2 have opposing actions (184). For example, a glycolipid agonist that stimulates NKT cells to produce Th1-type cytokines (such as interferon-γ and the interleukins IL-2 and IL-12) would be beneficial for...
the treatment of intracellular infections caused by viruses and some bacteria and for the stimulation of antihuman immunity. Several α-GalCer analogs have been synthesized that modulate the NKT cell responses by polarizing them to produce a desired cytokine profile. An N-acyl analog of α-GalCer that has a C20:2 cis-double bond in place of the C20:0 amide chain of the natural agelasphins induced a Th1-type response (185) as did analogs that have an aromatic group at the ω-terminus of the N-acyl chain (186). However, an analog of α-GalCer with a truncated phytosphingosine chain stimulated NKT cells to selectively produce Th2-type cytokines, such as IL-4, and suppressed autoimmunity in animal models of multiple sclerosis (187), collagen-induced arthritis (188), and autoimmune diabetes (189). An α-GalCer analog with an elongated phytosphingosine chain induced NKT cells to produce a weaker cytokine response, which resulted in the suppression of an arthritis disease model in mice (190).

A increased production of Th1 cytokines has also been achieved in mice with a C-glycosidic analog of α-GalCer, in which the linker region between the sugar and the backbone was altered by replacing the α-anomeric oxygen atom of galactose with a methylene group (isosteric C-α-GalCer, Fig. 16 (191). A nonisosteric C-glycoside of α-GalCer also was prepared in which the anomeric carbon is bonded directly to C-1 of the phytosphingosine backbone (nonisosteric C-α-GalCer, Fig. 16 (192). This analog elicited a higher ratio of Th1-type cytokine/Th2-type cytokine response in human NKT cells in culture than α-GalCer and the isosteric C-α-GalCer. The mechanism of the Th1 bias remains to be elucidated, but factors such as aqueous solubility, access to lipid transfer proteins, the stability of the CD1d:glycosphingolipid-TCR ternary complex, and receptor-mediated uptake may be involved. Thus, new α-GalCer analogs may have therapeutic relevance.

**Conclusion**

Many natural compounds, dietary constituents, and chemotherapeutic agents alter the activities of enzymes in the de novo biosynthetic pathway of sphingolipids, which thereby modifies the relative concentrations of endogenous sphingolipid species present in cells. During the last two decades the importance of sphingolipids in cell signaling pathways has been recognized, and many bioactivities of sphingolipid metabolites have been studied. The mechanisms by which sphingolipids exert their biological activities on intra- and extracellular targets have been elucidated, but the roles of sphingolipid-metabolizing enzymes and their isozymes in disease processes are not yet understood. The attention that has been devoted to the enzymes that metabolize sphingolipids has resulted in an improved understanding of the pathways for interconversions among sphingolipid messengers with opposing bioactivities. Manipulation of the levels of sphingolipid metabolites seems to be a means of determining cell fate. In vitro and in vivo studies have shown that several enzymes are potential targets for therapeutic applications, including sphingomyelinases, ceramidases, GlcCer synthase, sphingosine kinases, and S1P lyase. Other enzyme targets probably will be identified in future research, such as ceramide synthase, ceramide kinase, autotaxin, and dihydroceramide desaturase. In addition to enzymes involved in sphingolipid metabolism, lysosphospholipid GPCRs hold promise as attractive pharmacologic targets. Agents that disrupt lipid microdomains may also have therapeutic applications. Several human clinical trials have been undertaken to date with synthetic sphingolipids in the hope of assessing their efficacy and safety as chemotherapeutic agents for various diseases: with α-GalCer (193, 194), safingol (195), NB-DNJ (miglustat) (196–198), and FTY720 (175, 199, 200). It remains to be shown whether synthetic sphingolipids will be successful agents for the treatment of infectious and autoimmune diseases and hyperproliferative diseases such as cancer.
References


Synthetic Sphingolipids: as Bioactive Molecules: Roles in Regulation of Cell Function


Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function


Synthetic Sphingolipids as Bioactive Molecules: Roles in Regulation of Cell Function

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

---


---


Further Reading

Biochim. Biophys. Acta, Special Issue: Sphingolipids, Apoptosis, and Disease 2006;1758(12).


Website: sphingolipid classification list: http://www.lipidmaps.org/data/classification/sg.html

See Also

Extracellular Lipid Signals
Glycolipids, Synthesis of Lipidomics
The discovery of natural RNA catalysts has prompted chemical biologists to pursue artificial nucleic acids that have catalytic activities. Such artificial nucleic acid enzymes may comprise either RNA (ribozymes) or DNA (deoxyribozymes). The term “ribozyme” was first used in 1982, when Kruger et al. reported natural catalytic activity by RNA (1). This term is now used universally for catalytic RNAs, whether artificial or natural (see also the WECB review on natural ribozymes). The term “deoxyribozyme” was first used in 1994, when Breaker and Joyce reported the first artificial DNA catalyst (2); no natural deoxyribozymes have been identified. With current knowledge, we cannot design nucleic acid enzymes that have completely new catalytic activities from first principles or by rational modification of known enzymes. Instead, we must use combinatorial search techniques to identify functional RNA and DNA sequences by sifting through a large number of random sequences using an appropriate search strategy. This process of “in vitro selection” has proven useful for identifying nucleic acid enzymes with a wide range of catalytic activities. In certain cases, we can apply ribozymes and deoxyribosymes for practical purposes such as sensing the presence of an analyte and providing a visible response. This review describes how artificial nucleic acid enzymes are increasingly used in practical applications ranging from analytical chemistry to biology.

**Scope of Ribozyme Catalytic Activities**

The first artificial ribozyme was reported in 1990 (3). Since that time, in vitro selection has been used to discover many artificial ribozymes with a wide range of catalytic activities. The known natural ribozymes catalyze phosphodiester cleavage or ligation, with the exception of the ribosome—made of both RNA and protein—that catalyzes peptide bond formation. Many artificial ribozymes also catalyze phosphodiester exchange reactions (RNA/DNA cleavage or ligation), although a growing number of ribozymes catalyze other reactions. Artificial ribozymes for phosphodiester cleavage or ligation have been emphasized in part because nucleic acid catalysts can readily bind via Watson-Crick base pairs to oligonucleotide substrates. By segregating the binding and catalysis functions to distinct regions of the nucleic acid enzyme (Fig. 1), the difficulty of achieving catalysis is reduced. The substrate binding energy is designed directly into the system, and the enzyme needs only to catalyze the reaction. The scope of artificial ribozyme activities identified to date is provided in Table 1 (4–52); representative examples of ribozymes and the reactions that they catalyze are shown in Fig. 2. Such a compilation indicates that many different types of reactions are amenable to nucleic acid catalysis, which includes such prototypical “organic” reactions as the aldol reaction and the Diels–Alder cycloaddition, which form carbon–carbon
In vitro selection and application of nucleic acid enzymes (ribozymes and deoxyribozymes)

Figure 1 Schematic view of nucleic acid enzyme catalysis, showing separate binding and catalytic regions of the ribozyme or deoxyribozyme (lower strand). At the outset of selection, the binding regions are fixed in sequence, whereas the catalytic region comprises a random sequence. In this example, the two substrates (upper strands) are oligonucleotides, which interact with the binding regions by Watson–Crick base pairs; X and Y are the two functional groups that react with one another (solid arrows). Substrates may be small molecules rather than oligonucleotides, and some nucleic acid enzymes break rather than form bonds in the substrate. During most selection procedures, one substrate is attached covalently to the enzyme strand (e.g., via dashed loop at right), which enables selection to occur by linking genotype (sequence) to phenotype (catalysis). However, for practical application such attachment may not be required, particularly when the substrates are oligonucleotides.

General Considerations for In Vitro Selection Procedures

The identification of new artificial ribozymes is impossible without carefully designed in vitro selection methodologies. Although the details and even the fundamental elements of the selection procedure can differ for each ribozyme, some general considerations are common. (See also the WEBC review on construction and selection of nucleic acid combinatorial libraries.) An early approach to ribozyme selection was to identify RNA sequences that bind to a transition-state analog, as has been done for catalytic antibodies. Although such approaches do work in certain cases (45, 48, 51), in other cases the approach was unsuccessful (53), and it was suggested that selecting directly for catalysis is superior (8). Indeed, most ribozyme selections now aim directly for selecting catalytic RNA sequences using the desired substrates rather than transition-state analogs.

Sequence Space and Length of the Random Region

A ribozyme selection experiment almost always begins with solid-phase synthesis of a long DNA oligonucleotide that has two types of sequence elements: constant and random (Fig. 1). The constant regions either serve as primer binding sites for a polymerase chain reaction (PCR) step or—as after conversion to RNA—provide binding sites for nucleic acid substrates. The random region constitutes the nucleotides that, as RNA, will compose the catalytic portions of any functional ribozymes that emerge from the selection process. For some but certainly not all ribozymes, RNA nucleotides from the constant regions can also contribute to catalysis.

One critical consideration for any selection effort is the nucleotide length of the random region. This length directly determines the possible number of nucleotide sequences—i.e., the size of the “random pool”—in a mathematically straightforward way. Because N is the common designation for a random nucleotide (versus A, G, U, and C for the four standard RNA nucleotides), the random region is usually denoted as, for example, N20 for a 20-nucleotide region. For a statistically random N20 region, there are $4^{20} \approx 10^{6}$ possible nucleotide sequences. Successful ribozyme selections have used random regions that range in length up to N25, for which sequence space has the unimaginably large value of $4^{25} \approx 10^{77}$. The number of random-pool molecules actually used to initiate selection is limited by technical considerations such as a manageable PCR volume and is on the order of $10^{13}$–$10^{15}$, which corresponds to 0.01–10 nmol. Therefore, for all but the shortest random regions ($<N20$), sequence space is vastly undersampled. For a relatively small 40-nucleotide random region and starting with $10^{14}$ molecules, only $10^{-10}$ of the $10^{44}$-molecule sequence space is represented, for a typical N20 pool, only $10^{-28}$ of the $10^{44}$-molecule sequence space is covered. Despite the sparse sampling, many selection experiments are quite successful, which implies that catalytically active RNA sequences are relatively common in sequence space, at least for the investigated catalytic activities. Such a conclusion has been reached on the basis of experimental data numerous times (8, 23).

It can be difficult to choose the proper random-region length. This choice requires a compromise between coverage of sequence space (always favoring small random regions) and achieving the structural complexity necessary to support catalysis (probably but not necessarily favoring large random regions; note that excess sequence elements can inhibit catalysis). In selections for identifying aptamers (ligand-binding RNAs), the optimal random-pool length was found to reflect a balance between these factors (54). In ribozyme selections, the structural complexity of the catalytic motif was found to influence strongly the optimal pool size, with complex motifs particularly benefiting from long random pools (55). Of course, before undertaking a new selection effort, one does not necessarily know the structural complexity of the ribozymes that will be identified. Therefore, in practice, the random-pool length is usually chosen based on experience and educated guesswork. To handle the uncertainty, parallel selection experiments that use more than one random-pool length may be performed. Several random pools of various lengths may also be allowed to compete directly with one another in the same tube (19, 56).
In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes)

Table 1: Tabulation of artificial ribozymes by type of reaction catalyzed and type of bond formed or broken

<table>
<thead>
<tr>
<th>Reaction catalyzed</th>
<th>Bond</th>
<th># Rand nt</th>
<th>Rate enh</th>
<th>$M^{2+}$ req</th>
<th>Selection method</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,3-cyclic phosphate hydrolysis</td>
<td>O-P</td>
<td>0</td>
<td>50</td>
<td>Pb$^{2+}$</td>
<td>PAGE shift</td>
<td>(4)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>0</td>
<td>80</td>
<td>Pb$^{2+}$</td>
<td>PAGE shift</td>
<td>(5)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>100</td>
<td>200</td>
<td>Na$^{+}$</td>
<td>PAGE shift</td>
<td>(6)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>30</td>
<td>nd</td>
<td>None</td>
<td>PAGE shift</td>
<td>(7)</td>
</tr>
<tr>
<td>RNA ligation</td>
<td>O-P</td>
<td>220</td>
<td>$7 \times 10^6$</td>
<td>Mg$^{2+}$</td>
<td>seq tag acquisition</td>
<td>(8)</td>
</tr>
<tr>
<td>RNA ligation</td>
<td>O-P</td>
<td>220</td>
<td>$8 \times 10^6$</td>
<td>Mg$^{2+}$</td>
<td>seq tag acquisition</td>
<td>(9)</td>
</tr>
<tr>
<td>RNA ligation</td>
<td>O-P</td>
<td>210</td>
<td>$5 \times 10^6$</td>
<td>Mg$^{2+}$</td>
<td>seq tag acquisition</td>
<td>(10)</td>
</tr>
<tr>
<td>RNA ligation</td>
<td>O-P</td>
<td>115</td>
<td>250</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(11)</td>
</tr>
<tr>
<td>RNA ligation (branch formation)</td>
<td>O-P</td>
<td>80</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(12)</td>
</tr>
<tr>
<td>RNA phosphorylation</td>
<td>O-P</td>
<td>100</td>
<td>$1 \times 10^6$</td>
<td>Mg$^{2+}$</td>
<td>S-tag acquisition (beads)</td>
<td>(13)</td>
</tr>
<tr>
<td>RNA phosphorylation</td>
<td>O-P</td>
<td>0</td>
<td>$6 \times 10^6$</td>
<td>Mg$^{2+}$</td>
<td>S tag qac + PAGE shift</td>
<td>(14)</td>
</tr>
<tr>
<td>RNA capping</td>
<td>O-P</td>
<td>90</td>
<td>$-10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(15)</td>
</tr>
<tr>
<td>RNA capping</td>
<td>O-P</td>
<td>50</td>
<td>nd</td>
<td>Ca$^{2+}$</td>
<td>PAGE shift + run on bead</td>
<td>(16)</td>
</tr>
<tr>
<td>RNA capping</td>
<td>O-P</td>
<td>0</td>
<td>nd</td>
<td>Ca$^{2+}$</td>
<td>rational modification</td>
<td>(17)</td>
</tr>
<tr>
<td>amino acid acylation</td>
<td>O-P</td>
<td>80</td>
<td>nd</td>
<td>Ca$^{2+}$</td>
<td>S tag acquisition (beads)</td>
<td>(18)</td>
</tr>
<tr>
<td>cofactor synthesis</td>
<td>O-P</td>
<td>30, 60</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>biotin tag + PAGE shift</td>
<td>(19)</td>
</tr>
<tr>
<td>RNA polymerization</td>
<td>O-P</td>
<td>0</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>rational modification</td>
<td>(20)</td>
</tr>
<tr>
<td>template-directed pol</td>
<td>O-P</td>
<td>70-0</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>S tag qac + PAGE shift</td>
<td>(21)</td>
</tr>
<tr>
<td>RNA-protein conjugation</td>
<td>N-P</td>
<td>152</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(22)</td>
</tr>
<tr>
<td>Diels–Alder reaction (novo)</td>
<td>C-C</td>
<td>100</td>
<td>800</td>
<td>Ca$^{2+}$</td>
<td>biotin tag + PAGE shift</td>
<td>(23)</td>
</tr>
<tr>
<td>Diels–Alder reaction (novo)</td>
<td>C-C</td>
<td>0</td>
<td>$1 \times 10^3$</td>
<td>Ca$^{2+}$ + Na$^{+}$</td>
<td>biotin tag + PAGE shift</td>
<td>(24)</td>
</tr>
<tr>
<td>aldol reaction</td>
<td>C-C</td>
<td>120</td>
<td>$1 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads, br)</td>
<td>(25, 26)</td>
</tr>
<tr>
<td>alcohol oxidation</td>
<td>C-H</td>
<td>142</td>
<td>$4 \times 10^3$</td>
<td>Zn$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(27)</td>
</tr>
<tr>
<td>aldoldehyde reduction</td>
<td>C-H</td>
<td>0</td>
<td>$3 \times 10^3$</td>
<td>Mg$^{2+}$ + Zn$^{2+}$</td>
<td>rational modification</td>
<td>(29)</td>
</tr>
<tr>
<td>pyrimidine nt synthesis</td>
<td>C-N</td>
<td>228</td>
<td>$1 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>S + PAGE, biotin (beads)</td>
<td>(30, 31)</td>
</tr>
<tr>
<td>purine nt synthesis</td>
<td>C-N</td>
<td>95</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>S tag qac + PAGE shift</td>
<td>(32)</td>
</tr>
<tr>
<td>N7G alkylation</td>
<td>C-N</td>
<td>0</td>
<td>$3 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(33)</td>
</tr>
<tr>
<td>amide synthesis (novo)</td>
<td>C-N</td>
<td>100</td>
<td>$1 \times 10^3$</td>
<td>Ca$^{2+}$</td>
<td>biotin + PAGE (or beads)</td>
<td>(34)</td>
</tr>
<tr>
<td>uracil synthesis (novo)</td>
<td>C-N</td>
<td>100</td>
<td>$1 \times 10^3$</td>
<td>nd</td>
<td>reagent tag + PAGE (or beads)</td>
<td>(35)</td>
</tr>
<tr>
<td>peptide bond formation</td>
<td>C-N</td>
<td>142</td>
<td>$1 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(36)</td>
</tr>
<tr>
<td>peptidyl-RNA synthesis</td>
<td>C-N</td>
<td>0</td>
<td>$100$</td>
<td>Ca$^{2+}$</td>
<td>rational modification</td>
<td>(37)</td>
</tr>
<tr>
<td>acyl transfer</td>
<td>C-O</td>
<td>90</td>
<td>$1 \times 10^9$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(38, 39)</td>
</tr>
<tr>
<td>acyl transfer</td>
<td>C-O</td>
<td>120</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(40)</td>
</tr>
<tr>
<td>ammonolysis</td>
<td>C-O</td>
<td>50</td>
<td>$2 \times 10^3$</td>
<td>Mg$^{2+}$ + Ca$^{2+}$</td>
<td>reagent tag qac + HPLC</td>
<td>(41)</td>
</tr>
<tr>
<td>ammonolysis</td>
<td>C-O</td>
<td>70-0</td>
<td>nd</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(42)</td>
</tr>
<tr>
<td>ammonolysis</td>
<td>C-O</td>
<td>70</td>
<td>$2 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads)</td>
<td>(43)</td>
</tr>
<tr>
<td>ammonolysis</td>
<td>C-O</td>
<td>0</td>
<td>$6 \times 10^3$</td>
<td>Ca$^{2+}$</td>
<td>rational modification</td>
<td>(44)</td>
</tr>
<tr>
<td>carbonate hydrolysis</td>
<td>C-O</td>
<td>70</td>
<td>100</td>
<td>None</td>
<td>SELEX for TS analog</td>
<td>(45)</td>
</tr>
<tr>
<td>phosphorothioate alkylation</td>
<td>C-S</td>
<td>30</td>
<td>$2 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>S tag + PAGE (or beads)</td>
<td>(46)</td>
</tr>
<tr>
<td>Michael reaction</td>
<td>C-S</td>
<td>142</td>
<td>$3 \times 10^3$</td>
<td>Mg$^{2+}$</td>
<td>biotin tag acq (beads, hr)</td>
<td>(47)</td>
</tr>
<tr>
<td>proline mutarotation</td>
<td>C-N</td>
<td>50</td>
<td>500</td>
<td>Mg$^{2+}$ + Ca$^{2+}$</td>
<td>SELEX for TS analog</td>
<td>(48)</td>
</tr>
<tr>
<td>Pd nanoparticle formation (novo)</td>
<td>Pd-Pd</td>
<td>40</td>
<td>nd</td>
<td>None</td>
<td>nanoparticle binding</td>
<td>(49, 50)</td>
</tr>
<tr>
<td>biotin immobilization</td>
<td>None</td>
<td>28</td>
<td>88</td>
<td>Mg$^{2+}$</td>
<td>SELEX for TS analog</td>
<td>(51)</td>
</tr>
</tbody>
</table>

NOTES: This tabulation is not intended to be exhaustive, in that other examples of ribozymes with the indicated activities may have been reported, nd = not determined in published report. When the reaction type is marked with "(novo)," the ribozyme required nonstandard nucleotides for activity. The entry "# Rand nt" is the total number of original random nucleotides in the sequence pool, these may have been in two or more portions interspersed with constant regions. The entry "Rate enh" is the highest rate enhancement reported, often calculated as the ratio of rate constants for the observed versus background reactions (Fig. 4). In some cases, the authors reported the rate enhancement as the ratio of $k_{obs}/k_{0}$ for the ribozyme and the analogous value for the uncatalyzed reaction. The listed rate enhancement may be a lower limit on the true value, for at least one of two reasons: (1) Only an upper limit was possible on the uncatalyzed rate. (2) The uncatalyzed rate represents a spectrum of reactions, only one of which corresponds to the particular reaction catalyzed by the ribozyme. Under "Selection method," the entry "(beads)" is included if noncovalent binding to beads (or other solid support) of the tagged nucleic acid sequences was an integral part of the key selection step. In some other cases, the substrate was presented on beads to avoid aggregation, but the key selection step did not involve a solid support. The entry "(corex") is included if a photosensitizer release step was used.

8The experiment started with a known aptamer or ribozyme sequence, which in some cases was partially randomized. In some cases, a new random-sequence domain of indicated length was also included.

9The experiment used rational redesign of a known ribozyme sequence.
Tolerance of Ribozymes to Variation in Sequence

Many artificial ribozymes do not seem to require specific nucleotides at every position within their initially random regions. This is similar to how many proteins tolerate amino acid changes at numerous sites but not at certain key residues. Due to the hierarchical nature of RNA folding, in which a complex tertiary structure forms on a foundation of Watson–Crick secondary structure elements such as stem-loop structures (57), a ribozyme can probably tolerate nucleotide changes at certain positions as long as Watson–Crick interactions are maintained (see Reference 58 for similar findings with aptamers). At other positions, the identities of the nucleotides may not matter at all. These considerations increase substantially the possibility that a particular ribozyme will emerge from a selection experiment, because many variants of a ribozyme that differ at noncritical nucleotides are essentially equivalent in terms of catalytic activity.

Overview of Experimental Selection Strategies

In general, a successful in vitro selection experiment must physically link the information in the catalytic nucleic acid sequence with the desired reaction chemistry, such that the successful sequences can be isolated. Therefore, an experimental strategy must be devised by which the majority of catalytically incompetent random-pool sequences are discarded, whereas the small minority of functional random-pool sequences are retained. Many strategies have been used, and many variations are possible even within the framework of a single general strategy. One possible selection approach is represented by the strategy used to identify the first RNA ligase ribozymes (Fig. 3) (8). As a means of illustrating a selection process, the key aspects of the procedure are discussed below, using the RNA ligase ribozyme selection as the specific example.

Preparation of the Random-Pool DNA

The random-pool DNA is prepared by solid-phase synthesis (SPS), with the random (N) nucleotides provided simply by mixing the four standard DNA nucleotide phosphoramidites together in one bottle. During SPS, random coupling to each growing oligonucleotide chain ensures that the collection of synthetic oligonucleotides has an effectively random region that encompasses all N positions. In practice, random coupling to each growing oligonucleotide chain ensures that the collection of synthetic oligonucleotides has an effectively random region that encompasses all N positions. In practice, the phosphoramidites are mixed in non-1:1:1:1 mole ratio, because each phosphoramidite reacts during SPS with a different rate constant. After correcting for these known unequal coupling efficiencies, the random pool will have an approximately equal proportion of each standard DNA nucleotide at each position. Once the random-pool DNA has been synthesized, the random-pool RNA is made using this DNA as a template for in vitro transcription using T7 RNA polymerase (59).
...ribozymes (8), the selections were performed primarily at pH 7.4, 600 mM KCl, and 60 mM MgCl₂ at 25 °C for 16 h, at least for the initial selection rounds. Due to the large number of independent variables, selection experiments often take advantage of parallel processing. Multiple selection experiments that each use a specific set of incubation conditions are performed. By comparing the outcomes of several selections performed in parallel where the only difference is in the incubation conditions, one may gain information on which aspects of these conditions are essential to achieving the desired catalytic activity.

Once the catalytically active nucleic acid sequences have been separated by an appropriate physical method, the selection round must be completed by synthesizing the DNA pool for input into the next selection round, now enriched in those sequences that (as ribozymes) are competent for catalysis. For the RNA ligase ribozymes, this was achieved by reverse transcription and PCR amplification followed by transcription (Fig. 3). Similar steps are common to most ribozyme selection procedures.

The overall selection cycle is iterated multiple times until the activity of the pool is sufficiently high that identifying individual sequences within the pool is warranted. The necessary number of rounds can vary greatly but is typically between 5 and 15 (only rarely greater than 20). The need for multiple selection rounds may initially seem mysterious—how come catalytically active sequences do not emerge after just one selection round? The reason is that most of the astronomically large number of possible random-pool sequences are not truly catalytically active, yet they may accidentally survive a particular selection round merely by chance. Only by requiring reproducible catalysis over multiple selection rounds can active ribozyme sequences dominate the selection pool.

A hazard of any selection experiment is that certain nucleic acid sequences may be able to survive even though they do not actually catalyze the desired chemical transformation. There is no general solution to this problem, other than to take great care in designing the selection procedure to avoid giving such opportunistic sequences a means of survival. Occasionally “negative selection” pressure can be applied, in which sequences that catalyze an undesired reaction are intentionally discarded before the remaining sequences are offered the opportunity to catalyze the desired reaction.

Testing the Catalytic Activities of Individual Ribozymes

Once the activity of the pool has reached an acceptable level after multiple rounds, individual ribozymes are cloned (but not yet...
Ribozyme Optimization

After individual catalytically active ribozyme sequences have been identified, a “minimal” ribozyme is often devised by systematically deleting portions of the sequence. Usually this is guided by computer folding algorithms such as mfold (60) that can predict RNA secondary structure with good reliability, particularly for small RNAs. Empirically, it is often found that certain regions of an initially identified ribozyme (such as a large single-stranded loop) can be shortened or even eliminated without damaging catalysis. This is advantageous because smaller ribozymes are easier to synthesize and less likely to suffer out damaging catalysis. This is advantageous because smaller ribozymes are easier to synthesize and less likely to suffer nonspecific degradation, and they are also easier to study mechanistically because fewer nucleotides must have their chemical roles explored. However, as shown with the natural hammerhead ribozyme (61), one must not be overzealous in removing nucleotides that may be catalytically critical. A newer approach to determining a minimal ribozyme motif is to use nonhomologous randomization (typically changing the concentrations of divalent metal ions Mg^{2+} and Mn^{2+}) can intentionally elevate the polymerase error rate considerably, thereby introducing substantial variation during the DNA amplification step. An important consideration is the types of mutations that are made; some conditions favor certain nucleotide changes, whereas other conditions provide more randomness. Instead of altering the polymerase error rate, nonstandard nucleotide triphosphate analogs may be used to promote random mutations during PCR (64).

Second, a selection effort can be restarted from the beginning using a partially randomized sequence pool. In this “reselection” approach, a new pool is prepared by SPS on the basis of a known ribozyme sequence and with an enzyme region that is partially randomized, meaning that a marked asymmetry or mixture of the four nucleotide phosphoramidites is used for each N position. A typical reselection experiment might use a partially randomized pool in which all enzyme region nucleotides have the appropriate nucleotide of the parent ribozyme with 70% probability, or one of the other three nucleotides with 10% probability each. As for consideration of initial random-pool length, the extent to which each nucleotide position is randomized is based on both experience and guesswork. As a guide, one can compute the distribution of nucleotide changes relative to the “parent” sequence for any set values of randomization probabilities (65, 66), thereby knowing how many nucleotides (on average) will be changed relative to the parent sequence.

Continuous Evolution as an Alternative Method

One principal drawback of many conventional in vitro selection procedures is their tediousness, particularly when individual selection rounds are time-consuming. This also limits the number of selection rounds that may be performed. To obviate such issues, investigators have performed “continuous evolution” experiments. In continuous evolution, samples of replicating molecules are diluted serially (e.g., 10^3-fold dilution, often with >100 serial transfers) and at constant temperature (isothermal amplification), rather than selected in discrete rounds and with temperature cycling that is characteristic of PCR. The history of continuous evolution goes back to Spiegelman’s work in the 1960s on Qβ replicase, in which self-replication by minimal nucleic acid sequences was sought (67). More recently, Wright and Joyce broadened the approach of continuous evolution to encompass evolution of catalytic function (68). In one experiment, a continuous evolution approach was used to evolve an RNA ligase ribozyme that is “resistant” to the activity of an RNA-cleaving deoxyribozyme (69). The application of continuous evolution to ribozymes has advantages in terms of both speed and amplification power, but there are disadvantages in terms of susceptibility to contamination and limitations on the types of reactions that may be catalyzed (70).
Multiple Turnover Using In Vitro Compartmentalization (IVC)

An inherent limitation of most selection approaches is the absolute requirement for a covalent linkage between the catalytic nucleic acid and its substrate during selection. Although in many cases the emergent ribozymes can be converted into intermolecular (trans-acting) forms, during the selection process each candidate ribozyme is restricted to two possible chemical yields, 100% or 0%: i.e., each molecule has either performed or not performed the desired chemical reaction. Consequently, selection for multiple turnover is not possible. (In some cases, artificial ribozymes are capable of multiple turnover anyways, but this must be considered as fortuitous.) To alleviate this difficulty, candidate ribozymes have been encapsulated along with unattached substrates within individual droplets in a water-in-oil emulsion. This process is termed in vitro compartmentalization (IVC) (71, 72). If the system is arranged such that each droplet contains just one type of ribozyme candidate, then the encapsulation serves the same conceptual purpose as a covalent bond in terms of linking information (“genotype”; the ribozyme sequence) to catalytic ability (“phenotype”). An IVC strategy has been applied to develop several ribozymes with multiple turnover ability (73, 74).

Fundamental Insights Into Nucleic Acid Catalysis from Artificial Ribozymes

A primary motivation for studying artificial nucleic acid enzymes is to gain insight into natural nucleic acid catalysis. These insights include implications for prebiotic chemistry and the RNA World hypothesis, as well as a fundamental mechanistic understanding of nucleic acid enzymes.

Implications for Prebiotic Chemistry and the RNA World

The RNA World hypothesis posits that before the advent of proteins, there was a period of prebiotic evolutionary history in which RNA both carried information and performed catalysis (75) (see also the WECB review on the origins of life: emergence of an RNA world). Although it is probably impossible to reconstruct a complete RNA World in the modern laboratory, exploring the capabilities of artificial ribozymes can provide information and constraints. For example, the identification of a ribozyme for processive template-directed RNA polymerization (21) provides evidence that relatively small RNA molecules (although they do have hundreds of nucleotides) are capable of catalysis that would have been important in a prebiotic RNA-based era. Many studies have focused on similar considerations (e.g., see References 76 and 77).

Mechanisms of Artificial Ribozymes

Little is known about the mechanisms of most artificial ribozymes. The catalytic rate and other basic features of each new ribozyme are generally determined as part of its initial characterization. Detailed mechanistic analyses are infrequent, although on occasion the tools of physical organic chemistry (e.g., kinetic isotope effects (78)) have been applied. Artificial ribozymes are usually, but not always, selected in the presence of divalent metal ions such as Mg$^{2+}$ or many others (Table 1). Not surprisingly, artificial ribozymes typically require a modest-to-high concentration of such metal ions for their function. The catalytic tasks of these metal ions are often unknown, and a major challenge is to separate the roles of metal ion participation in structure versus catalysis.

For each ribozyme, it is usually possible to identify an appropriate “background” reaction that has a relatively low rate. For an RNA ligase ribozyme, a relevant background reaction is the analogous ligation reaction when the functional groups are held together by a complementary splint that has no enzyme region whatsoever (Fig. 4). The rate enhancements determined for various ribozymes in comparison with suitable background reactions can be low (∼10$^2$), but they can also be as high as 10$^{10}$ (Table 1). Although ribozyme rate enhancements are often modest relative to those of protein enzymes, the appropriate benchmarks are the natural ribozymes, and on this basis, artificial ribozymes compete well with their natural counterparts.

The substantial rate enhancements observed for artificial ribozymes imply that they do more than passively hold together their substrates. Therefore, ribozyme catalysis is more than an “effective molarity” phenomenon. Consistent with this, almost all ribozymes make just one product even when multiple products are possible. For example, although the initially reported RNA ligase ribozymes could potentially have made either 3′-5′ or 2′-5′ linear RNA linkages, each particular ribozyme was observed to make just one linkage (8). Depending on geometrical constraints, effective molarity alone could lead to a mixture of products, which is not observed. In most cases, it is not yet clear whether ribozymes generally work by preferential lowering of transition-state energies or by precise positioning of the reacting molecules.
moieties, i.e., orientation effects. Given that such issues are not yet settled for most of the natural ribozymes [such as the ribosome (79, 81)], mechanistic studies of all kinds of ribozymes will continue to be an active research field.

Structural Biology of Artificial Ribozymes

Most (but not all) natural ribozymes have X-ray crystal structures available, and nuclear magnetic resonance (NMR) spectroscopy has also been extensively applied (see also the WECE Reviews on crystallographic techniques: nucleic acids and on NMR: nucleic acids). Perhaps surprisingly, only two artificial ribozymes have been studied using these methods. First, the Phe-dependent RNA-cleaving “leadzyme” has been examined by both X-ray crystallography (82, 83) and NMR spectroscopy (84, 85). Second, a ribozyme for the Diels–Alder reaction has been studied by X-ray crystallography (86), revealing a pre-formed hydrophobic substrate-binding pocket that is capable of enantioselective catalysis. The application of structural biology methods to understand artificial ribozyme function is a field ripe for increased activity.

Evolving New RNA Catalytic Activities Starting From Known Ribozymes

One interesting issue that is relevant both for a fundamental appreciation of RNA catalysis and for practical applications is to understand the circumstances in which a known ribozyme activity may be evolved to provide a different type of catalysis. In one report, evolving a self-aminoacylating ribozyme into a self-phosphorylating ribozyme required a substantial number of mutations, such that the new ribozyme could adopt a distinct structure (14). This implies that escape from the parent ribozyme’s fold is required for evolution of new activity. Another study found that one designed RNA sequence can adopt either of two different folds, each of which catalyzes a different reaction (87). In nature, bifunctional sequences presumably represented transitional structures for the evolution of one catalytic activity into another. However, most bifunctional sequences would probably not exist today, because they are likely suboptimal for the emergent ribozymes. For example, artificial variants of the hammerhead ribozyme are resistant to ribonuclease due to 2′-amino and 2′-fluoro modifications (88). Increased stability to natural RNA-cleaving enzymes is useful for certain practical applications, particularly in vitro (see below).

Ribozymes with Nucleotides Other Than \( A, G, U, \) and \( C \)

The first in vitro selection effort with nonstandard nucleotides identified RNA that catalyzes the Diels–Alder reaction (23, 24). In this work, a pyridine-modified uridine derivative replaced all uridines throughout the RNA. The pyridine moieties were presumed to assist catalysis by some combination of hydrogen bonding, hydrophobic and dipolar interactions, and metal coordination. Most of the new ribozymes were strictly Cu(II)-dependent, suggesting a role for Lewis acid catalysis. Other artificial ribozymes with nonstandard nucleotides (each of which has an unmodified Watson–Crick face that permits use with T7 RNA polymerase and reverse transcription) include an RNA ligase ribozyme with N6-amino-modified adenosines (11), amide syntheses ribozymes that incorporate 5-imidazolyl-U (34), and a urea syntheses ribozyme that also incorporates 5-imidazolyl-U (35). One of the most intriguing artificial ribozyme activities is the ability to induce palladium nanoparticle crystallization using pyridine-modified uridine nucleotides (49). The various ribozymes create different crystal shapes such as hexagonal plates versus cubes, and they require the pyridine-bearing RNA nucleotides for their activities (50). Presumably the pyridines interact directly with metals during crystal formation. Such ribozymes present an intrinsic mechanistic challenge due to the heterogeneous nature of their reaction.

Incorporating nonstandard RNA nucleotides into a selection effort can enhance the chemical (or biochemical) stability of the emergent ribozymes. For example, artificial variants of the natural hammerhead ribozyme are resistant to ribonuclease due to 2′-amino and 2′-fluoro modifications (88). Increased stability to natural RNA-cleaving enzymes is useful for certain practical applications, particularly in vitro (see below).

Ribozymes with Fewer Than Four Kinds of Nucleotide: Minimal Informational Systems

In the opposite of adding nonstandard functional groups to RNA’s chemical repertoire, ribozymes having three or fewer kinds of nucleotide have been investigated. In one study, RNA ligase ribozymes were identified that exclude cytidine (C) from the enzyme region (89). When one such ribozyme was re-selected with inclusion of C, the catalytic rate improved about 20-fold (90), providing a quantitative measure of the importance of variety in ribozyme components. Going to the extreme, an RNA ligase ribozyme was identified that uses only two different nucleotides—a binary informational system (91) comprising only uridine (U) and 2,6-diaminopurine (D) nucleotides (Fig. 5). This ribozyme was inefficient, with only about 8% ligation yield in 80 h at pH 9.0 and 23°C (4.0 × 0.05 h−1). Nevertheless, its activity demonstrates that a dramatically minimized and biologically related informational system can encode catalysis. Such types of ribozymes could have been relevant in an RNA World, at a point in time when only one base pair had evolved.
Practical Synthesis of Ribozymes

For practical synthesis of a ribozyme, two options are generally available. The first method is SPS, which is readily available to nonexperts via several commercial sources and via some academic service facilities. When nonstandard RNA nucleotides are desired, SPS is generally the best approach. The second method is in vitro transcription from a DNA template (59), which in most cases is limited to the four standard RNA nucleotides. A new method (albeit one that requires organic synthesis of the components of a nonstandard base pair) allows enzymatic transcription of RNA with unnatural nucleotides (92).

In Vitro Selection of Deoxyribozymes

Many considerations that are important for ribozymes also apply to deoxyribozymes, which differ solely by the lack of a 2′-hydroxyl group at each nucleotide. Natural DNA is almost always found in double-stranded form, which is structurally uniform and does not support catalysis. In contrast, artificial deoxyribozymes are largely single-stranded, and their conformational flexibility allows catalytic activity. Because RNA already seems to have a paucity of functional groups relative to proteins, early speculation was that the lack of a 2′-hydroxyl would render DNA catalytically inferior to RNA (93), if not altogether incompetent. Nevertheless, experiments have shown that DNA is quantitatively as functional as RNA in terms of catalysis. This is particularly evident when comparisons are made between RNA and DNA enzymes that catalyze the same type of reaction. For example, deoxyribozymes that cleave RNA are at least as competent as analogous ribozymes, if not better (94). This similarity between RNA and DNA may relate to common classes of mechanisms. Evidence increasingly suggests that ribozymes often rely on acid-base catalysis involving the nucleobases (95), which are present in both RNA and DNA. Combining this observation on catalysis with the finding that both RNA and DNA aptamers are highly competent for binding to ligands (96), it is unsurprising that the lack of 2′-hydroxyl groups on DNA is still compatible with its catalytic activity.

Scope of Deoxyribozyme Catalytic Activities

To date, many fewer types of reactions have been explored with deoxyribozymes than ribozymes. This reflects both the later discovery of DNA catalysis and possibly the greater fundamental appeal of RNA catalysis as more relevant to the RNA World hypothesis. Nevertheless, when they have been examined, deoxyribozymes have proven quite competent catalytically (Table 2) (97–122). This extends not only to the types of reactions that have been investigated but also to important characteristics such as rates. For example, the RNA-cleaving 10–23 deoxyribozyme can achieve a kcat of 10 min⁻¹ (rivaling natural ribozymes) and a kcat/Km of 10⁵ M⁻¹ min⁻¹, which is higher than that for the protein enzyme ribonuclease A (97).

Other Considerations for Deoxyribozymes

The same experimental selection approaches that are used for ribozymes are generally applicable to deoxyribozymes. For deoxyribozymes, there is no need to transcribe a sequence pool from DNA to RNA and then reverse-transcribe the successful RNA back into DNA for amplification during each selection round. Instead, the DNA is both the information and the catalyst throughout the selection process. Mechanisms of deoxyribozymes have been probed even less often than for ribozymes; in general, much work remains to understand how DNA catalyzes reactions. Intriguingly, one RNA-cleaving deoxyribozyme uses ascorbate as a cofactor (116) and another uses histidine (123), indicating that nucleic acid enzymes can use cofactors to expand their limited functional group repertoire. Structural biology efforts have also been very modest for deoxyribozymes. In the sole published experiment, the 10–23 deoxyribozyme crystallized in a catalytically inactive 2:2 stoichiometry complex with its RNA substrate (124). Although this was an interesting four-way structure with implications for understanding Holliday junctions, it highlights the need for future efforts to understand the structure of catalytic DNA.

Structural data on DNA aptamers show that they can form ligand-recognition motifs that resemble their RNA counterparts, even without 2′-hydroxyl groups (125).
Table 2. Tabulation of artificial deoxyribozymes

<table>
<thead>
<tr>
<th>Reaction catalyzed</th>
<th>Bond</th>
<th># Rand m</th>
<th>Rate enh</th>
<th>M^{2+} req</th>
<th>Selection method</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>50</td>
<td>$\sim 10^6$</td>
<td>Pb^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(2)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>40</td>
<td>$\sim 10^5$</td>
<td>Mg^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(98)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>50</td>
<td>nd</td>
<td>Mg^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(97)</td>
</tr>
<tr>
<td>RNA cleavage</td>
<td>O-P</td>
<td>40</td>
<td>$1 \times 10^8$</td>
<td>None</td>
<td>biotin tag loss (beads)</td>
<td>(99)</td>
</tr>
<tr>
<td>RNA cleavage (nontid)</td>
<td>O-P</td>
<td>50</td>
<td>nd</td>
<td>Zn^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(100)</td>
</tr>
<tr>
<td>RNA cleavage (nontid)</td>
<td>O-P</td>
<td>20</td>
<td>nd</td>
<td>None</td>
<td>biotin tag loss (beads)</td>
<td>(101)</td>
</tr>
<tr>
<td>RNA cleavage (nontid)</td>
<td>O-P</td>
<td>50</td>
<td>$\sim 10^6$</td>
<td>None</td>
<td>biotin tag loss (beads)</td>
<td>(102)</td>
</tr>
<tr>
<td>RNA ligation (2–5')</td>
<td>O-P</td>
<td>40</td>
<td>300</td>
<td>Mg^{2+}</td>
<td>PAGE shift</td>
<td>(103)</td>
</tr>
<tr>
<td>RNA ligation (3–5') and other</td>
<td>O-P</td>
<td>40</td>
<td>$\sim 10^6$</td>
<td>Mg^{2+}</td>
<td>PAGE shift</td>
<td>(104)</td>
</tr>
<tr>
<td>RNA ligation (3–5')</td>
<td>O-P</td>
<td>40</td>
<td>$\sim 10^5$</td>
<td>Mg^{2+}</td>
<td>PAGE shift</td>
<td>(105)</td>
</tr>
<tr>
<td>RNA ligation (branch formation)</td>
<td>O-P</td>
<td>40</td>
<td>$5 \times 10^5$</td>
<td>Mn^{2+}</td>
<td>PAGE shift</td>
<td>(106, 107)</td>
</tr>
<tr>
<td>RNA ligation (branch formation)</td>
<td>O-P</td>
<td>40</td>
<td>$\sim 10^5$</td>
<td>Mg^{2+}</td>
<td>PAGE shift</td>
<td>(108, 109)</td>
</tr>
<tr>
<td>RNA ligation (lariat formation)</td>
<td>O-P</td>
<td>40</td>
<td>$\sim 10^5$</td>
<td>Mn^{2+}</td>
<td>PAGE shift</td>
<td>(109, 110)</td>
</tr>
<tr>
<td>DNA phosphorylation</td>
<td>O-P</td>
<td>70</td>
<td>$\sim 10^5$</td>
<td>Mn^{2+}</td>
<td>PAGE shift</td>
<td>(112)</td>
</tr>
<tr>
<td>DNA adenylation (capping)</td>
<td>O-P</td>
<td>70</td>
<td>$2 \times 10^{10}$</td>
<td>Mg^{2+} + Cu^{2+}</td>
<td>PAGE shift</td>
<td>(113)</td>
</tr>
<tr>
<td>DNA ligation</td>
<td>O-P</td>
<td>116</td>
<td>$3 \times 10^3$</td>
<td>Cu^{2+} or Zn^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(114)</td>
</tr>
<tr>
<td>DNA ligation</td>
<td>O-P</td>
<td>150</td>
<td>$\sim 10^5$</td>
<td>Mn^{2+}</td>
<td>PAGE shift</td>
<td>(115)</td>
</tr>
<tr>
<td>oxidative DNA cleavage</td>
<td>C-O</td>
<td>50</td>
<td>$\sim 10^6$</td>
<td>Cu^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(116, 117)</td>
</tr>
<tr>
<td>DNA deglycosylation</td>
<td>C-N</td>
<td>85</td>
<td>$9 \times 10^3$</td>
<td>Cu^{2+}</td>
<td>PAGE shift</td>
<td>(118)</td>
</tr>
<tr>
<td>thymine dimer photoreversion</td>
<td>C-C</td>
<td>40</td>
<td>$3 \times 10^5$</td>
<td>None</td>
<td>PAGE shift</td>
<td>(119)</td>
</tr>
<tr>
<td>phosphoramidate cleavage</td>
<td>N-P</td>
<td>72</td>
<td>$\sim 10^6$</td>
<td>Mg^{2+}</td>
<td>biotin tag loss (beads)</td>
<td>(120)</td>
</tr>
<tr>
<td>porphyrin metalation</td>
<td>Cu-N</td>
<td>228</td>
<td>$1 \times 10^3$</td>
<td>Cu^{2+} or Zn^{2+}</td>
<td>binding to TS analog</td>
<td>(121, 122)</td>
</tr>
</tbody>
</table>

NOTES: See Table 1 legend for description of column headings.

As for ribozymes, several deoxyribozymes that incorporate nonstandard DNA nucleotides have been identified. For example, an RNA-cleaving deoxyribozyme with three imidazole functional groups has been obtained (100); other examples include RNA-cleaving deoxyribozymes that have both imidazole and amine nonstandard functionalities (101, 102). DNA can be completely modified by incorporation of solely nonstandard nucleotides during PCR (126). This suggests that future efforts should be able to expand considerably the use of chemically modified DNA for in vitro selection. A different purpose of nonstandard nucleotides is for regulation of catalytic activity. For example, appending an azobenzene moiety onto an RNA-cleaving deoxyribozyme allows the catalytic activity to be switched photochemically (127, 128).

An interesting conceptual question is whether a ribozyme and deoxyribozyme can be active with the same nucleotide sequence (with, of course, the U nucleobases of RNA replaced with T in DNA). In one study in which a deoxyribozyme with hemin-dependent peroxidase activity was identified, exchanging all DNA for RNA reduced but did not destroy the catalysis (129). More recently, a selection approach was used to convert a known ribozyme sequence into a deoxyribozyme (130). Curiously, this deoxyribozyme was nonfunctional when made as either RNA or DNA is intriguing for its implications regarding “crossover” between two types of informational macromolecule.

For practical synthesis of deoxyribozymes (i.e., single-stranded DNA), SPS is the typical approach. If nonstandard nucleotides must be incorporated, SPS is generally the sole viable approach. In contrast, for DNA containing only the four standard nucleotides, conventional PCR may be used. If so, a main challenge is to separate the desired single-stranded DNA product from its complement. This may be achieved by using one primer with a nonamplifiable 5'-tail, such that the two product strands are of unequal length and therefore separable by PAGE (131). Asymmetric PCR that includes only one primer may instead be applied (132), although because DNA synthesis is linear and not exponential in this process, only a limited amount of single-stranded DNA can be generated. As another option, methods have been described for taking advantage of biotin-streptavidin technology to isolate single-stranded DNA from conventional PCR reactions (133, 134).

Terminology for Nucleic Acid Enzymes

An alternative definition of “nucleic acid enzyme” is a protein enzyme that modifies nucleic acids, e.g., T4 polynucleotide kinase (T4 PNK) for phosphorylation of RNA and DNA. Unfortunately, the term “nucleic acid enzyme” can therefore refer either to ribozymes and deoxyribozymes or to protein enzymes.
that modify RNA and DNA. In practice, both definitions of nucleic acid enzymes are used by different groups of authors, and the intended meaning must be gleaned from context.

Occasionally, the term “RNA enzyme” is used as a synonym for ribozyme. In contrast, the contraction “RNAz” is rarely used, perhaps because “ribozyme” is approximately the same length and is just as simple to write. Similarly, “DNAz” is often used as a synonym for deoxyribozyme. The contraction “DNAzyme” is used by some authors to replace the polysyllabic “deoxyribozyme.” However, there is often confusion about the capitalization (i.e., DNAzyme, dNAzyme, or DNAzyme), with only the first of these variants generally considered proper.

RNA and DNA molecules with catalytic activity are related to aptamers. Ribozymes and deoxyribozymes bind to their substrates and catalyze chemical reactions, whereas RNA and DNA aptamers simply bind to their ligands (135, 136). Aptamers are most commonly identified by a procedure originally termed “systematic evolution of ligands by exponential enrichment” or SELEX (135). SELEX is a form of in vitro selection, but not all in vitro selection is SELEX. In particular, SELEX is not generally used to identify nucleic acid enzymes—note the word ligands (not catalysts) in the full form of the acronym. Therefore, except when SELEX is performed on a transition-state analog, RNA and DNA catalysts are identified by “in vitro selection.”

One relatively subtle but important distinction for identifying catalytic nucleic acids is the difference between screening in vitro and in vitro selection. In a screen, all candidates (such as small-molecule compounds from a combinatorial library) are surveyed individually for a desired property. This requires an efficient screening procedure, which for small-molecule combinatorial chemistry often involves multi-well plate experiments or miniaturized systems that allow parallel examination of thousands of spatially segregated samples. (See also the WECB reviews on screening of chemical libraries and on high-throughput screening.) However, statistical considerations for nucleic acid enzymes—e.g., 10^15 sequences examined in a single experiment—obviate any realistic possibility of screening individual sequences one at a time. Instead, selection approaches must be used. The distinguishing feature of a nucleic acid enzyme selection is that a successful candidate sequence must survive through a stringent experimental step based on its catalytic ability. Even though most random-pool sequences have no catalytic activity, the sequences that are functional emerge due to their immense selective advantage, magnified over multiple selection rounds.

Finally, for in vitro experiments, the difference between “selection” and “evolution” is that evolution requires the introduction of variation after the start of the experiment. In many selection procedures, all variation is present within the random region of the initial pool at the outset of the experiment. Because such procedures merely serve to discard inactive sequences and retain active ones without introducing new variation, they are in vitro selection and not in vitro evolution. In contrast, when variation is introduced intentionally by mutagenic PCR or by restarting the experiment with a partially randomized pool, in vitro evolution is being performed.

**Practical Applications of Nucleic Acid Enzymes**

As the field of artificial nucleic acid enzymes continues to develop, attention has increasingly turned from conceptually oriented experiments to those with more immediate practical utility. Ribozymes and deoxyribozymes have been used for analytical, biochemical, biological, and chemical applications. In addition to several specific applications that are described below, ribozymes and deoxyribozymes have been used for many additional purposes in biochemistry, nanotechnology, and even molecular computation (94).

**In Vitro Engineering of Signaling Ribozymes and Deoxyribozymes as Analytical Sensors**

Analyte detection is one of the primary research motivations in many disciplines. Numerous efforts have focused on nucleic acid enzymes as the basis for sensors (see also the WECB review on biosensors). One of the most productive approaches combines in vitro selection with rational design to create allosteric nucleic acid enzymes. In such a case, catalytic activity is regulated by binding of a small-molecule ligand remote to the catalytic site in a modular fashion (Fig. 6b). Many artificial allosteric ribozymes have been identified, such as several hammerhead ribozyme variants (Fig. 6b) (137–141). Additional work has shown that nuanced behavior such as responsiveness to more than one ligand is possible (142), as is catalysis that is regulated by oligonucleotides (143–145). Laboratory exploration of allosteric ribozymes presaged the discovery of riboswitches, which are naturally occurring RNA regulators of gene expression that generally function by allosteric mechanisms (146). Allosteric deoxyribozymes may be even more practical than their ribozyme counterparts, due to the increased chemical and biochemical stability associated with DNA. A small number of deoxyribozymes have been reported that use various strategies to achieve allosteroy (147–150).

Many systems have been engineered specifically to optimize their practical signaling properties, often on the basis of regulated RNA cleavage activity. Liu et al. have developed deoxyribozyme sensors for metal ions and small organic molecules whose sensing ability is based on fluorescence or colorimetric signals (Fig. 6c) (151–154). Li et al. placed a fluorophore and a quencher close together on the substrate, thereby identifying new RNA-cleaving deoxyribozymes that synchronize fluorescence signaling with catalysis (Fig. 6d) (147) and collectively function at a wide range of pH values (155). These deoxyribozymes are advantageous because they were developed specifically to cleave the fluorophore-containing substrate, and therefore catalysis is optimal with the substrate that is directly relevant to sensing applications. Two research groups (156, 157) have placed ribozymes into arrays that offer the potential for investigating complex biochemical phenomena such as gene expression patterns.
In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes)

Figure 6. Artificial ribozymes and deoxyribozymes as sensors. (a) Schematic depiction of an allosterically regulated nucleic acid enzyme. (b) Examples of allosterically controlled hammerhead ribozymes that cleave RNA in response to binding of the depicted ligands (137, 139); other examples are known (140, 141). The communication modules are highlighted in gray. (c) Example of a signaling ribozyme that is sensitive to adenosine binding with a colorimetric response due to gold nanoparticle dissociation (153). (d) Example of a signaling deoxyribozyme that synchronizes ligand (ATP) binding and catalysis with a fluorescence response (147).

Cleavage of RNA In Vitro and In Vivo by Deoxyribozymes

RNA-cleaving deoxyribozymes have been particularly useful as in vitro laboratory reagents for RNA cleavage (94). Because such deoxyribozymes bind to their RNA substrates via extensive Watson–Crick base pairs (e.g., 10–23 deoxyribozyme in Fig. 2), selectivity for the substrate sequence is an inherent part of each interaction. With the development of a nearly complete collection of related RNA-cleaving deoxyribozymes (158), it is now possible to cleave almost any desired RNA target sequence with an appropriate DNA enzyme. In some experiments, the secondary structure within the RNA target can interfere with binding of the deoxyribozyme. Measures to improve the RNA–DNA interactions by chemical modification of the DNA can enhance the cleavage yield (159), as can inclusion of “disruptor” oligonucleotides that interfere with the target’s secondary structure. DNA is not susceptible to the cellular ribonucleases that quickly destroy RNA. Due to this stability and to other advantages in cost, toxicity, and potency, deoxyribozymes are particularly useful for in vivo mRNA cleavage (see summary in Reference 94), although chemically modified ribozymes may also be employed (160). Such approaches using nucleic acid enzymes are a valuable counterpart to other mRNA-targeting strategies, most notably the application of small interfering RNA (siRNA) or antisense oligonucleotides (see also the WECB reviews on gene silencing techniques and on the chemistry of small interfering RNAs).

RNA Ligation by Deoxyribozymes: Synthesis of Linear and Branched RNA

Several studies from the Silverman research group have identified deoxyribozymes that ligate two RNA substrates. Several deoxyribozymes join an RNA 3′-hydroxyl group with an RNA 5′-triphosphate, forming a native 3′–5′ linear RNA linkage (105). Synthesis of RNA incorporating internal chemical modifications often requires ligation of two or more fragments (see also the WECB review on labeling techniques: nucleic acids). Therefore, RNA ligation deoxyribozymes should be an important alternative to methods such as “splicing ligation” using T4 DNA ligase (161), which often does not work well in particular systems. Other deoxyribozymes create 2′,5′-branched RNA or lariat RNA by mediating the reaction of an internal RNA 2′-hydroxyl group with a 5′-triphosphate (108, 111). In addition to the very high site-selectivity exhibited by these deoxyribozymes (which
is interesting in a fundamental sense), the branched RNA products can be applied to enable biochemical experiments that would otherwise be impossible (162).

**Catalysis of Bimolecular Small-Molecule Reactions by Ribozymes and Deoxyribozymes**

A particularly desirable yet challenging goal is to identify nucleic acid enzymes that mediate bimolecular reactions between two small molecules, i.e., when the substrates are not themselves nucleic acids. For this purpose, no substrate should be covalently attached to RNA or DNA. Very few nucleic acid enzymes have been identified by selection for binding to a transition-state analog (Tables 1 and 2), the resulting catalysis requires no covalent attachment to the substrate. However, in most selections for small-molecule catalysis, one substrate is tethered covalently to the RNA or DNA. This tether can be very short (17, 20, 28, 30), or a long flexible tether such as PEG can be used to mimic a substrate free in solution (25, 34, 35, 73). In either case, most of the resulting nucleic acid enzymes are nonfunctional when the substrate lacks the tether. Indeed, for only two types of in vitro-selected nucleic acid enzymes has catalysis of small-molecule chemistry been achieved with no tethering whatsoever (in trans). First, an RNA capping ribozyme is active when the reacting nucleotide is not attached to the remainder of the ribozyme (17). Second, several Diels–Alder ribozymes (25, 26, 73) are active when the PEG tether that links the anthracene substrate to an oligonucleotide is absent. These limited examples are insufficient to draw general conclusions about what is necessary to achieve bimolecular catalysis of small-molecule reactions by nucleic acid enzymes.

**Perspective on Artificial Ribozymes and Deoxyribozymes**

As noted by Joyce, in the arena of nucleic acid enzymes, “you get what you select for... and sometimes a whole lot more” (163). (Battled-hardened experimentalists might add, “...but if things don’t work, a lot less.”) Currently, the study of artificial ribozymes and deoxyribozymes is just under 20 years old and about to leave behind its teenage years. The original motivation to study nucleic acid enzymes was for the fundamental purpose of exploring the scope of nucleic acid catalysis. Artificial ribozymes and deoxyribozymes have already offered substantial insight into such catalysis, and we may expect additional success from ongoing research in this area. As more types of reactions well beyond nucleic acid cleavage and ligation are increasingly the goal of artificial nucleic acid enzymes, including efforts toward catalysis of small-molecular reactions, a practical and important question is to decide what are the most appropriate reactions to be targeted by RNA and DNA catalysis? Which reactions truly need nucleic acid catalysts, and how can nucleic acid enzymes be developed as practical chemical reagents? In contrast, what reactions are best performed by other catalysts such as protein enzymes or more traditional organic and organometallic catalysts? The responsibility falls to those of us working actively on selection experiments to develop and demonstrate the synthetic utility of ribozymes and deoxyribozymes for the broader chemical biology community. In parallel with these considerations, elegant experiments have already shown that practical applications such as analytical sensing can be accomplished in spectacular fashion using nucleic acid enzymes. Such techniques will surely be refined and extended in the near future.

**References**


In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes)


117. Carmi N, Shultz LA, Breker RR. In vitro selection of self-cla

125. Bock LC, Grif

126. Thum O, Jager S, Famulok M. Functionalized DNA: a new repli

127. Lin CH, Patel DJ. Structural basis of DNA folding and recogni

147. Mezi SH, Liu Z, Brennan JD, Li Y. An efficient RNA-cleaving DNA enzyme that synchronizes catalysis with fluorescence sig

153. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

154. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

155. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

156. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

157. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

158. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

159. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

160. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

161. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

162. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

163. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

164. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

165. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

166. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

167. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

168. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col

169. Liu J, Lu Y. Adenosine-dependent assembly of aptazyme-functionalized gold nanoparticles and its application as a col
In Vitro Selection and Application of Nucleic Acid Enzymes (Ribozymes and Deoxyribozymes)


Further Reading


See Also

Natural Ribozymes
Origins of Life: Emergence of an RNA World
Crystallographic Techniques: Nucleic Acids
Chemical Libraries for Biologically Active Small Molecules, Screening of Biosensors
Gene Silencing Techniques
Chemistry of Small Interfering RNAs
Labeling Techniques: Nucleic Acids
Advanced Article

Locked Nucleic Acids

Zenia M. Starling and Troels Koch, Santaris Pharma A/S, Hørsholm, Denmark
doi: 10.1002/9780470048672.wecb516

The intensive research and development of antisense oligonucleotides in the past decades has refined our knowledge of this class of molecules, and they are now recognized as potent reagents for gene-target validation and therapeutic uses. Within this field, a very promising class is locked nucleic acid (LNA). In recent years, the number of applications for LNA in basic research, molecular diagnostics, and therapeutics has expanded continuously. In this article, the general biophysical, biochemical, and pharmacological properties of LNA will be summarised. The ability of LNA oligonucleotides to target both coding and noncoding RNAs effectively and specifically via different modes of action will be discussed. Several LNA oligonucleotides have completed preclinical toxicology and are now tested in clinical trials. The status of one drug candidate will be presented briefly at the end of the article.

Background

Native oligonucleotides cannot be used as active antisense reagents and therefore in the history of antisense, a variety of chemically modified nucleotides has been developed to improve properties such as target affinity, nuclease resistance, and pharmacokinetics of the oligonucleotides. The first generation of chemically modified antisense drugs was the phosphorothioates (PSs). The phosphorothioate internucleoside link showed improved nuclease stability and bioavailability, but it resulted in lower RNA affinity. To compensate for that, the length of PS was kept rather long and 18–20-mers were frequently used. This is significantly longer than the antisense principle requires for specific mRNA recognition. The ideal antisense reagent ought not to be longer than required for specific mRNA recognition which is dependent on sequence typically 12–16-mers. However, PSs have been used extensively to inhibit mRNA targets in vitro, but the combination of a relatively low target affinity and limited nuclease stability reduce the in vivo potential of PS via specific antisense mechanisms. On the contrary, the therapeutic significance of PS is high in combination with other chemical modifications (see below) or via their mode of action other than classical antisense.

The second-generation antisense drugs are represented classically by the nucleotides where the 2′ position has been modified. Examples of this class are the 2′-O-Me, 2′-O-MOE, and 2′-F-modified nucleotides (Fig. 1). By incorporating these modifications into oligonucleotides, the affinity was improved—compared with the PSs—but to get sufficient potency and efficacy, the length of the second-generation antisense drugs was unchanged compared with the first generation (18-20-mers) drugs illustrating that the second-generation antisense oligonucleotide did not increase the affinity adequately.

Locked Nucleic Acid (LNA) was presented in 1998. LNA represents a true member of the third generation of antisense drugs. When LNA nucleotides were incorporated into oligonucleotides, they induced the highest binding affinity for both complementary RNA and DNA reported in the field (2). This opened completely new perspectives for the development of antisense drugs. The remaining part of this article will focus on LNA and its therapeutic aspects.
Locked Nucleic Acids

Biophysical Properties of LNA

LNA oligonucleotides obey Watson-Crick hydrogen bonding rules and form right-handed helices, and actually the most intriguing property of the LNA nucleotide is that it fits perfectly in the Watson-Crick framework. LNA nucleotides can be included in any combination with DNA/RNA nucleotides, and any analog thereof that fits within the Watson-Crick framework. The immediate property described for LNA was an extraordinary high affinity (2)—a property later confirmed in numerous papers and recently reviewed (7). When LNA monomers are incorporated in either DNA or RNA each residue adds to affinity in an additive manner. insertions of few LNA-ribonucleotides lead to increase of melting temperature (T_m) of 6-7 °C per modification; however, the largest increase is observed when LNA nucleotides are included in PS, which provides a record high increase of 9-10 °C per modification (8). This feature of LNA is unique because PS modification is known to reduce the affinity of DNA and 2'-modified oligonucleotide analogs. However, when PSs are used as internucleoside link in LNA oligonucleotides affinity is almost not affected and therefore, the affinity of LNA diesters and LNA PSs is almost the same.

The affinity increase per LNA modification is the greatest when only few LNA nucleotides are incorporated into oligonucleotides because LNA nucleotides direct the conformation of nearby desoxyribose-nucleotides to attain a greater degree of the high affinity North conformation. Therefore, the affinity increase is driven by both the LNA nucleotides and the North-directed DNA nucleotides. Fully modified LNA oligonucleotides—with no RNA nucleotides—do have the highest affinity, but the affinity per LNA modification is lower. This unique property of LNA is called "structural saturation." Structural saturation is a very important property of LNA because only a few LNA modifications can transform the entire oligonucleotide into a high-affinity structure.

The high affinity of LNA makes it possible to design shorter antisense oligonucleotides with retained high potency, and it means that the "hit rate" of LNA antisense oligonucleotides designed to complement different sequences along the mRNA are greatly compared with traditional antisense chemistries (9). This finding indicates that the lead identification process in drug discovery becomes faster and that fewer LNA antisense oligonucleotides are needed to find highly potent mRNA binders.
LNA Oligonucleotide Designs

LNA and most LNA-analogs can be mixed in any combination with nucleic acids and analogs. Depending on the application of LNA, the design may be divided in five categories: mixmer, gapmer, headmer, tailmer, and fully modified (Fig. 3, a–e). In the mixmer, the LNA residues are dispersed along the sequence of the oligonucleotide (a), whereas in the gapmer two continuous LNA segments are attached in the 5′- and 3′-ends of a central nucleic acid fragment (b). A headmer is a continuous nucleic acid fragment with a continuous LNA segment in the 5′-end (c) and vice versa for a tailmer (d). The fully modified LNA oligonucleotide is self-explanatory (e) (Fig. 3).

The length and specific design, however, may vary depending on the intended use and on how important the affinity, nuclease resistance, and the RNase H/RISC recruitment (see below) is for the proposed application. In the classic designs described above, the total length of LNA oligonucleotides is 16 mers, and the oligonucleotides are fully PS-modified.

In conclusion, freedom of design offered by LNA has enabled the development of LNA oligonucleotides with the ability to target both coding and noncoding RNAs effectively and specifically in vivo, using different modes of action, such as target cleavage, steric block of translation and inactivation by sequestration.

Biochemical Properties of LNA

Nuclease resistance

One important aspect of antisense technology is the stability of antisense agents against extracellular and intracellular endonucleases and exonucleases. A series of studies has investigated the serum and nuclease stability of LNA-modified oligonucleotides, and the studies have shown that incorporation of LNA segments into oligonucleotides increases their stability against nucleases. It is the number and the position of the LNA residues as well as insertion of other stabilizing entities—such as PSs—that determines the rate of degradation of the oligonucleotides. For example, a single LNA residue at the 3′-end is not sufficient to increase stability substantially against exonucleolytic attack, whereas one LNA residue at the 3′-penultimate position or two LNA modifications at the ultimate position of the 3′ end efficiently stabilize against 3′-endonucleases (10, 11). An effective way to secure exonucleolytic protection is obtained with the gapmer design, but the central DNA segment is sensitive to endonucleolytic activity. If the central DNA segment becomes longer than four nucleotides, then cleavage starts to occur and the cleavage rate increases as the DNA segment increases.

A preferred way to protect degradation of the central DNA segments is to use PS internucleoside links. The PS link is easy to introduce during LNA synthesis and it serves as a substrate for RNase H (see below). A series of fully PS-modified gapmers with gap-sizes that range from 9 to 11 PS links have been examined in human and rat plasma and they showed no sign of endonucleolytic activity (7). Fully modified LNA is essentially resistant to exonucleolytic and endonucleolytic degradation either with fully diester or PS internucleoside links (11).

RNase H recruitment

It is well established that oligonucleotides targeting mRNA are most potent if they are designed to recruit RNase H. Under normal circumstances, this enzyme cleaves mRNA hybridized to a DNA complement. Nevertheless, because LNA nucleotides adopt an RNA-like structure, they cannot recruit RNase H. However, LNA gapmers, which contain a stretch of DNA—or PS—nucleotides, can recruit RNase H. More specifically, several studies have investigated various LNA/DNA gapmers and concluded that a gap of six DNA nucleotides is necessary for initiating RNase H activity, and that a gap of seven DNA nucleotides allows for high RNase H activity (11–14). The conclusion is that antisense LNA oligonucleotides can be designed to elicit RNase H activity while still containing LNA monomers for improved binding and target accessibility (15). Furthermore, it has been shown that the highest activity was observed if the entire LNA gapmer was phosphorothioate (not just the DNA segment) (11).

LNA Applied In Vitro

Most antisense experiments made with LNA have been focused mechanistically on mRNA inhibition by RNase H recruitment (7). Expression studies in a wide variety of human epithelial and cancer cell lines are made and the overall picture of these experiments is that LNA antisense oligonucleotides are very potent with IC₅₀ values for mRNA inhibition obtained frequently in the low to sub nanomolar range—and actually more potent than competing oligonucleotide analogs (7).

The potency of LNA oligonucleotides is related to the design of the oligonucleotide and to the target site. Several studies have found that the gapmer design is slightly more potent than the mixmer design, in which inhibition of mRNA in the translational start site is required (7). However, when the target site is in the coding region, the gapmer design is superior (16).
This finding is caused by the fact that a high-affinity LNA oligomer can prevent ribosome binding to mRNA and thereby stop translation initiation; but once the ribosome is bound and initiates translation, mRNA has to be cleaved by RNase H to stop its action (17). LNA gapmers have been used intensively in cellular-based assays to inhibit mRNA expression. This work has been reviewed recently (7), and the conclusion is that LNA gapmers ranging from 12–16mers inhibit gene expression with IC₅₀ values in the low to sub-nanomolar range (18).

Another mechanistic action for mRNA inhibition is a non-RNase H based mechanism in which the LNA antisense oligonucleotide acts via very tight binding to the mRNA and thus, interferes with its additional processing. This mechanism is called "steric block." Steric block includes prevention of ribosome binding by blocking the 5'-UTR that acts as a "road block" during translation by binding to the message or by redirection of splicing sites to alter the production of splice variants (19). When antisense oligonucleotides act according to this mechanism, the mRNA remains intact. LNA oligomers have been used effectively as steric blocks to inhibit the expression of several therapeutically interesting targets for instance inhibition of viral replication (11,20) and redirect splicing in samples from patients with Duchenne muscular dystrophy (21,22).

Pharmacology

Currently, pharmacological activity of LNA antisense oligonucleotides has been demonstrated in many different tissues (liver, kidney, colon, jejunum, small intestine, lung, spleen, and brain) and species (mice, rats, and monkeys) using a variety of administration routes (subcutaneous infusion, intravenous injection, intraperitoneal injection, and direct injection into brain) (7). Using different designs of LNA oligonucleotides, it has been possible to target both mRNAs and noncoding miRNAs effectively (see below), which causes either a reduction in protein synthesis, a shift in the splicing pattern, or inactivation by biochemical sequestration.

The pharmacological effect of LNA oligonucleotides depends on biodistribution and cellular uptake of the LNA and LNA-analog oligonucleotides. Biodistribution studies have shown equal uptake of LNA and ω-LNA, except in kidneys, where the uptake of ω-LNA is greater than for LNA. Amino- and thio-LNA have increased uptake in liver, whereas the amino-LNA has the broadest distribution showing significantly greater uptake in heart, lung, muscle, and bone compared with other analogs (23). Once absorbed by tissues, LNA oligonucleotides are taken up by cells by internalization, and a strong relationship between tissues uptake and endogenous effects have been demonstrated (7).

The first in vivo antisense experiment with LNA-antisense involved two different 15-mer LNA designs that target the delta 16-mer LNA, which targeted the RNA polymerase II gene product, inhibited tumor growth in mice at low doses (less than 5 mg/kg/day) and was tolerated well (24). Since these early studies, the pharmacological properties of LNA have been documented in numerous publications. Many improvements have been made, and the LNA oligonucleotide design and how LNA oligonucleotides interact with the intracellular molecular machinery are better understood. The conclusion of these works is that pharmacological relevant doses of LNA range typically from 0.5-5 mg/kg, and that the oligonucleotides are generally well tolerated (7).

An interesting observation is that the pharmacological potential can be increased with reduced size of the LNA oligonucleotides. A shorter design (oligonucleotides shorter than 16-mers) has been used to inhibit ApoB-100 expression in liver and jejunum. ApoB-100 is the major cholesterol-carrying protein in low-density lipoprotein (LDL) and plays a central role in cholesterol metabolism and atherosclerosis. The study showed that 12-14mers reduced the target expression with 80-90% where a reduction of only 30-40% was seen with the 15 and 16mers. The reduction in ApoB-100 expression was followed almost linearly with reduction in plasma cholesterol levels (25). These data suggest that use of shortmers may become a new approach for systemic gene targeting with significantly improved pharmacological effects.

Inhibition of MicroRNA

MicroRNAs (miRNAs) are an abundant class of short endogenous noncoding RNAs that act as posttranscriptional modulators of gene expression. To date, more than 5000 miRNAs have been identified in invertebrates, vertebrates, and plants according to the miRBase miRNA database release 10.0 in August 2007. Growing evidence shows that miRNAs exhibit a wide variety of regulatory functions and exert significant effects on cell growth, development, and differentiation.

A major obstacle to specific and sensitive detection of mature miRNAs in animals is their small size (~19-23 nucleotides). Endlabeled LNA-probes, in which every third position is an LNA residue, have been used in Northern blot analysis to detect miRNAs with high specificity and showing at least 10-fold increased sensitivity compared with DNA probes (26), which is a major achievement in miRNA research. Furthermore, LNA probes with uniform Tₘ values of 72°C have been used in high-sensitivity arrays for miRNA expression profiling (27).

Such arrays allow discrimination between single nucleotide variations as demonstrated by the specific discovery of related miRNA family members. LNA probes have been used successfully to enhance the sensitivity and specificity of in situ hybridizations to miRNAs in zebrafish (28,29) and in human brain (30).

The expanding list of human miRNAs along with their highly diverse expression patterns and high number of potential target miRNAs suggest that miRNAs are involved in a wide variety of human diseases. Perturbed miRNA expression patterns are reported in many human cancers (31), and more than 50% of the human miRNA genes are located in cancer-associated genomic regions or at fragile sites (32).
Several studies have shown that LNA oligonucleotides can be used to inhibit miRNA function specifically and effectively (LNA-anti-miRs). One study showed inhibition of miR-21 activity with an LNA gapmer that leads to induction of apoptosis in glioblastoma cell lines, which illustrates that aberrantly expressed miRNAs can function as oncogenes (33). Another study showed inhibition of the bantam miRNA in Drosophila by a complementary LNA-modified oligonucleotide resulting in sequence specific derepression of Hid protein synthesis in cell culture (34). Recent findings by Elmén et al. (35) show that high-affinity antagonism of the liver-specific microRNA 122 (miR-122) in mice, by steric hindrance using LNA-anti-miRs, can result in potent functional inhibition. This finding suggests that oligonucleotides composed of LNA may be valuable tools for identifying miRNA targets in vivo and for studying the biological role of miRNAs and miRNA-associated gene-regulatory networks in a physiological context. In addition, the high metabolic stability of LNA-anti-miRs, caused in part by increased nuclease resistance, their small size and apparent lack of acute toxicity or changes in liver morphology (35), imply that LNA-anti-miRs may be well suited as a novel class of potential therapeutics for disease-associated microRNAs.

Three other LNA-based compounds are at preclinical development studies. ENZ3042 is a specific inhibitor of survivin, which plays a vital regulatory role in both apoptosis and cell division. Survivin is overexpressed heavily in many cancers and in newly formed endothelial cells engaged in angiogenesis but almost absent in normal adult differentiated tissue. Clinically, survivin expression is associated with poor prognosis, increased cancer recurrence, and resistance to therapy. The compound therefore has the potential to improve the response to therapy in a wide variety of malignant solid tumors. SPC3833 is in preclinical development and it is designed to effectively affect the expression of apolipoprotein B100 (ApoB100). The drug effectively down-regulate ApoB100 synthesis leading to a reduction of VLDL secretion from the liver. The down-stream effect of this is that SPC3833 efficiently lowers LDL and cholesterol levels (37). Finally, SPC3649 is the first LNA oligonucleotide developed as a microRNA inhibitor. SPC3649 is a lead candidate targeted to the liver-specific microRNA-122. The LNA anti-miR is developed for treatment of Hepatitis C and hypercholesterolemia and is expected to commence human clinical pharmacology and safety studies in 2008.

LNA Drugs in Clinical and Preclinical Development

In 2006, the LNA oligonucleotide SPC2996 completed its first human clinical trial, which was a Phase I/I study in patients with chronic lymphocytic leukemia (CLL). SPC2996 acts by inhibiting the synthesis of B-C1, which is a key sensor protein that protects cells against apoptosis. The protein is expressed in most cancers including CLL, and high expression levels have been firmly correlated with low response rates, resistance to chemotherapy, faster time to relapse, and shorter survival time (36). Thus, downregulation of B-C1 expression provides an attractive means by which CLL and other cancers can be resensitized to natural apoptotic stimuli and to chemotherapeutic agents that provoke apoptosis.

The results of this study are encouraging for such an early stage trial. All patients who received the top dose of the drug had a rapid reduction in the number of cancerous white blood cells in circulation, and expression of B-C1 was reduced with 33% as measured by qPCR in total blood RNA.

The results of this first human clinical trial provide the basis for additional development of SPC2996, and it underlines that repeated doses of the drug can be administered safely.

The second LNA oligonucleotide that entered a human clinical Phase I/I trial in patients with solid tumors is ENZ2968—a potent inhibitor of HIF-1α. HIF-1α serves as the key sensor of cellular hypoxia in response to which it transcriptionally upregulates a host of genes that play important roles in promotion of cancers, which include angiogenesis, apoptosis, cell migration, and metastasis. HIF-1α is essentially undetectable in normal cells, but it reaches high intracellular concentrations in a variety of cancers in which it is correlated strongly with poor prognosis and resistance to therapy.

Three other LNA-based compounds are at preclinical development studies. ENZ3042 is a specific inhibitor of survivin, which plays a vital regulatory role in both apoptosis and cell division. Survivin is overexpressed heavily in many cancers and in newly formed endothelial cells engaged in angiogenesis but almost absent in normal adult differentiated tissue. Clinically, survivin expression is associated with poor prognosis, increased cancer recurrence, and resistance to therapy. The compound therefore has the potential to improve the response to therapy in a wide variety of malignant solid tumors. SPC3833 is in preclinical development and it is designed to effectively affect the expression of apolipoprotein B100 (ApoB100). The drug effectively down-regulate ApoB100 synthesis leading to a reduction of VLDL secretion from the liver. The down-stream effect of this is that SPC3833 efficiently lowers LDL and cholesterol levels (37). Finally, SPC3649 is the first LNA oligonucleotide developed as a microRNA inhibitor. SPC3649 is a lead candidate targeted to the liver-specific microRNA-122. The LNA anti-miR is developed for treatment of Hepatitis C and hypercholesterolemia and is expected to commence human clinical pharmacology and safety studies in 2008.

Final Remarks

The remarkable hybridization properties of LNA, both with respect to affinity and specificity, position LNA as an enabling molecule for molecular biology research, diagnostic applications, and therapeutics. The fact that LNA nucleotides can be mixed in any combination with DNA/RNA residues and phosphorothioate internucleoside linkages is a very attractive property. Of particular relevance for the therapeutic aspects of LNA is the combination of LNA/DNA residues and PS residues. These “chimeric” oligonucleotides exert their action via the individual components: The LNA residue mediates very high target affinity and stability, and the PS’s provide the necessary PK properties and cellular uptake. We believe that the combination of these properties in one oligonucleotide—such as we have described here—provides enabling properties of LNA as RNA inhibitor.

References


See Also

MicroRNA (miRNA)
mRNA Localization and mRNA Levels, Control of
Oligonucleotide-Directed Inhibition of Gene Expression
Drug Discovery
DNA-Based Structures, Applications of Chemical Biology
Nucleosides in RNA, Modified

N. Dinuka Abeydeera and Christine S. Chow, Wayne State University, Detroit, Michigan
doi: 10.1002/9780470048672.wecb688

Natural and unnatural modified nucleosides play important roles in chemistry and biology. Over 100 different functional group modifications occur on natural nucleosides, which range from sugar methylations to base isomerization reactions. Unnatural modifications are also broad in their range of possible structures and applications. Both types of modifications have been incorporated site specifically into RNAs to understand their biological roles or to be used as biophysical, structure, and mechanistic probes. More recently, modified RNAs have been applied as therapeutic agents. This article summarizes the approaches toward generation of modified nucleosides and their incorporation into RNA using chemical, biochemical, or combined (semisynthesis) approaches. Some examples of applications with modified RNAs will also be discussed.

Background to Modified Nucleosides in RNA

The expansive array of RNA functions discovered to date is highly dependent on the ability of RNA to fold into unique structures, undergo large conformational changes, or participate in specific interactions with macromolecules (e.g., RNA and proteins), metal ions, and small organic ligands. Although the core nucleotide structures are sufficient for many biological roles of RNAs, over 100 natural modifications in RNA have been discovered over the past six decades, which seem to modulate RNA function (1, 2). In addition, a wide variety of unnatural nucleosides has been synthesized and incorporated into RNA. These modifications often provide RNA with unique properties, such as expanded or altered hydrogen-bonding; van der Waals, base-stacking, or electrostatic interactions; as well as unique metal-binding sites, chemical reactivity, or fluorescent characteristics. The synthetic routes to generate modified nucleosides vary considerably, as do the routes for incorporation of the nucleosides into small oligonucleotide fragments or large, full-length RNAs. The syntheses of modified RNAs serve many purposes, such as developing a better understanding of the biological roles of natural modifications, generating tools for studies of RNA structure and function, and producing novel therapeutic agents.

The types of nucleoside or nucleotide modifications range from alterations of the base component to changes in the sugar or phosphate moieties. The natural modifications include pseudouridylation (generation of a C-glycoside), methylation, deamination, thiolation, and alkylation (1, 2). Hypermodifications also exist in which multiple components of the nucleotide are altered. Some unnatural modifications include fluorescent bases, such as 2-aminopurine, or novel nucleobases to expand the genetic code. This review summarizes some of the most common approaches taken by chemical biologists toward generation of modified nucleosides and their incorporation into RNA using chemical, biochemical, or combined (semisynthesis) approaches. It is possible to use these techniques to construct RNA molecules with a wide variety of functional groups at highly specific positions. Although the applications of modified RNAs are numerous and far reaching, it is not possible to give a comprehensive view in this short article; therefore, representative or recent examples of applications have been selected for discussion.

Synthesis of Modified Nucleosides

The chemical synthesis of novel nucleosides or those that mimic natural modifications will provide biologists with new probes for RNA function. Thousands of modified bases have been described in the literature, which range from synthesis of both natural and unnatural nucleosides to their incorporation into short oligonucleotides and full-length nucleic acids. Numerous different approaches have been taken to generate modified nucleosides (3–6), which clearly cannot all be described in this brief review. Therefore, several examples have been selected to highlight some general strategies for preparing nucleoside analogs. Traditional methods for modified nucleoside synthesis...
Nucleosides in RNA—Modified

Nucleosides—coupling strategies

Most nucleosides contain N-glycosidic linkages (i.e., linkages through the base nitrogen atom). The N-glycosides can be generated by reactions with a suitably protected and activated sugar component and a silylated base (purine or pyrimidine) fragment. The most commonly used reaction in N-glycoside synthesis is the Hillbert-Johnson-type sugar-base condensation in the presence of a Friedel-Crafts catalyst, such as SnCl₂, which can be used to produce pyrimidine, purine, as well as sugar and/or base-modified nucleosides (Fig. 1a, b, i). Several improvements to these methods have been made since the original reports, which include the use of microwave-assisted reactions to generate nucleoside libraries (9). Maydanovych et al. (10) have used similar coupling strategies to generate adenine analogs for the study of RNA editing, and they reported the detailed procedures in a recent review article. The less-common C-glycosides contain a C-C linkage between the carbohydrate moiety and the heterocyclic base. A wide range ofaryl C-glycosides can be generated by using the Heck coupling reaction, as recently reviewed by Wellington and Bennet (11). Pseudouridine (Fig. 2a, c), which is the most abundant natural modification found in RNA, is a C-glycoside generated by posttranscriptional isomerization of uridine residues at specific sites along the oligonucleotide chain. Although numerous groups have reported on the synthesis of pseudouridine, Hanesian, and Machaalani’s stereocentered synthesis is the most convenient and efficient route to date (12). Their synthesis involves a reaction between a suitably protected D-ribose aldehyde and lithiated pyrimidine, followed by sequential ring-opening and ring-closing steps to produce either α or β pseudouridine in high yields. The ratio of α/β anomers is controlled by the presence of L-selectride and zinc chloride.

Nucleoside modifications

An alternative route to modified nucleosides involves derivatization of the standard nucleosides, guanosine, adenosine, cytidine, and uridine. Among the large repertoire of modifications identified to date, the methylated nucleosides are highly abundant and occur frequently in the functionality important regions of RNA (1, 2). Methylation typically occurs on the sugar or base portion of the nucleoside. Although many approaches have been taken to methylate nucleosides, site-selective introductions of a methyl group can be challenging. A recent report by Höbarth et al. (13) showed the synthesis of a wide variety of methylated ribonucleosides, which included 1-methylguanosine (m2G), 1′-methylguanosine (m1E), 2′-methylguanosine (m2G), 1-methylinosine (m1I), 3-methyluridine (m3U), N²-methylcytidine (m2C), N²-methyladenosine (m2A), and N²,N⁶-dimethyladenosine (m2A). Guanosine methylation at the amino nitrogen to yield m1G can be achieved by treatment of guanosine with NaI in DM SO followed by addition of methyl iodide (Fig. 2a, steps i, ii) (13). Methylations at the N⁷ position on guanosine to yield m2G and m22G can be achieved following sugar and O³ protection and generation of a 2-fluoro intermediate. The 2-fluoro intermediate is treated with CH₃NH₂ or (CH₃)₂NH to yield m2G or m22G, respectively (Fig. 2a, steps iii–viii) (13). Alternatively, m2G can be generated through reduction of a suitably protected p-thiocresol intermediate (Fig. 2a, steps iii, viii-x) (14). Methylated adenosine analogs can be obtained using isonine as a starting material. Generation of the 6-chloro intermediate allows for conversion into m2A or m22G with CH₃NH₂ or (CH₃)₂NH, respectively (Fig. 2b, steps xi, vii, viii, ix) (13).

Methylated cytidine analogs can be obtained from 5-methyluridine (m1U) as the starting material (15). Suitable protected m1U is reacted with P₂S₅ to yield a 4-thio intermediate that is then converted to 5-methylcytidine (m2C) during treatment with NH₃ (Fig. 2c, steps xiii–xiv). Uridine can be used as a starting material for the synthesis of N⁴-methylcytidine (m⁴C), N⁴,2′-O-dimethylcytidine (m⁴Cm), and 2′,O-methyl-cytidine (mC) (Fig. 2d, steps xv–xx). A tetrazole intermediate is generated (16, 17) that can be converted to m2C or m6Cm with the appropriate methylating agent. The Z-2′-O-methyl group is installed through an anhydronucleoside intermediate and opening with Mg(OH)₂, in methanol (18).

Methylated uridine analogs can be generated using related methods, such as treatment of uridine with NaI and MeI (Fig. 2e, steps i, ii) (12) to give m2U or conversion to 2′-O-methyluridine (mU) through an anhydronucleoside intermediate and opening by the appropriate nucleophile (18).

Several unnatural nucleoside phosphate modifications (Fig. 3) have been generated, which include phosphorothioate, phosphororoamidate, phosphorothioate, and methylphosphonate derivatives. The syntheses of these modifications have been reviewed by Verma and Eckstein (19). A modification of those procedures can be used for the generation of phosphoroxyenolate RNAs (20).

Convertible nucleotides

Convertible nucleotides are modified nucleotides that contain reactive functionalities. They are incorporated into RNA at specific locations and allow for convenient posttranslational introduction of chemical probes, isotope labels, or cross-linking agents, such as disulfide linkages. In this approach, the nucleotide derivative contains a leaving group on its base moiety and is incorporated into the oligonucleotide at a defined location (21). Once the preparation of the desired precursor RNA is completed, the oligonucleotide is treated with an appropriate nucleophile (e.g., amine), which can displace the leaving group and become attached to the specific base moiety (Fig. 3). As will be discussed in more detail later, the convertible nucleosides have many applications, and they can be used to control the RNA structure and function, as well as to regulate RNA conformational switching and dynamics.

Enzymatic approaches to nucleoside triphosphate synthesis

Nucleosides can also be prepared by using enzymatic reactions. Modified nucleoside triphosphates are useful as precursors for
the synthesis of RNA through a modified DNA template and T7 RNA polymerase. The chemoenzymatic synthesis of nucleoside triphosphates was recently reviewed by Wu et al. (22). A recent example is the synthesis of β-azaguanosine (BetaGuo), which is a guanosine analog with N8 in place of C8 that displays fluorescent properties, and the conversion to the corresponding triphosphate (BetaGTP) (23). The strategy for BetaGTP synthesis involves the use of enzymes from the pentose phosphate and nucleotide salvaging pathways, many of which are commercially available. The presence of the triphosphate allows for the incorporation of BetaGuo into RNA using a DNA template and T7 RNA polymerase.
Nucleosides in RNA, Modified

[Image: Figure 2 showing nucleoside base, sugar, and phosphate modifications.]

**Figure 2**  Nucleoside base, sugar, and phosphate modifications are shown. The syntheses of methylated guanosine (a), adenosine (b), cytidine and uridine (c and d) nucleosides are highlighted. Reaction conditions: (i) NaH, DMF, rt, 2 h; (ii) MeI, rt, 5 h (13); (iii) Ac₂O, DMAP, Et₃N, CH₃CN, 0.5 hour; (iv) NPE, Ph₂P, DEAD, dioxide, rt, 2 hours; (v) HBr, NaN₂O₃, acetone/water, −20°C to rt, 3 hours; (vi) Br₂CH₂NH₂, ethanol, 7–12 hours; (vii) CH₃₂NH in ethanol/water, rt, 3–9 hours (3); (viii) p-thiocresol, CH₂CO₂H, HCHO, ethanol, reflux, 4 hours; (ix) NaN₃, DMSO, 100°C, 1 hour; (x) H₂ in CH₃OH, rt, 16 hours (14); (xi) Ac₂O, DMF, pyridine, rt, 4 days; (xii) (xiii) benzoyl chloride, pyridine, 55°C, 3 days; (xiv) F₂, pyridine/water, reflux, 4 hours; (xv) N₂H₅, CH₃OH (sealed), 100°C, 24 hours; (xvi) tetrazole, TsCl, diphenyl phosphate, pyridine, rt, 36 hours (16, 17); (xvii) CH₃NH₂ + Cl⁻, KOH, (C₂H₅)₃N, CH₃CN/water, rt, 24 hours; (xviii) (PhO)₂CO, NaHCO₃, DMSO, 80°C, 3 hours; (xix) Mg(OCH₃)₂, CH₃OH, reflux, 3 hours (18); (xx) NaHCO₃, KOH, (C₂H₅)₃N, CH₃CN/water, rt, 24 hours. e) Variations of nucleoside phosphate modifications are shown.
Figure 2 (continued)
Nucleosides in RNA, Modified

Single site incorporation

\[ \text{Nu (amine, probe, etc.)} \]

Multiple site incorporation

\[ \text{Nu-S-S-Nu} \]

**Figure 3** The convertible nucleoside approach is shown (21). On the left, a single convertible nucleoside that has been incorporated site-specifically into the RNA is treated with a nucleophile to displace the leaving group, X. On the right, an RNA with multiple convertible nucleosides can be used to generate a cross-linked analog.

Incorporation of Modified Nucleosides into RNA

Through a combination of chemical and biochemical approaches, modified nucleosides can be incorporated into RNA either globally (i.e., at many sites) or in a site-specific manner. Depending on the particular application of interest, either a global or site-specific incorporation strategy may be desirable.

**Enzymatic incorporation**

Modified nucleotides in RNA can be generated by using site-specific enzymes. Several naturally occurring modifying enzymes have been isolated and used to modify either small oligonucleotides or full-length RNA templates at specific sites (Fig. 4a) (24). The enzyme reactions are useful for understanding the biological roles of the modified nucleosides; however, they are not as useful for generating modified RNAs for biophysical studies or as therapeutic agents because the reactions do not go to completion, and the modified RNAs are often difficult to separate from the unmodified RNAs. RNAs in bacteria and eukaryotes are modified by different pathways. Individual modifications in bacterial RNAs are mediated by one of many unique protein enzymes that are typically site specific and target particular structures or sequences of RNA (24). The most common enzymes are methyltransferases and pseudouridine synthases. Because of the greater number of modifications in RNAs from higher organisms, an alternative pathway of insertion is used that employs fewer site-specific enzymes. Noncoding RNAs known as small nucleolar RNAs (snoRNAs) combine with a set of proteins to generate a small nuclear ribonucleoprotein (snRNP) complex that guides the site of enzyme modification on the target RNA (25). The discovery of snoRNAs has been extremely important for understanding the biological roles of RNA modifications, but the use of snoRNPs is not as convenient for generation of site-specifically modified RNAs for biophysical studies.

RNA segments can be modified at their 3'-ends by using a bisphosphate analog of the modified nucleoside and T4 RNA ligase (Fig. 4b) (26). This reaction is general and essentially can be applied to any RNA sequence and modification type that can be converted into a bisphosphate. The 3'-modified RNA segment can then be dephosphorylated with calf intestinal phosphatase and ligated with another 5'-phosphorylated RNA fragment to generate larger RNAs with site-specific modifications (Fig. 4b) (26). Methods for ligation include the use of a DNA "splint" with either T4 RNA ligase or T4 DNA ligase (27, 28). The presence of the complementary DNA fragment promotes formation of the desired RNA product and minimizes dimer or circular RNA formation.

A second enzymatic approach employs T7 RNA polymerase and a DNA template that directs the addition of specific ribonucleotides according to standard Watson-Crick base-pairing rules. This transcription method can be used to generate RNAs of any length and sequence. The T7 RNA polymerase will often accept modified nucleoside triphosphates. The method is useful for global modification, but typically it does not allow for site-specific incorporation of modified nucleotides. This limitation has been overcome in several cases, in which nonstandard nucleotides are used to direct the insertion of modified nucleotides through alternative base-pairing schemes (Fig. 4c) (29).

**Chemical approaches**

The chemical synthesis of RNA has the distinct advantage of allowing for site-selective incorporation of modified nucleosides into the oligonucleotide chain. The early phosphodiester method developed by Khorana (30) was followed by the H-phosphonate and phosphoramidite approaches (31). Although the phosphodiester approach has distinct advantages over phosphoramidite synthesis, it is not as widely used. Therefore, the focus of this section will be on the more commonly employed methods.
Nucleosides in RNA, Modified

(a) Modifying enzymes (E) convert standard nucleotides (N) into modified versions (N-X) (e.g., methylation, pseudouridylation) (24). Eukaryotes require a snoRNP complex along with the modifying enzyme to direct the site of modification, whereas bacteria use unique enzymes to modify each site (25). (b) T4 RNA ligase is used to add a single modified nucleotide (X) in the form of a bisphosphate (pXp), or to ligate a modified fragment to another piece of RNA (26, 27). (c) Enzymatic incorporation employs T7 RNA polymerase and a modified DNA template, in which deoxynucleotide Y makes a unique base pair with ribonucleotide X (in the form of XTP) to incorporate the modified nucleotide at a single, specific site on the RNA chain (29). (d) The steps for incorporation of modified nucleotides by solid-phase oligonucleotide synthesis are shown (32–34). (e) The most common protective groups for the phosphoramidite building blocks are highlighted.

Figure 4: The site-specific incorporation of modified nucleotides into RNA is depicted. a) Modifying enzymes (E) convert standard nucleotides (N) into modified versions (N-X) (e.g., methylation, pseudouridylation) (24). Eukaryotes require a snoRNP complex along with the modifying enzyme to direct the site of modification, whereas bacteria use unique enzymes to modify each site (25). b) T4 RNA ligase is used to add a single modified nucleotide (X) in the form of a bisphosphate (pXp), or to ligate a modified fragment to another piece of RNA (26, 27). c) Enzymatic incorporation employs T7 RNA polymerase and a modified DNA template, in which deoxynucleotide Y makes a unique base pair with ribonucleotide X (in the form of XTP) to incorporate the modified nucleotide at a single, specific site on the RNA chain (29). d) The steps for incorporation of modified nucleotides by solid-phase oligonucleotide synthesis are shown (32–34). e) The most common protective groups for the phosphoramidite building blocks are highlighted.
Nucleosides in RNA, Modified

The chemical synthesis of RNA occurs on a solid support, such as a controlled-pore glass or polystyrene bead (Fig. 4d). The synthesis typically goes in the 3′ to 5′ direction, building from the 5′-hydroxyl group at each cycle. The 5′-OH is protected with a suitable organic moiety that can be removed quickly and efficiently. The 2′-hydroxyl is also protected until the end of the synthesis. The monomer unit that reacts with the 5′-OH is a nucleoside with a phosphoramidite moiety at the 3′ hydroxyl. The phosphoramidite method involves four major steps after deprotection of the 5′-hydroxyl group from the 3′ nucleoside attached to the solid support: 1) coupling at the 5′-OH with a suitably protected phosphoramidite, 2) capping the unreacted 5′-OH groups, 3) oxidation of the phosphite linkage to generate a phosphate, and 4) removal of the 5′-protecting group of the newly added nucleotide unit (Fig. 4d). The final steps involve removal of the oligonucleotide product from the solid support and deprotection of the 2′-OH groups and phosphate moieties. The final RNA product is then purified using standard methods, such as polyacrylamide gel electrophoresis or high-performance liquid chromatography.

The choice of 5′ and 2′ protecting groups is critical for the successful generation of an oligonucleotide product. The chosen phosphoramidites must have high coupling efficiencies and be stable during the entire oligonucleotide synthesis, yet easily removed at the end of the synthesis. The synthetic route to generate the phosphoramidites should also be efficient in order to reduce the cost of the oligonucleotide synthesis. A wide range of protecting groups has been introduced for RNA synthesis over the past three decades; however, only the most commonly used methods will be discussed. One of the first 2′-OH protecting groups to be employed was tert-butyldimethylsilyl (TBDMS or TBS) (32). More recently, [(triisopropylsilyl)oxy]methyl (TOM) (33) and benzhydryloxy-bis(trimethylsilyloxy)silyl (BzH) (34) have been developed for 2′-OH and 5′-OH protection (Fig. 4e).

Whereas each method has distinct advantages and disadvantages, they complement one another because unique chemical modifications on the nucleoside may be more compatible with one particular method. Modifications that are chemically incompatible with these protecting groups would require alternative methods. Several key factors in choosing the type of phosphoramidite to use include: 1) fast coupling times (<2 minutes), 2) high coupling efficiencies (>99.8%), 3) high stability in solution, and 4) the ability to generate long RNA fragments (>20 nucleotides in length). Many recent reviews on RNA synthesis protocols are available that provide more details regarding phosphoramidite preparation (6, 34).

The rationale for choosing chemical synthesis to generate modified RNAs is that a wide range of both natural and unnatural modifications can be incorporated at single or multiple points in the RNA sequence.
sites, and highly pure RNAs can be produced on a large scale (>mg). As mentioned earlier, convertible nucleotides can be generated in which a group of modified nucleosides that contain reactive functionalities are incorporated into RNA and allow for subsequent attachment of chemical probes, such as isotopic labels, tools for disulfide cross-linking, or fluorescent moieties (21). This postsynthetic modification allows for generation of site-specifically modified RNAs in fewer steps compared with complete enzymatic or chemical synthesis (Fig. 3).

Semisynthesis approaches
The enzymatic and chemical approaches can be easily combined to generate >100-nucleotide-length RNAs that contain either natural or unnatural modifications at specific sites. Either chemically or enzymatically produced fragments can be joined or ligated with T4 DNA ligase or T4 RNA ligase along with a DNA split, as mentioned above (27, 28). Together, these methods can be applied to generate large, full-length RNA molecules, such as 235 ribosomal RNA (approximately 3000 nucleotides in length), with site-specific modifications to study RNA function (35).

Applications of Modified RNAs
Numerous motivations can be cited for generating site-specifically modified RNAs. The first reason is to use modified RNAs for structure-function studies in either natural or model systems. For this purpose, many different biophysical approaches are employed, such as fluorescence spectroscopy, NMR spectroscopy, and X-ray crystallography. For mechanistic studies, site-specific changes to the oligonucleotide allow for single-atom or functional-group changes and subsequent biochemical studies to be conducted and compared with wild-type RNA. A second major reason for constructing modified RNAs is to use them as therapeutic agents with enhanced stability or improved biological activity. Antisense oligonucleotides were recognized long ago as having potential therapeutic properties for downregulating gene expression. More recently, RNA interference (RNAi) has been discovered as another tool for regulation of RNA activity. Other therapeutic agents include RNA aptamers and ribozymes. A topic of recent interest is the use of modified nucleosides to expand the genetic code, as well as to understand the basic principles of Watson-Crick base pairing and catalysis. A variety of nucleoside modifications can be used to monitor RNA folding, stability, and structure, as well as ligand interactions. Fluorescent analogs, such as 2-aminopurine (2-AP) (Table 1), have proven to be highly useful to monitor RNA conformational changes and dynamics in solution (40-42). The 2-AP modification causes very little perturbation to the RNA structure, and it has been used to monitor drug (e.g., aminoglycoside antibiotic) interactions with target RNA (53). In contrast, attachment of a bulky substituent, such as 8-bromoguanosine (8-BG) (Table 1), to the RNA base may limit rotation about the glycosidic bond, favoring one conformation (e.g., syn) over another (e.g., anti). Such an approach has been used to stabilize YNMG RNA hairpin motifs and enhance the RNA folding rate, because of preorganization of the denatured state of the RNA (43).

Biophysical probes
A variety of nucleoside modifications can be used to monitor RNA folding, stability, and structure, as well as ligand interactions. Fluorescent analogs, such as 2-aminopurine (2-AP) (Table 1), have proven to be highly useful to monitor RNA conformational changes and dynamics in solution (40-42). The 2-AP modification causes very little perturbation to the RNA structure, and it has been used to monitor drug (e.g., aminoglycoside antibiotic) interactions with target RNA (53). In contrast, attachment of a bulky substituent, such as 8-bromoguanosine (8-BG) (Table 1), to the RNA base may limit rotation about the glycosidic bond, favoring one conformation (e.g., syn) over another (e.g., anti). Such an approach has been used to stabilize YNMG RNA hairpin motifs and enhance the RNA folding rate, because of preorganization of the denatured state of the RNA (43).

Generation of site-specifically modified RNAs that can perform disulfide cross-linking allows for RNA folding pathways to be examined (21, 54). Photoactivated cross-linking is another strategy that uses naturally occurring or synthetic modifications, such as 4-thiouridine, 6-thioguanosine, 5-bromouridine, and 5-iodouridine (Table 1), to study RNA-RNA interactions in folded RNAs (19).
<table>
<thead>
<tr>
<th>Modification</th>
<th>Application</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-methylseleno ribonucleoside</td>
<td>phasing tool in X-ray crystallography</td>
<td>36–38</td>
</tr>
<tr>
<td>8-bromoguanosine</td>
<td>fluorescent probe of RNA conformational changes</td>
<td>40–42</td>
</tr>
<tr>
<td>2-amino-8-azapurine ribonucleoside</td>
<td>stabilize nucleoside conformations about the glycosidic bond</td>
<td>43</td>
</tr>
<tr>
<td>4-thiouridine, 5-thiouridine</td>
<td>photoactivated cross-linking probe of RNA folding to study RNA-RNA interactions</td>
<td>19</td>
</tr>
<tr>
<td>6-thio-2-aminopurine</td>
<td>natural modified nucleoside</td>
<td>44–47</td>
</tr>
<tr>
<td>pseudouridine</td>
<td>mechanistic probe of RNA structure and function</td>
<td>48</td>
</tr>
</tbody>
</table>
Modified nucleosides incorporated into small RNA model systems can also be used to investigate the global versus individual effects of modified nucleotides on natural RNAs, such as rRNA or tRNA. For example, in some early studies, Yarian et al. (44) demonstrated that pseudouridine (Table 1) leads to increased thermal stability of the tRNA anticodon stem-loop region. Later, Meroueh et al. (45) demonstrated that pseudouridines have opposing effects on rRNA helix 69 stability, which depends on their specific locations and sequence contexts. These effects on stability may be important for conformational switching mechanisms in functional RNAs (46, 47).

Mechanistic probes

Site-specifically modified RNAs have been used in many applications to examine RNA structure-function relationships, RNA–protein interactions, RNA–ligand interactions, and RNA-catalysis mechanisms. Some earlier studies demonstrated the use of synthetic oligonucleotides to probe the roles of specific functional groups and detailed mechanisms in ribozyme catalysis (55). The synthesis of nucleoside analogs allows for a full range of chemical diversity (e.g., inductive effects, space-filling capacity, etc.) to be explored, such that quantitative structure activity relationships can be determined for RNA enzymes and other biologically important RNAs (56).

Nucleotide analog interference mapping (NAIM) is an alternative method to conventional single-atom substitution experiments, which provides biochemical information related to the RNA structure and function (48). NAIM has the ability to screen rapidly the effects of chemical group substitution rapidly on RNA function. This method uses an array of modified 5′-O-(1-thio)-nucleoside triphosphates (Table 1) to assess concurrently the contributions of functional groups at every nucleotide position along the RNA backbone. The applications of this method range from mapping of RNA tertiary contacts and substrate contacts in ribozymes to quantification of the energetic contributions of RNA–ligand interactions.

Functional RNAs

RNAs have an enormous range of functions. Those functions have been the subject of numerous publications over the past several decades. One example is RNA aptamers, which are oligonucleotides that have been selected for and tailored to bind with high affinity and selectivity to a variety of ligands, which include nucleotides, peptides, proteins, vitamins, and drugs (57). Aptamers are engineered from random nucleic acid libraries through repeated rounds of in vitro selection. Modified nucleotides play important roles in aptamer function, because they provide enhanced stability to the RNA-based aptamers and can also endow them with desirable chemical reactivities or control of their folded structures, which would be beneficial for the development of novel targeting agents. Several modified aptamers have already been developed as potential therapeutics (57).

Nucleic acid selection methods have also been exploited for the development of novel RNA enzymes or ribozymes (58). An in-vitro-selected RNA that contains the modified nucleotide 5′-(4-pyridylmethyl)-uridine (Table 1) can catalyze carbon–carbon bond formation in a Diels-Alder cycloaddition, with an 800-fold rate acceleration compared with a random RNA (49). Modified RNAs that contain the same uridine modification have also been selected to mediate metal–metal bond formation in the synthesis of palladium nanoparticles (59). Modified RNAs are likely to have many other applications as novel ribozymes that catalyze important biological reactions or can be used to create novel materials.

RNAi/siRNA

Short interfering RNA (siRNA) is a class of molecules composed of double-stranded RNA molecules (20–25 nucleotides) that are involved with the interference of specific genes (i.e., RNA interference or RNAi). During the RNAi process, the siRNA is incorporated into a protein complex, which unwinds the siRNA by an ATP-dependent mechanism to generate an
active RNA-induced silencing complex (RISC) (60). The unwound antisense strand of RNA will then guide the RISC to the target mRNA and bind according to standard Watson-Crick base-pairing schemes. Strand scission of the target mRNA is induced by RISC, which leads to a decreased level of protein expression by the corresponding mRNA. The overall success of gene downregulation depends on the stability and structure of the duplex formed between the siRNA and target mRNA. This idea has led to the design and synthesis of numerous modified siRNAs with the anticipation that appropriate changes in the nucleoside composition will lead to enhanced stability, both in terms of chemical stability (i.e., to avoid nuclease degradation) and thermal stability (i.e., to generate a duplex with more favorable free energy, enthalpy, and/or entropy contributions). The modified oligonucleotides are also designed to have increased cellular uptake and properties that will enhance their biodistribution or entry into specific cell types (61, 62). Another type of nucleotide modification involves the addition of a photocurable moiety (Table 2), such that temporal and spatial control of siRNA targeting is possible (50).

Other applications

RNAs that employ the four standard bases have more limited opportunities for hydrogen-bonding interactions and functionality, even though the range of tertiary interactions is broad (63). The use of additional base pairs would not only increase the structural complexity of RNA but would also allow for enzymatic replication. This concept has been exploited by several research groups to expand the genetic code, as well as to generate novel, site-specifically modified RNAs (e.g., biotinylated, fluorescent, cross-linked RNAs) (6, 29, 64).

Summary and Outlook

This article highlights methods for synthesis of modified nucleosides and their site-specific incorporation into RNA. RNAs of practically any size can now be engineered with a broad range of natural or unnatural modifications to the base, sugar, or phosphate portions of the nucleotides. Using chemical, biochemical, or combined approaches, a change of a single atom at a desired location can be accomplished, as well as global modification. Because of the diversity of modifications accessible through chemical or enzymatic synthesis, the modified RNAs also have a tremendous range of applications, some of which have been summarized in the last portion of this review. Future efforts will continue to give researchers a better understanding of the importance of natural modified RNAs, provide new tools for biophysical and structure analysis for structure-function studies of RNA, and lead to novel RNA therapeutics or catalysts. On a more practical note, novel methodologies for cost-effective, large-scale preparation of highly pure modified RNAs will also be of importance (65), and improved catalysts for RNA ligation (66) will allow for larger, biologically relevant RNAs to be generated with greater ease.

Acknowledgments

We are grateful for support from the National Institutes of Health (Grants GM 054632 and AI061192).

References

13. Holmström C, Knuts C, Flecker A, Ottensmayer E, Pils W, Grumbach K, Mocura K. The synthesis of 2-0-[tris(4-methylphenyl)oxy] methyl (trm) phosphorodiamidates of methylated dioxolanes (mg2, mg3, mg2f, mg3f, mg2m, mg3m, mg2a, mg3a) for use in automated RNA solid-phase synthesis. Monatshefte fur Chemie 2003;134:853-871.


Nucleosides in RNA, Modified


See Also

Fluorescence to Study Nucleic Acids
NMR Tools for Nucleic Acids
RNA Interference, Mechanisms and Proteins Involved in
Peptide Nucleic Acids
Catalina Achim, Bruce A. Armitage, Danith H. Ly and James W. Schneider, Carnegie Mellon University, Pittsburgh, Pennsylvania
doi: 10.1002/9780470048672.wecb435

Peptide nucleic acid (PNA) is a chimeric molecule that consists of hydrogen-bonding purine and pyrimidine heterocycles attached to a pseudopeptide backbone. The bases allow the recognition of specific DNA or RNA sequences, which results in hybrid multihelical structures. The lack of a negative charge on the PNA backbone eliminates Coulombic repulsion from target DNA and RNA strands, which results in high affinity hybridization. These properties have led to a diverse set of applications for PNA, which includes antisense/antigene inhibition of gene expression, various DNA/RNA detection assays, nucleic acid labeling and purification technologies, and programmed assembly of nanoscale materials. The modular design of PNA has led to the synthesis of several “next-generation” analogs that promise to improve PNA’s performance in many existing applications as well as to open doors to new applications. Although most prior research on PNA has focused on its nucleic acid-like character, recent developments of the peptide-like aspects of PNA promise to stimulate additionally the work on this fascinating DNA mimic.

The DNA double helix is the first molecular icon, a structure that is readily associated with its function, even by people with little or no science education. Over the past half century, the structure has inspired chemists along three lines that are familiar to those who have worked at the interface between chemistry and biology. The first group of chemists synthesized in the laboratory what nature produces enzymatically, and they were motivated by both fundamental and practical needs. The second group of chemists understood the relationship between structure and function, where synthetic chemistry allows atomic-scale changes in the DNA structure. The third group of chemists identified DNA as a potential target for molecular recognition, with the ability to regulate gene expression using synthetic compounds as the ultimate goal. These lines of inquiry are not necessarily parallel, but rather they become intertwined at various points, most notably in the area of synthetic oligonucleotides. A variety of DNA analogs have been synthesized and characterized, which feature modifications to the deoxyribose ring, the phosphodiester linkage, and the heterocyclic nucleobases. Many of these modifications have been combined within the same structure to optimize the oligonucleotide for a specific application.

Most synthetic DNA analogs represent logical departures from the natural structure. However, one of the most radical structural modifications of DNA to appear in the literature is peptide nucleic acid (PNA, Fig. 1), in which the sugar-phosphate backbone is abolished entirely in favor of a pseudopeptide. PNA was first reported by Nielsen et al. in 1991 from the University of Copenhagen (1). The structure is reminiscent of both proteins and nucleic acids, hence the name, and has led to considerable interest in its possible role in prebiotic chemistry.

An intriguing molecular structure and origins-of-life scenarios hardly explain the intense research effort dedicated to PNA, whereas the original 1991 Science article has been cited over 1100 times as of September 2006. Rather, it is what PNA does in the presence of DNA and of RNA that has captured the imagination of chemists and of biologists. Although PNA was conceived originally as a triplex-forming oligomer that would bind in the major groove of double-stranded (ds) DNA, the earliest reports demonstrated that homopyrimidine PNA binds to ds DNA by a strand invasion mechanism. First, one PNA strand displaces the homologous homopyrimidine DNA strand locally and forms Watson-Crick base pairs with the complementary homopurine DNA target (Fig. 2) (1, 2). A second PNA strand then binds to the major groove of the hybrid to form a very stable PNA-DNA triplex. This discovery was very exciting because...
Peptide Nucleic Acids

Figure 2

Most common DNA/RNA binding modes. Left: Strand invasion of homopyrimidine PNA into duplex DNA yields a PNA2–DNA triplex. Right: Hybridization of mixed sequence PNA with complementary DNA or RNA produces Watson-Crick base-paired duplex structures.

targeting ds DNA opens the door to the regulation of gene expression at the transcriptional level. Several variations on this “strand invasion” recognition mechanism that ease the sequence requirements have been reported in the ensuing 15 years, and research is ongoing (see Further Reading).

PNA is also capable of binding to complementary single-stranded (ss) DNA and to RNA mixed-sequence (i.e., purine-containing and pyrimidine-containing) targets by Watson-Crick base pairing to form hybrid duplex structures (Fig. 2) (3). As with the triplex structures mentioned above, these hybrids exhibit high thermodynamic stability partly because of the lack of electrostatic repulsion between the PNA and DNA/RNA. High affinity and excellent mismatch discrimination led to many PNA applications that range from the inhibition of mRNA processing and translation to sensing, diagnostics, and imaging. The original Nature article (3) that describes Watson-Crick PNA–DNA and PNA–RNA duplex formation has been cited over 700 times to date.

PNA recognition of DNA and RNA has been extended recently to a new binding mode, named “homologous hybridization” (4, 5). In this format, guanine-rich PNA probes recognize homologous DNA or RNA targets by forming hybrid guanine quadruplexes. Great interest exists in G-quadruplexes for their suspected roles in the regulation of gene expression at the transcriptional and translational levels, so this added binding mode could expand the range of biologic targets for PNA.

This article focuses on the interplay between chemical properties and biologic applications of PNA. Besides strategies to regulate gene expression, PNA-based biotechnologies will be presented. Although the vast majority of these studies have used the original PNA structure, the next-generation PNAs that feature backbone and nucleobase modifications are likely to be used in future applications. Finally, to illustrate the versatility of this fascinating molecule, the use of PNA in materials science will be summarized. The growing footprint of nanotechnology in biologic research warrants careful consideration of the opportunities presented by PNA’s unique properties. As space limitations prevent an exhaustive description of PNA research, readers are referred to several recent reviews for additional information (see Further Reading).

PNA as Antisense and Antigene Agents

Compounds that bind sequence selectively to single-stranded RNA or double-stranded DNA have the potential to regulate gene expression by antisense or antigene mechanisms, respectively. The high affinity and sequence selectivity exhibited by PNA sparked much interest in the possible use of DNA in these applications. Furthermore, the biochemical stability of PNA, as it is not susceptible to cleavage by nuclease nor by protease enzymes, raised hopes that PNA's could be effective in vivo as well as in vitro. Experiments in which the PNAs were targeted either to RNA or to DNA, are described in this section.

Antisense applications

In the classic antisense approach, the binding of an oligonucleotide to a complementary region of a mRNA prevents translation via three possible mechanisms: 1) binding of the antisense agent blocks ribosome assembly on the mRNA, 2) an actively translating ribosome cannot proceed past the steric block presented by an antisense agent bound within the coding region of the mRNA, or 3) the hybrid duplex formed by the antisense agent and the mRNA recruits RNase H, which degrades the mRNA strand. PNA–RNA hybrids are not substrates for RNase H, which leaves the first two mechanisms available for PNA antisense agents. Scattered reports indicate that PNAs targeted to the coding region of an mRNA can inhibit translation, presumably by mechanism 2). However, two separate groups have reported that PNAs targeted to either the translation start codon or the 5′-untranslated region exhibited antisense effects, whereas PNAs targeted to the coding region did not (6, 7). The failure of the latter PNAs to block translation could be caused by the weak binding of the PNA to its target sequence, which could occur if the target sequence is part of a stable, folded region of the mRNA. Alternatively, the PNA could bind but be displaced by the ribosome as it translates through the PNA binding site. Regardless of the explanation, these studies indicate that PNAs that interfere with ribosome assembly are most likely to be effective antisense agents.
Three other examples of RNA-targeted PNAs that block gene expression should be mentioned. In the first case, the PNA is not targeted to the mRNA but to the unspliced pre-mRNA present in the nucleus immediately after transcription. Splicing eliminates introns from the pre-mRNA and results in the ligation of exons to produce the mature mRNA. Alternative splicing, in which the same pre-mRNA produces different RNAs based on the use of different splice sites, increases the complexity of the genome. This process has been implicated in certain genetic diseases, whereas the use of a splice site leads to a normal protein, whereas the use of a different splice site can lead to a defective protein (and the absence of the normal protein). Cell culture experiments with complementary PNAs targeted to a splice site used in place of adenine and thymine (12). (See the “Nucleobase Modifications” section.) A significant challenge to using homopyrimidine PNAs in vivo is the basic process that underlies several biotechnologic techniques that constitute a multibillion-dollar-per-year industry. DNA/RNA microarrays, polymerase chain reactions (PCRs), single nucleotide polymorphism (SNP) assays, and fluorescence in situ hybridization (FISH) are only some techniques that rely on Watson-Crick base pairing. Although each of these methods works when using a synthetic DNA oligonucleotide as a capture or detection probe, the higher affinity and nuclease stability of PNA offers obvious advantages (except in PCR, where a pure PNA oligomer cannot serve as a primer). Besides allowing lower concentrations of probe to be used, which helps to minimize hybridization to unintended targets that have a similar sequence, PNA can invade stable secondary or tertiary structural elements to gain access to its target. These folded structures impose thermodynamic and kinetic penalties to any hybridization probe; the higher affinity of PNA combined with its lack of electrostatic repulsion from DNA and RNA leads to significant advantages of PNA over most oligonucleotides based on polyanionic backbone chemistries. This section illustrates a few applications of PNA in biotechnology.

Fluorescent PNA probes

PNA probes have been used for FISH experiments, and commercial kits are available for microbial targets (17). Typically, the probes present in these kits hybridize to the abundant ribosomal RNA from bacteria. A recent paper illustrated the value

Probes, Labels, and Sensors

Hybridization between complementary oligonucleotide strands is the basic process that underlies several biotechnologic techniques that constitute a multibillion-dollar-per-year industry. DNA/RNA microarrays, polymerase chain reactions (PCRs), single nucleotide polymorphism (SNP) assays, and fluorescence in situ hybridization (FISH) are only some techniques that rely on Watson-Crick base pairing. Although each of these methods works when using a synthetic DNA oligonucleotide as a capture or detection probe, the higher affinity and nuclease stability of PNA offers obvious advantages (except in PCR, where a pure PNA oligomer cannot serve as a primer). Besides allowing lower concentrations of probe to be used, which helps to minimize hybridization to unintended targets that have a similar sequence, PNA can invade stable secondary or tertiary structural elements to gain access to its target. These folded structures impose thermodynamic and kinetic penalties to any hybridization probe; the higher affinity of PNA combined with its lack of electrostatic repulsion from DNA and RNA leads to significant advantages of PNA over most oligonucleotides based on polyanionic backbone chemistries. This section illustrates a few applications of PNA in biotechnology.
of PNA as a hybridization probe to detect Legionella pneumophila (18). Discrimination between this and other species of Legionella bacteria required a probe that could bind to sites that typically are classified as “low affinity” because of the poor hybridization of DNA probes. PNA probes showed absolute specificity, which identified correctly 47 different strains of bacteria and distinguished between pneumophila and other Legionella species.

PNA FISH probes also have been used for telomere analysis (19). Telomeres consist of hundreds of repeats of a short sequence (e.g., 5′-TTAGGG-3′ in humans). Therefore, many copies of a complementary nucleic acid probe can hybridize to each telomere in a cell, which leads to a bright fluorescence signal. The high affinity of PNA allows the use of probes shorter than DNA probes, which means that more copies of the probe can hybridize to the telomeres.

A third application for fluorescent PNA probes is to introduce fluorescent labels into the mRNA site specifically (20). For example, PNA probes have been hybridized to exonic sites that flank consecutive splice sites in a pre-mRNA from yeast. Förster resonance energy transfer (FRET) donor and acceptor dyes were attached covalently to the PNA, and low FRET efficiencies were observed when the PNAs were hybridized to the pre-mRNA. However, when hybridized to the mRNA produced by splicing, large increases in FRET have been observed in both in bulk solution and on a glass slide where single-molecule measurements could be made. As in the FISH applications, the ability to use short PNA probes to deliver the fluorescent dye to a desired location decreased the likelihood that the PNA will disrupt the structure and the function of the RNA that is under investigation.

PNA can also be functionalized with “fluorogenic” dyes, that is, dyes that exhibit enhanced fluorescence in response to a change in the environment. Unsymmetrical cyanine dyes developed originally as DNA stains can be attached covalently to PNAs to create fluorogenic probes, and chemistries have been developed that allow attachment of the dye at internal (21) as well as at terminal positions (22).

Finally, the intrinsic properties of PNA led to an interesting variation on the molecular beacon concept. As devised originally, a DNA probe that bears a 5′-terminal and 3′-terminal fluorophore (F) and quencher (Q) groups, respectively, can be designed to fold into a hairpin in which the fluorescence is quenched because of the close proximity of F and Q (23). Subsequent hybridization with a complementary DNA opens the hairpin and the increase in distance between F and Q restores the fluorescence. This design necessitates that the terminal DNA sequences are complementary to stabilize the hairpin conformation by Watson-Crick base pairing. If F and Q groups are attached to a PNA probe, the terminal sequences are not required (24), which is most likely caused by the tendency of PNA to adopt a compact rather than an extended structure in the absence of a complementary strand. These “stemless” PNA beacons ease the sequence constraints that complicate the DNA molecular beacon design.

### Probes for nonfluorescent DNA/RNA detection

PNA has been tested in a wide variety of DNA detection modes where unmodified DNA can also be used. However, besides the standard advantages noted above for PNA (high affinity, good mismatch discrimination), many of these assays exploit the ability of PNA to hybridize at low ionic strength, conditions that can denature competing secondary/tertiary structure in the target nucleic acid. Moreover, use of PNA with its neutral backbone as a probe for hybridization at surfaces eliminates the substantial charge repulsion between the DNA/RNA target and immobilized DNA probes.

An interesting new approach to DNA detection involves probes attached to polymer microspheres loaded with a scintillator (25). The hybridization of a complementary, radiolabeled target DNA leads to strong signal amplification because of the close proximity of the scintillant and the radiolabel. Another appealing aspect of this method is that the PNA probe can be synthesized directly on the scintillant-loaded microspheres.

PNA probes have also been used for in vivo imaging of mRNAs (26). In this application, the PNA is functionalized with a cell-penetrating peptide and a magnetic resonance imaging (MRI) contrast agent. Cell uptake and hybridization to the mRNA target allowed MRI imaging of gene expression both in cell culture and in live rats.

Another method for nonfluorescent detection of DNA involves the noncovalent binding of an inexpensive cyanine dye to PNA–DNA duplexes (27). The dye assembles into a helical aggregate in the presence of PNA–DNA, which results in a vivid blue-purple color change. This phenomenon has been used in SNP detection assays.

### PCR clamping

An early application for PNA is in a method known as “PCR clamping,” which improves the detection of single nucleotide polymorphisms (SNPs) in samples where mutant DNA of interest is present in a very low amount relative to wild type (28). This situation can occur, for example, when only a few mutant cancer cells are present along with a large excess of healthy cells in a tissue biopsy. The challenge is to detect the low abundance of mutant sequence against the high background of wild-type sequence. One approach to improve the signal-to-noise ratio in such situations is to amplify selectively the mutant DNA by PCR. This is where PNA is particularly helpful. PCR primers are designed to overlap the SNP site present in the cancer cells and then a PNA oligomer that is fully complementary to the wild-type DNA (and therefore has a single mismatch to the mutant DNA) is synthesized. The PNA preferentially hybridizes to the wild-type DNA in competition with the PCR primer that is complementary to the same site, which reduces the amplification of the wild type relative to the mutant DNA. This method allows SNP detection when the wild-type DNA is present in up to 20,000-fold excess relative to the mutant.
Bioseparations

A growing need exists to develop sequence-specific means to purify nucleic acids, both for genomic analysis and for the larger scale purification of plasmid DNA for gene therapy. The sequence-specific binding properties of PNA can be used in affinity separations of DNA as an alternative to standard DNA purification methods such as anion-exchange chromatography and density-gradient techniques. This section describes several such applications of PNA.

Analytical applications

The hybridization of PNA to a complementary DNA or RNA leads to changes in the properties of the nucleic acid. One of these changes occurs in the electrophoretic mobility, where the perturbation in the charge: size ratio because of the PNA hybridization causes significant shifts in the rate at which a given DNA or RNA migrates in an electric field. Adjusting conditions such as temperature allows preferential hybridization of the PNA to fully complementary targets versus those bearing even single mismatches, and it results in large separation during electrophoresis.

Purification applications

In addition to the analytical-scale applications described above, large-scale purification of nucleic acids using PNA probes has also been explored. Applications range from the purification of plasmid DNA to be used in gene therapy to the isolation of RNA and RNA–protein complexes from complex biologic samples. One approach to using PNA in this way is to link it to a solid support such as a polymer bead and then use the PNA-functionalized support to capture complementary DNA from solution. However, both biosensor and surface plasmon resonance data point to a reduction in the specificity and the binding kinetics to PNA probes that are attached directly to solid surfaces. The most effective “purification by hybridization” strategies circumvent this by two-step methods, in which PNA probes with some attachment functionality (e.g., hexahistidine peptide or biotin) bind targets in solution and the resulting PNA–DNA (RNA) hybrids are captured on beads or surfaces. The most effective “purification by hybridization” strategies circumvent this by two-step methods, in which PNA probes with some attachment functionality (e.g., hexahistidine peptide or biotin) bind targets in solution and the resulting PNA–DNA (RNA) hybrids are captured on beads or surfaces. The target DNA thereby is modified with a nonpolar tag that may adsorb selectively to nonpolar media, including alkyl-modified Sepharose and surfactant micelles (Fig. 3). In general, the covalent attachment of n-alkanes and other lipophiliic materials impacts neither the duplex stability nor the sequence selectivity of PNA–DNA duplexes. The attachment of a 12-carbon alkane to PNA provides adequate resolution to separate 60-mer DNA targets from both noncomplementary oligomers and calf thymus DNA in hydrophobic interaction chromatography (HIC), an aqueous-based method. The resolution is improved greatly when 18-carbon alkanes are attached to PNA probes, even when the PNA is targeted to internal sequences of the target. The strand-invasion ability of PNAs can also be leveraged to purify dsDNA oligomers in HIC. Far greater resolution can be achieved when targeting dsDNA because DNA in duplex form interacts very weakly with the HIC media unless the alkylated PNA is attached.

Finally, using PNA as an affinity capture reagent recently was extended to probing RNA–protein complexes (RNP)s in cells (33). In this application, the PNA is functionalized with a peptide that allows uptake into cells and is complementary to an RNA component of an RNP. The PNA also bears two affinity tags, the first of which is a benzophenone-modified phenylalanine residue that can photodissociate the PNA to a protein present in the RNP. The second tag is a biotin group, which allows the purification of the cross-linked PNA–protein. Subsequent analysis by mass spectrometry identifies both the protein and its cross-linking site. As is the case for PNA used to deliver a fluorophore to a specific site in an RNA, this method requires that the PNA not disrupt the structure being probed.

PNA-encoded libraries

Another application of PNA that has emerged recently is a tag to encode and to screen combinatorial libraries of small molecules (34). For example, fluorescent PNA tags have been attached to individual members of a library of potential protease inhibitors. After incubating the library with a target protein, bound versus free compounds were separated by size-exclusion chromatography and the PNAs obtained from the bound fraction were hybridized to a DNA microarray. Any DNA to which fluorescent PNA was bound allowed decoding of first the PNA and then the compound from the library tagged by the respective PNA. The authors suggested alternative screening, profiling, and assay formats for this promising technology.

Next-Generation PNAs

PNA and its structural analogs are synthesized conveniently by standard solid-phase methods (35). The backbone chemistry of PNA allows the attachment of a diverse array of fluorescent dyes, metal-binding ligands, cell-penetrating peptides, and many other moieties that increase the functional capacity of the PNA. In addition to PNAs that simply are modified by attaching functional groups to the N-terminus or C-terminus, large numbers of PNA analogs that feature integral modifications, i.e., alterations in the backbone or nucleobase components, have been reported (36). Most of these analogs were made in attempts to better understand the structure-function relationship of PNA so that the hybridization, solubility, and/or cellular uptake properties of PNA could be improved. In this section, we introduce several second-generation PNAs, classified as either having backbone or nucleobase modifications, and discuss their potential impact on the applications described above.

Backbone modifications

Most backbone-modified PNA analogs fall into one of four classes (Fig. 4). The first class was made by inserting a methylene group into the PNA backbone or the linker that connects the nucleobase to the backbone (Fig. 4, I). These modifications were made to assess the effect of chain length, in the backbone...
Peptide Nucleic Acids

Figure 3 PNA amphiphiles (representative structure shown) can form mixed vesicles with surfactants such as SDS. Hybridization to complementary DNA or RNA allows separation from complex mixtures using open channel capillary electrophoresis.

Figure 4 Examples of backbone-modified PNAs.

as well as in the linker, on the hybridization properties of PNA. Both structural changes induced significant destabilization of the PNA–DNA and PNA–RNA hybrid duplex. A single backbone modification lowered the $T_m$ of a PNA–DNA duplex by 8–20°C and a PNA–RNA by 6–15°C, depending on where the modification was made. These results suggest that the original PNA design, N-(2-aminoethyl) glycine backbone and carboxymethyl linker, which has the same number of atoms in the backbone and in the linker as the DNA, is optimal for hybridization to DNA and RNA.

The second group of backbone modifications was made by introducing methylene bridges that connect the various functional groups in the backbone and in the linker (Fig. 4 II). These modifications were made in attempts to introduce structural preorganization into an otherwise randomly folded PNA backbone structure (37) and led to improvements in the binding affinity (38) and RNA versus DNA selectivity (39). Although this result supports the hypothesis that preorganization of the strands reduces the entropy penalty for duplex formation, additional structural analyses will be required to determine whether these backbone modifications do preorganize PNA into helical structures and whether these helical structures bear close resemblance to that of the targeted molecules, DNA and RNA.

The third group of backbone modifications was made by installing amino acid side chains with R or S configuration at the α-position of the N-(2-aminoethyl) glycine unit (Fig. 4 III) (40). Generally, these chiral PNAs were found to form slightly less stable PNA–DNA complexes than the original unsubstituted PNA for backbones that contain amino acid side chains with bulky, apolar groups. PNAs that contain negatively charged amino acid side chains, such as aspartic or glutamic acid, were shown to induce even greater destabilization, presumably because of the electrostatic and/or steric repulsion. On the other hand, incorporation of positively charged amino acid side chains, such as lysine or arginine, led to the stabilization of the PNA–DNA duplex, with R having a larger effect than S.
amino acids. Furthermore, the arginine-modified PNAs, also referred to as "GPNAs" because of the presence of guanidinium groups on the PNA backbone, exhibited substantially improved cell uptake relative to unmodified PNAs (41). This is likely because of the similarity between GPNAs and various natural and synthetic compounds that feature multiple guanidinium groups and readily enter mammalian cells.

The fourth and perhaps most intriguing class of backbone modifications feature amino acid side chains installed at the γ-position of the N-(2-aminoethyl) glycine unit (Fig. 4 IV). Unlike the unmodified PNA or other PNA derivatives that have been developed so far, generally that do not fold into well-defined conformations, γ-modified chiral PNAs assume helical conformation (42). These chiral PNAs preorganize into either a right-handed or left-handed helix, which depends on the backbone configuration of the amino acids from which the PNA analogs were derived. Those PNAs that are derived from naturally occurring (and less costly) L-amino acids preorganize into a right-handed helix and bind to DNA and RNA with high affinity and sequence selectivity, whereas those derived from unnatural D-amino acids preorganize into a left-handed helix and do not bind effectively to either DNA or RNA.

**Nucleobase modifications**

Several modified nucleobases have been incorporated into PNA—many of which have been covered in a recent review (36). In general, any modified nucleobases can be incorporated into PNA as long as they can withstand the conditions used commonly in Boc or Fmoc solid-phase peptide synthesis. One of the first modified nucleobases to be incorporated into PNA was pseudo-isocytosine (J, Fig. 5). J is a synthetic analog of N-3 protonated cytosine and forms Hoogsteen base-pairs with guanine in triplexes at neutral pH. This enhances the stability of PNA2–DNA triplex invasion complexes significantly under physiologic conditions, although the kinetics of strand invasion at physiologic ionic strength is relatively slow.

Another important nucleobase to be incorporated into PNA was 2,6-diaminopurine (D, Fig. 5). D is a synthetic analog of adenine (A) that contains an extra exocyclic amine. When base-paired with the complementary thymine nucleobase, D forms three H-bonds as opposed to two H-bonds in base pairs of T with A. The extra H-bond provides additional stability to the hybrid duplex, with a gain in $T_{m}$ of 3–5°C/bp for PNA–DNA. In addition to providing duplex stability, D has been exploited, in combination with another modified nucleobase, 6-thiouracil (sU), in the recognition of mixed-sequence, double helical DNA in a binding mode called “double-duplex invasion” (12). In this binding mode, both strands of the ds DNA are targeted simultaneously by two PNA strands. Replacement of the naturally occurring T and A nucleobases with D and sU, which form a less stable base pair than T-A and T-U, prevents the two complementary PNA strands from hybridizing to one another to form a PNA–PNA complex. This strategy has been used successfully in the sequence-specific recognition and cleavage of mixed-sequence, ds DNA (43).

More recently, another modified nucleobase termed “guanidino G-clamp” (X, Fig. 5) has been incorporated into PNA (44). Guanidino G-clamp is an analog of cytosine, but unlike cytosine, which can form only three H-bonds with guanine, it can form five H-bonds—three through Watson-Crick and two through Hoogsteen base pairing. Besides the extra H-bonds, the guanidino G-clamp stack effectively with the adjacent nucleobases because it contains a tricyclic phenoxazine ring, and thus, it provides additional stability to the bound complex. No binding studies have been performed on guanidino G-clamp, but the introduction of a similar nucleobase analog that contains an amino instead of a guanidino group enhanced the stability of PNA–DNA and PNA–RNA duplexes by 23°C and 18°C per modification, respectively (45). Guanidino G-clamp is expected to confer even greater stabilization because of its ability to form five H-bonds as opposed to four in the case of amino G-clamp.

**Nanotechnology and Materials Science**

The information storage ability conferred to PNA by the nucleobases together with the chemical and biologic robustness of this synthetic nucleic acid have made PNA appealing particularly for use in the organization of chemical objects ranging
in size from molecular-scale to nanoscale. The outcome of the interaction between PNA and these chemical entities are materials whose chemical composition and three-dimensional structure are encoded in the nucleobases and/or in the secondary structure of the PNA. The information stored in the PNA is “expressed” into the new structures by relatively weak chemical and physical forces, such as hydrogen or coordination bonding and hydrophobic interactions. The synthesis and the study of the properties of these pre-programmed materials provides information relevant for the understanding of fundamental biologic processes, such as molecular recognition and self-assembly, and of the rules based on which artificial enzymes and molecular electronic devices can be built. The PNA-based materials have the potential of being useful in biology, diagnostics, and therapeutics.

Toward the molecular end of the scale range, PNA has been shown to act as scaffold for transition metal ions situated at the core of PNA duplexes (46) (Fig. 6). The metal ion incorporation was realized by the chemical substitution of nucleobases with ligands. This process is site specific because the ligands have higher affinity for metal ions than the nucleobases. The existence of the metal ions within the PNA duplex opens the possibility of directional electron transfer mediated by the metal ions, in a manner similar to that in which electron transfer metallo-proteins work in Nature. Also, some of the metal ions enhance electrochemical, optical, or magnetic resonance detection of the PNA strands or duplexes. At the nanoscale end, the predictable molecular recognition process based on hybridization of ssPNA attached covalently to carbon nanotubes (47) or shell cross-linked nanoparticles (48) has been used to achieve controlled mechanical manipulation of nanosize objects and to organize them in large-scale architectures or on surfaces. The strong interaction between PNA and DNA has also been exploited in the synthesis of DNA-based nanostructures. For example, bisPNAs have been used as tools for rational design of nanosize, DNA-based locked pseudorotaxanes or on surfaces. The strong interaction between PNA and these chemical entities are materials or on surfaces. The strong interaction between PNA and DNA has also been exploited in the synthesis of DNA-based nanoscale structures. For example, bisPNAs have been used as tools for rational design of nanosize, DNA-based locked pseudorotaxanes and catenanes. The strand invading ability of homopyrimidine bisPNAs allows them to be used as sequence-specific “openers” of DNA duplexes (49). The resulting single-stranded region of the DNA can in turn bind opposite ends of another single-stranded DNA that is circularized by DNA ligase to produce a catenane-like structure (50).

Biotinylated PNA oligomers have also found use as cross-linking agents to connect avidin with DNA three-way junction nanostructures, which yield hydrogels that grow and shrink in a temperature-dependent manner (51). Avidin-linked enzymes can be integrated into the hydrogel structure, where they can catalyze reactions of small substrates that diffuse into the pores of the gel (52). Cooling below room temperature led to precipitation of the gel and allowed the separation of the product and recovery of the enzyme.

**Conclusion**

The early reports of PNA’s abilities to bind to single-stranded and double-stranded nucleic acids with high affinity and selectivity generated much excitement that led to both in-depth characterization of the binding process as well as an impressive array of biologic, biotechnologic, and nanotechnologic applications. The pace of PNA research as measured by number of publications has grown steadily for the past 15 years and shows no sign of abating. Additional advances are expected to come as next-generation PNAs are developed, particularly those that feature backbone modifications and take advantage of the peptide-like character of PNA. Recent successes in improving antigenic/antisense activity and cell uptake bode particularly well for future use of this most versatile of DNA mimics in chemical biology.

**References**


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.

Further Reading

See Also
Nucleic Acids, Chemistry of Nucleic Acids, Design and Engineering of Nucleic Acid Recognition by Peptides and Drugs Synthetic Chemistry: Building Molecules to Modulate Biological Systems Synthetic Nucleic Acids to Study Biologic Function
Artificial Metalloproteins, Design and Engineering of
Yi Lu, Dewain K. Garner and Jun-long Zhang, University of Illinois Urbana-Champaign, Illinois
doi: 10.1002/9780470048672.wecb397

Engineering of artificial metalloproteins is an expanding field with potential impacts in many areas from fundamental understanding of protein structure and function to industrial production of specialty chemicals such as chiral drug intermediates. Incorporation of unnatural amino acids and non-native metal cofactors into proteins is an emerging field in the area of protein design, as it offers the tantalizing prospect of introducing new functionality and provides exquisite probes for and fine-tuning of native protein properties. Although it is a relatively young field, engineering of non-native metalloproteins boasts a myriad of techniques for design, construction, and/or expression of the desired artificial protein. Additionally, many groundbreaking studies exist in this field that have enhanced basic scientific understanding of protein function or generated promising artificial enzymes. Discussion of the techniques prominent for incorporation of unnatural amino acids and non-native cofactors is followed by some of the most interesting applications of these techniques reported to date.

Approaches to Unnatural Amino Acid Incorporation into Proteins

Many methods have been developed to incorporate unnatural amino acids into metalloproteins, and these methods can be classified as either biological or synthetic in nature. Biological approaches rely on cellular machinery for placement of the unnatural amino acid or cofactor with its unnatural analog, while altering a specific structural or electronic element, may be accomplished without gross perturbation of the protein site it replaces. As the title suggests, metal-containing proteins are the focus of this article, and although many wonderful examples of protein incorporation of unnatural amino acids and cofactors exist, only those for metalloproteins will be considered in this review. Because the elementary nature and function of amino acids and cofactors differ, it is sensible to discuss them separately. Furthermore, discussion of the techniques prominent for incorporation of unnatural amino acids and non-native cofactors will be followed by some of the most interesting applications of these techniques reported to date.

Approaches to Unnatural Amino Acid Incorporation into Proteins

Many methods have been developed to incorporate unnatural amino acids into metalloproteins, and these methods can be classified as either biological or synthetic in nature. Biological approaches rely on cellular machinery for placement of the unnatural amino acid within the protein sequence. Recent advances in molecular cloning and biology are the essence of these techniques. The simplest of these is cavity complementation in which a metal ligand is mutated to Gly or Ala to create a cavity that can then be titrated by exogenous ligands. The use of auxotrophs (bacteria lacking the ability to produce a specific amino acid) can be used for global replacement of a natural amino acid with a structurally similar unnatural analog. Finally, and more excitingly, it is now possible to use the amber stop codon and tRNAs that have been specially

ARTICLE CONTENTS

• Approaches to Unnatural Amino Acid Incorporation into Proteins
• Approaches for Non-Native Cofactor Attachment
• Conclusions
selected to charge with an unnatural amino acid to place the unnatural amino acid exactly at the coded position in the peptide sequence (13).

Synthetic approaches use chemical transformations on either native proteins, peptides obtained by solid-state synthesis, or both. Chemical modification is the simplest of these and requires the treatment of the selected peptide with a chemical reagent that alters reactive amino acid side chains such as Ser, Lys, or Cys (14). Because that particular amino acid at all positions of the protein is modified, it is difficult to make the chemical modification site-specific. Total solid phase peptide synthesis allows for site-specific incorporation of unnatural amino acids at any specific location (13), but it is limited in length, because the synthesis efficiency decreases with increased length of the peptide chain. This limitation has been eased by native chemical ligation which exploits the reactive properties of an N-terminal cysteine to connect two or more shorter synthetic peptides covalently via a peptide bond (16). The biological and synthetic approaches are complimentary and the combination of these two approaches has proven very useful. One excellent example is the semisynthetic technique of expressed protein ligation that couples a bacterially expressed peptide with a synthetic peptide in a fashion similar to that used for native chemical ligation (17). This hybrid of synthetic and biological techniques allows easy site-specific introduction of unnatural amino acids into much longer peptide sequences (18).

**Design and engineering of metallocproteins that contain unnatural amino acids using the biological approach**

The method of cavity complementation has been employed effectively to introduce unnatural amino acids as metal ligands into the oxygen storage protein myoglobin (11, 19) and into the electron transfer protein azurin (20). In the case of myoglobin, the proximal histidine that coordinates to the iron atom was removed, which generated a cavity that was titrated by neutral thiol and thioether ligands. Introduction of these unnatural amino acids demonstrated that ferrous heme iron proteins can be ligated by a neutral cysteine. These observations are important because they demonstrate that ferric heme proteins that apparently lose cysteinate ligation on reduction may in fact be ligated by neutral cysteines. Thus, ligation of the heme iron by unnatural amino acids has demonstrated the existence of a ligation state previously only hypothesized to exist and which significantly alters our understanding of the electronics and reactivity of the very important P450 class of enzymes. In the case of azurin, removal of one of the copper ligating histidines generated a cavity near the solvent exposed surface of the protein. It was shown that altering the type of unnatural amino acid titrated into the protein produced variations in the spectroscopic characteristics that encompass the known range of mononuclear copper-thiolate proteins. From this study, it became apparent that the protein scaffold has a preference for a few distinct geometries of the metal center. It is also important that the spectroscopic properties of the native protein could be restored by titration of the cavity with imidazole, which indicates that the introduction of the cavity had not altered the protein structure irreversibly. The use of unnatural amino acids has thus demonstrated that despite the wide variation in naturally occurring copper proteins, a few specific metal ligand geometries may be responsible for all variations in function.

The auxotrophic method has been used in the study of the monoxygenase P450 enzymes to replace all of the methionine residues with norleucine (Nle) (21). Although the Met to Nle substitution resulted in decreased thermal stability of the enzyme, a roughly two-fold increase in peroxysyngase activity was observed. This was noteworthy because peroxysyngase activity, which is the oxidation of organic substrates using hydrogen peroxide, is normally inefficient in the monooxygenase P450s. The increase in activity was attributed to the removal of the methionine residues, which are oxidized easily by hydrogen peroxide, resulting in enzyme inactivation. This study demonstrates the power of unnatural amino acids to tune the properties of native enzymes by introducing novel properties without sacrificing natural catalytic function.

The incorporation of unnatural amino acids into metallocproteins via RNA by using the amber stop codon has also contributed to the design of artificial metalloproteins. Using this technique, a para-cyano-L-phenylalanine was incorporated into myoglobin in place of the distal histidine (22). The cyano group on this unnatural amino acid proved to be a probe of ligand binding and local electronic environment, which distinguishes between linear and bent configurations of different diatomic heme ligands (i.e., O2, NO, and CO). This report demonstrated site-specific incorporation of a very useful probe via unnatural amino acids that might also find application in the study of protein folding and conformation, electronic fields within proteins, or biomolecular interactions. Site-specific modification of cytochrome c3 with a redox mediator was also accomplished using the amber stop codon encoding for unnatural amino acids (23). To achieve this goal, a para-propargyloxyphenylalanine unnatural amino acid introduced on the surface of the protein was reacted with an azido-viologen. The covalent anchoring of the redox active viologen near the hemes of cytochrome c3 was shown to increase the amount electron transfer both to the solution and the protein cofactor by three fold. This report expands the number of modifications that can be made site specifically in a metalloprotein because the unnatural amino acid used can react with any azide-bearing molecule. The development of a system for introducing a selectively reactive unnatural amino acid near a c-type heme is important and applicable to the study of other heme proteins. These studies demonstrate the use of the RNA method for incorporating unnatural amino acids, and an increase in the number of studies using this technique is anticipated.

**Design and engineering of metallocproteins that contain unnatural amino acids using synthetic techniques**

An excellent example of unnatural amino acid introduction via a synthetic technique may be found in work performed on the protease subtilisin. Chemical conversion of a catalytically active serine to a selenocysteine altered the catalytic activity...
significant (24). Instead of native hydrolysis activity, which is greatly suppressed, selenosubtilisin catalyzes the reduction of peroxides in a highly enantioselective fashion, and it has even been used for resolution of functionalized racemic peroxides with up to 99% enantiomeric purity in the alcohol produced (25). The methodology for incorporation of selenium on the catalytically active serine position has even been extended down the periodic table to incorporate tellurium (26) at this position. The new tellurometallase displays similar reactivity to selenosubtilisin but with altered kinetic properties. The dramatic changes in reactivity observed in subtilisin on incorporation of unnatural amino acids. Although limited by the size of protein it can make, the total synthesis of proteins has contributed significantly to the study and the design of metalloproteins. Replacement of two histidine residues in a de novo designed four-a-helix bundle with 4-L-(pyridyl)-L-alanines resulted in −6 × 104 weaker binding of the ferric heme and a near 300 mV increase in the reduction potential (28). Changing this same position to 1-methyl-L-histidine causes >7 × 103-fold decrease in ferric heme affinity but only 125-fold decrease for ferrous heme (29). The methylated histidine unnatural amino acid also prevented the favored heme coordination geometry and resulted in a five-coordinate high-spin ferrous heme similar to decysmyoglobin that could bind CO reversibly but not O2 (28). These substitutions of unnatural amino acids reveal the preferential binding of different redox states of the heme cofactor and provide insight into the importance of the protein environment for the reduction potential (28). A stabilizing effect of such a hydrogen bond on the reduced form of the protein had been predicted to increase the reduction potential. The replacement of the amide in question with an ester linkage lowered the reduction potential of the iron cluster by ∼100 mV, in agreement with the predicted stabilizing effect for this hydrogen bond. A gain, the use of unnatural amino acids confirmed the existence of subtle interactions between the protein environment and the metal center, and it emphasizes the necessity of including such interactions in protein design and engineering efforts. It is also important to note that an alteration such as exchanging an amide for an ester was impossible with conventional mutagenesis because all natural amino acids contain an amide linkage.

Figure 1 Controlling coordination geometry and metal binding properties using unnatural amino acids in de novo designed four-a-helix bundles. (a) Substitution of histidine with 4-L-(pyridyl)-L-alanines or 1-methyl-L-histidine. (b) Substitution of cysteine with penicillamine. (Adapted from Reference 1, p. 120 copyright 2005 and Reference 27. Cover art copyright 2006 with permission from Elsevier and Wiley, respectively.)
Artificial Metalloproteins: Design and Engineering of

Design and engineering of metalloproteins containing unnatural amino acids using semisynthetic techniques

Expressed protein ligation (EPL) has been used recently to introduce unnatural amino acids into metalloproteins to examine the effect of individual amino acids on fine-tuning the properties of the metal binding site. In the first report of EPL on a metalloprotein, the Cys in the type 1 blue copper electron transfer protein azurin has been replaced with selenocysteine, which resulted in marked changes in the electronic absorption and electron paramagnetic resonance (34, 35). However, little change was observed in the reduction potential. Conversely, when the axial methionine ligand in azurin is substituted by its structural analogs selenomethionine, norleucine, oxomethionine, trifluoromethionine, or difluoromethionine (Fig. 3), very little change is observed in the spectroscopic properties, but large changes in reduction potential are reported (>200 mV) (36, 37). A linear relationship between the reduction potential and the hydrophobicity of the axial ligand demonstrates that hydrophobicity is the dominant factor contributed by the axial ligand to tuning the reduction potential (36). Extrapolation of these results to studies with conventional mutagenesis demonstrate nearly identical effects of the axial ligand hydrophobicity on the reduction potential in all type 1 blue copper proteins (36). Type 1 blue copper proteins are one of the most common classes of electron transfer proteins, which play critical roles in processes such as photosynthesis and oxidative respiration. The role of the conserved axial methionine has been probed by mutagenesis using all other 19 natural amino acids, but the exact role remained undefined because conventional mutagenesis changes more than one factor, such as electronic and steric factors, at the same time. Using isostructural amino acids, the exact effect of the methionine is more readily defined, and observation of the above trend is only possible by introducing unnatural amino acids using EPL.

Although it is not used to study the metal binding site, EPL has been used to expand the DNA recognition abilities of a zinc-finger protein (32) by replacement of arginine with citrulline (Cit) (Fig. 2b). DNA-binding specificity of a zinc-finger is determined primarily by three amino acids. One such amino acid residue, Arg, recognizes the 5′-most guanosine (G) via donation of two hydrogen bonds. Replacement of the Arg with natural amino acids resulted in similar (as with Lys, which like Arg has only hydrogen bond donors) or decreased specificity.
Similar imidazole terminated rhenium molecular wires of a photogenerated radical with the active site of P450cam covalent electron tunneling wire that allows direct interaction between the heme and ruthenium complex, intermediates that are very difficult to see using the protein’s native redox partners were observed. Implementation of these unnatural cofactors may be far reaching, and it may allow detailed study of many biological electron transfer processes by removing the diffusion limits and time resolution restrictions of more conventional techniques.

An artificial catalyst that consists of a monoclonal antibody, 1G8, and the achiral rhodium complexes against which it was raised, exploits noncovalent interactions to promote enantioselective catalysis (41). Not only did the antibody bind tightly to the non-native cofactor, but it also provided a chiral environment. Hydrogenation of 2-acetamidoacrylic acid proceeded with 98% ee. The substrate specificity was also very high, and the antibody prevented the hydrogenation of larger substrates. Inclusion of a substrate mimic during antibody generation should improve the activity and substrate scope for the artificial catalyst. Additionally, rational design of the original antibody isolated is also possible. As the first report of asymmetric hydrogenation by a transition metal-immunoglobulin complex, this approach holds promise for future development of tightly bound non-native cofactors as catalysts.

The noncovalent approach has been used to replace the native heme in heme proteins such as myoglobin, with non-native protoporphyrin IX modified at the propionate groups. Removal of the native heme followed by reconstitution with the propionate-modified protoporphyrin IX has led to myoglobins that display increased negative charge (42), flavin redox cofactors (43), and glycosylation (44) on the heme edge (Fig. 4). These modified hemes modulate a host of properties such as protein-protein interactions, protein-protein recognition, and electron transfer. Extensive reviews that contain the specifics of each heme modification are recommended to the reader (9). The propionate-substituted porphyrins allow much more direct contact between the heme and the solution outside the protein than occurs with only native heme. In a similar strategy, introduction of a porphycene in place of the native heme led to a myoglobin with dramatically enhanced dioxygen affinity (45). These authors show that direct alteration of the cofactor may be more effective than conventional mutagenesis for improving the properties of a native metalloprotein. The role of the formyl substituent on heme a has also been probed by comparing heme b and a heme a mimic inside a four-α-helix bundle (46). It was demonstrated that a major function of the formyl substituent is to raise the reduction potential of the heme, which it does by destabilizing the binding of the ferric state (47). This work with non-native cofactors is valuable because it is difficult to probe the thermodynamics of binding, oxidation reduction events, and variance of the heme type simultaneously in the native system. Lastly, a four-α-helix bundle that binds a nonbiological metalloporphyrin selectively over native heme has been reported, which demonstrates that proper design can actually favor the non-native cofactor (48, 49). The design of a peptide that can bind a non-native cofactor selectively takes us closer to the rational design of a metalloprotein from the ground up and provides an important testing ground for our knowledge of protein structure and function.
Artificial Metalloproteins: Design and Engineering of

R1 = Me, Et or n-Pr
R2 = H or t-Bu
R3 = H or Me
M = Mn or Fe

Figure 4 Introduction of non-native metal Schiff base and substituted heme cofactors by noncovalent attachment in myoglobin. (Adapted from Reference 8, p. 6, 7, copyright 2007 with permission from Elsevier.)

Reconstitution of heme proteins with modified hemes relies heavily on the strong protein-heme interaction and incorporation of non-native cofactors that bear less structural similarity to heme is more challenging. The metal ligand affinity of the proximal His in myoglobin for a Cu(II) N-salicylideneaminoacido compound has been used for noncovalent attachment of this complex (50) demonstrating that noncovalent attachment to a protein is possible even if the new metal cofactor does not resemble heme. Although the metal Schiff base complexes Mn(III)salen and Cr(III)salophen are roughly planar and similar in dimension to heme, a structural-based design to modify both the metal complex (through substitution of the 5 and 5′ positions) and protein (mutation of Ala71 with Gly) was required (Fig. 4) (51) to obtain cofactor incorporation. Crystal structures with an Fe(III)salophen derivative reveal disorder of the non-native cofactor induced by steric repulsion from Ala71 (52). In catalytic sulfoxidation, the modification of His64 to Asp allowed the Mn(III) derivative to produce enantioselectivity up to ∼30% (53). Such detailed analysis of an artificial cofactor in a protein has allowed these authors to propose substrate-binding locations and to delineate substrate and cofactor interactions that may be responsible for the catalytic activity. Application of similar techniques allowed the heme oxygenase incorporation of Fe(III)salen (54). This non-native cofactor composite demonstrated the importance of a hydrogen bond to Arg177 for rapid electron transfer in heme oxygenase. Furthermore, the direct participation of the non-native cofactor in electron transfer reactions via the conserved interactions of the native system demonstrates the potential for regulation of non-native cofactors by the native protein environment.

A more general noncovalent attachment method is the use of biotin for attaching cofactors to avidin or streptavidin. This technique was first demonstrated by Wilson and Whitesides (55) who generated a biocatalyst capable of asymmetric hydrogenation of α-acetamidoacrylic acid with up to 44% ee using a biotinylated achiral diphosphinerhodium(I) complex and avidin. Lin et al. expanded this concept with a chiral Pyrphos-Rh(I) complex and hydrogenated itaconic acid with up to 48% ee (56). Recently, Letondor et al. (57) have demonstrated the use of this system by the noncovalent attachment of various diphosphine rhodium, ruthenium and iridium complexes to avidin and streptavidin (Fig. 5). The subsequent tuning of the linker and
Design and engineering of metalloproteins containing non-native cofactors through covalent attachment

The covalent attachment of non-native cofactors to a protein host has the advantage of site-specific incorporation and high yield and is usually achievable with minimal structural modification to the complex or protein host, even if the protein affinity for the cofactor is minimal. A Cu(I)-I. 130 phenanthroline, an efficient DNA cleavage agent, was covalently attached via a cysteine to several DNA-binding proteins (Fig. 6a). The resulting combination of efficient DNA cleavage by the cofactor and selective DNA positioning by the protein makes an excellent artificial nuclease. The covalent attachment of this same complex to an active site Cys in the adipocyte lipid-binding protein cavity has provided high enantiomeric selectivity for several reactions and good conversion. Examples include hydrolysis of acetamidocarbamic acid with 96% ee (58), transfer hydrogenation of p-methyl acetoephone with 94% ee (59), and oxidation of aromatic alcohols (60). A similar system has also been subjected to directed evolution for hydrolysis of acetamidocarbamic acid (61). The stepwise improvement of selectivity to 65% ee demonstrates that the noncovalent attachment of the metal cofactor can be coupled to biological techniques for improvement of the artificial metalloprotein.

**Conclusions**

Great progress has been made in the design and engineering of artificial metalloproteins. The introduction of unnatural amino acids has created new probes of physiologic activity and provided methods for fine-tuning protein properties not available with conventional biochemical techniques. Such techniques have also been used to elucidate the subtle role of key residues in protein metal binding sites as was never possible before. The incorporation of non-native cofactors has enhanced our understanding of protein design by revealing factors, both covalent and noncovalent, that govern cofactor binding and enzyme enantio- and chemoselectivity. Last, novel function has been introduced in metalloprotein constructs, with directed evolutionary, rational design, and combinatorial techniques for the design of non-native metalloproteins has a bright future in biotechnological and pharmaceutical applications.

![Diagram](image-url)
Artificial Metalloproteins: Design and Engineering of Non-Native Metal Cofactors

Figure 6. Introduction of non-native metal cofactors by covalent attachment. (a) Computer model of a phenanthroline complex bound to adipocyte lipid binding protein. (b) Computer model of a dual covalently attached Mn Salen. (Reproduced from Reference 62, p. 11644, copyright 1997 and Reference 63, p. 10812, copyright 2004 with permission from The American Chemical Society.)

References


Further Reading


See Also

Expanding the Genetic Code Through Chemical Biology

Metallo-enzymes and Metallo-Proteins, Chemistry of

Proteins, in vivo Chemical Modification of

Protein Engineering: Overview of Applications in Chemical Biology
Click Peptides, Design and Applications of
Yoshiaki Kiso and Atsuhiko Taniguchi  Department of Medicinal Chemistry, Kyoto Pharmaceutical University, Yamashina-ku, Kyoto, Japan
Youhei Sohma  Department of Physical Chemistry, Kyoto Pharmaceutical University, Yamashina-ku, Kyoto, Japan
doi: 10.1002/9780470048672.wecb632

A "click peptide" is a chemically modified peptide analog used to understand the biologic function of peptides. A click peptide does not exhibit the inherent biologic activity of the original peptide because of a simple chemical modification of the peptide backbone and, by adding an exogenous action ("click") such as pH-change or photo-irradiation, easily affords the native peptide in situ with a quick and easy one-way conversion via a native amide bond-forming reaction. For example, the designed click peptides of Alzheimer's amyloid β peptide (Aβ) 1–42, which did not exhibit a self-assembling nature because of a backbone isomerization from a native Gly25-Ser26 bond to a β-ester bond, could migrate to the intact Aβ1–42 under physiologic conditions by pH- or photo-triggered "click" via an O→N intramolecular acyl migration reaction. This click peptide overcomes the difficulties in handling Aβ1–42 in syntheses and biologic experiments, which clarifies the currently unexplained processes of Alzheimer's disease.

The use of intact peptides in an in vitro experiment sometimes causes considerable discrepancies in the biologic data because of the difficulties in handling and controlling the properties of peptides such as low water-solubility and highly aggregative nature. A "click peptide" is a chemically modified peptide analog used to understand the biologic function of peptides while overcoming the problems associated with the peptide properties (Fig. 1) (1–3).

A click peptide has the ability to:

- Control the following natures of the original peptide.
  - Physicochemical property (e.g., water-solubility, self-assembly, aggregation, or folding)
  - Biologic activity (e.g., ligand-receptor binding affinity, or enzyme-substrate binding affinity)
- Convert to the native peptide in situ by an exogenous action ("click").
- Examples of such an action include pH-change and photo-irradiation.
- Because a rapid amide bond formation is required, an intramolecular reaction is desired.
- A n atom-economical reaction should be selected as the conversion reaction to avoid a side effect derived from coreleased products in biologic experimental conditions.

Additionally, in the design of the click peptide, a simple modification of the peptide backbone is desired to minimize difficulties in chemical synthesis of the click peptide and the conversion process to the native peptide.

Biologic Background

The highly aggregative feature of peptides is a significant obstacle against establishing a reliable in vitro biologic experimental system to investigate the major causative agents of diseases. Recent amyloid β peptide (Aβ)-related studies have encountered such problems because of their highly aggregative nature. A more clear understanding of the pathologic mechanism of Aβ would be of great value to discover novel drug targets against Alzheimer's disease (AD). Aβ is the main proteinaceous component of amyloid plaques found in the brain as a pathognomonic feature of AD (4), and it has been found to be neurotoxic in vitro and in vivo (5). Several studies have supported the hypothesis that neurotoxicity and the kinetics of Aβ1–42 aggregation are related directly to the assembly state in solution. However, the pathologic self-assembly of Aβ1–42 in amyloid plaque formation, a process currently unexplained, is very difficult to demonstrate in vitro because of its uncontrolled self-assembly. The synthesized Aβ1–42 in
Itself contains various oligomeric forms (6), and Aβ1–42 undergoes time- and concentration-dependent aggregation in aqueous TFA (trifluoroacetic acid)-acetonitrile solution used in high performance liquid chromatography (HPLC) purification (7). Moreover, the Aβ1–42 monomer forms aggregates easily even in a standard storage solution such as dimethylsulfoxide (8). This uncontrolled self-assembly in an in vitro experiment causes considerable discrepancies in the biologic data. As a result of its highly aggregative nature, difficulties in handling Aβ1–42 have hampered the progress of Aβ1–42-related AD research such that constructs have been used instead of the native peptide.

Recently, it has been ascertained that the pathologic self-assembly nature of inherent peptides or proteins is one major event that leads to the development of many diseases such as prion protein in prion disease, α-synuclein in Parkinson’s disease, and islet amyloid polypeptide in type 2 diabetes, as well as Aβ1–42 in AD (9).

**Chemistry**

We designed and synthesized the click peptide based on the O-acyl isopeptide (Fig. 21–3). The click peptide has an O-acyl isopeptide structure of the corresponding native peptide. Hydrogen bond interactions between peptide chains often play a crucial role in the onset of physicochemical and biologic actions through their contributions to the conformational stability of the higher order structure. In the click peptide, a newly formed ester bond results in the inhibition of the ordered hydrogen bond interactions, which lead to conformational changes, and mask the activity of the native peptide. Additionally, the target peptide was generated via an O→N intramolecular acyl migration reaction. The O→N intramolecular acyl migration is a very attractive chemical reaction for the click peptide. The O→N intramolecular acyl migration is capable of rapid amide bond formation under physiologic conditions (pH 7.4) because of the energetically favorable intramolecular five-membered ring intermediate. It is an atom-economical reaction; thus, no byproduct is released during conversion to the parent peptide, which is a great advantage in toxicology in biologic experimental systems.

**O-Acyl isopeptide method**

In 2003, we discovered that the presence of an O-acyl instead of N-acyl residue within the peptide backbone changed the secondary structure of the native peptide significantly (10). The coupling and protection efficiencies improved during solid-phase peptide synthesis of peptide derivatives that possess “difficult sequences,” most likely because the O-acyl isopeptide prevented the formation of secondary structures that disfavor coupling. Generally, the difficult sequences are hydrophobic and are prone to aggregate in solvents during synthesis and purification. This aggregation is attributed to inter/intra-molecular hydrophobic interactions and the hydrogen bond network among resin-bound peptide chains, which results in the formation of extended secondary structures such as β-sheets (11). In addition, the target peptide was generated subsequently by an O→N intramolecular acyl migration reaction. This finding marked the beginning of the development of the “O-acyl isopeptide method” for peptide synthesis. The method has begun to be used by several other groups, especially Mimna et al. (12), Coin et al. (13) and Börner et al. (14), which indicates that the O-acyl isopeptide method is widely advantageous for peptide preparation.

**Synthesis of O-acyl isopeptide**

The synthetic procedure of O-acyl isopeptides has been well established (10). O-acyl isopeptides can be prepared easily using Fmoc-based solid-phase peptide synthesis. After Boc-Ser(Thr)-OH is coupled to a peptide-resin to obtain Boc-Ser(Thr-peptide-resin), Fmoc-Xaa-OH is esterified to the β-hydroxy group of Ser(Thr) residue using a DIC (1,3-diisopropylcarbodiimide–DMAP (4-dimethylaminopyridine) method. A thorough esterification on the resin might induce epimerization at the esterified amino acid residue, the use of an O-acyl isopeptide unit, which can be synthesized easily in solution, could avoid this problem (15). The protected O-acyl isopeptide-resin is obtained through coupling of additional amino acid residues using conventional techniques. During Fmoc removal of the second amino acid residue at the N-terminal side next to the ester bond, caution should be used to prevent the formation of diketopiperazines. Finally, the desired O-acyl isopeptide is obtained by
An application of the method to Alzheimer’s Aβ1–42 revealed that the O-acyl isopeptide of Aβ1–42 (in which the Gly10-Ser18 sequence of Aβ1–42 was isomerized to a β-ester bond) could be synthesized effectively and stored without spontaneous self-assembly (1-3). Intact monomer Aβ1–42 could then be obtained from the isopeptide under physiologic experimental conditions. The water solubility of the TFA salt of the isopeptide was 100-fold higher than that of Aβ1–42 (0.14 mg mL⁻¹). Purified isopeptides could be converted quantitatively to Aβ1–42 via O–N intramolecular acyl migration in phosphate buffered saline (pH 7.4) at 37°C with a half-life of 1 minute (Fig. 3), whereas the TFA salt of the isopeptide was stable at 4°C in both solid state and dimethylsulfoxide solution. The Aβ1–42 isopeptide proved to be easier to synthesize and purify than Aβ1–42; moreover, it provided an opportunity to prepare Aβ1–42 in situ rapidly under physiologic conditions. The spontaneous self-assembly of Aβ1–42 could thus be more studied effectively to elucidate the inherent pathologic functions of the aggregative Aβ1–42 in AD. Moreover, slower migration (t1/2 = 3 hours) was observed at pH 4.9, and no migration at pH 3.5 after incubating for 3 hours. These results suggest that this pH-dependent O–N intramolecular acyl migration enables pH-triggered click for controlled in situ production of an intact Aβ1–42 from the click peptide.

Photo-triggered click peptide of Aβ1–42

Furthermore, we have synthesized a photo-triggered click peptide of Aβ1–42, in which a photocleavable 6-nitroveratryloxy-carbonyl group (16) was introduced at the ε-amino group of Ser14 in the Aβ1–42 isopeptide (Fig. 4, 2, 3). The photo-triggered system had been employed to provide well-defined, time-controllable, and position-selective activation by light irradiation when studying biologic systems, because light is the fastest medium and can be focused to a restricted area. In size-exclusion chromatography, oligomers of Aβ1–42 increased in quantity with incubation time (pH 7.4, 37°C) at the expense of the monomer. On the other hand, the photo-triggered click peptide remained in the monomeric form after 24 hours of incubation. Similarly, thioflavin-T fluorescence intensity (17), which corresponds to the extent of fibril formation, increased with incubation time in the case of Aβ1–42, yet unchanged in the case of the photo-triggered click peptide after 24 hours of incubation. Our results indicated that the click peptide was nonaggregative. The isopeptide structure resulted in complete inhibition of the aggregative nature of Aβ1–42. Photo-irradiation of the click peptide and subsequent O–N intramolecular acyl migration afforded intact Aβ1–42 rapidly in situ. In the absence of light, the click peptide was stable during storage. This method may provide a useful system to investigate the biologic dynamics of Aβ1–42 in AD with high spatial and temporal resolution by photochemically inducible activation of peptide self-assembly.

Chemical Tools and Techniques

An experimental system that can prepare a bioactive peptide in situ is advantageous to investigate biologic functions. The click peptide would be a useful tool to establish such an experimental system. As a tool with a similar concept, a “caged” compound is a synthetic molecule whose biologic activity is masked by a covalently attached photocleavable protecting group, and affords the bioactive molecule by photo-irradiation. Generally, such a photo-triggered system is considered to be advantageous for studying the dynamic processes of peptides and proteins, because upon photoactivation, only a short duration of time is required to control the spatiotemporal dynamics.
of the native compounds (18, 19). A fundamental drawback of the caged strategy on large peptides and proteins is that a small photocleavable group does not always mask biologic activity, because high potency may be attributed to large sections of the peptide structure (19, 20). This drawback can be overcome by the “click peptide” strategy because the native properties are masked in the isopeptide and released by O-acyl to N-acyl migration. This advantage should open new doors for the development of novel and useful photo-triggered tools to probe systems in chemical biology and medical science.

Because difficulties in handling Aβ1–42 in syntheses and bio-
logic experiments would hamper the progress of Aβ1–42-related AD research, we expect that the “click peptide” method will help to clarify the currently unexplained processes of AD. Moreover, many amyloidogenic diseases, as well as Aβ1–42 in AD, have recently attracted much attention and are being studied to discover novel drug targets (9). Thus, we hope that the “click peptide” strategy would be applied widely to these amyloid-related peptides or proteins. Click peptide can serve as a tool to study peptide folding and aggregation in chemical biology-oriented research, because acyl migration can be induced chemically or photochemically to release the native peptide.

References


Figure 4 A photo-triggered click peptide. The production of Aβ1–42 by photo-triggered click followed by O→N intramolecular acyl migration reaction.

Peptide-bond formation requires activation of a carboxyl group followed by aminolysis of the activated carbonyl. The reagents and methods employed for activation, the intermediates and activated forms generated and their aminolysis are described. Rationalization of the strategy employed to avoid isomerization during coupling is presented. Approaches for minimizing aggregation that interferes with coupling are also given.

Peptides are employed for therapeutic purposes and for research in the development of new drugs, vaccines, biomarkers, diagnostic agents, and so on. The demand for these peptides has experienced double-digit annual growth over recent years (1). Most peptides are obtained by chemical synthesis. Chain assembly is effected by successive addition of single residues always starting from the carboxy-terminus, which eliminates the danger that isomerization might occur during coupling. If the peptide is long enough, then two or more segments may be combined to produce the final product. Moreover, the chain may be assembled in solution or on a solid phase. Many methods are available for linking two amino acids together. Each has its own peculiarities as well as favorable and unattractive features. Some methods of coupling are of general use, some are not applicable to solid-phase synthesis, and some are employed only for specific purposes. The choice of method depends on circumstances that vary considerably; every method seems to find its place among the armaments employed for synthesis. An intriguing observation is that five different coupling methods are involved in the synthesis in kilogram amounts of the nine-residue drug Atisoban (2).

**Peptide-Bond Formation**

A peptide bond is formed by reaction of the carboxyl group of one amino acid with the amino group of a second amino acid (Fig. 1A). A combination of the two occurs when the carboxyl group is activated, which means it has been rendered deficient in electrons or electrophilic, and the amino group is electron rich or nucleophilic. The amino group is nucleophilic if it is not protonated; the carboxyl group is activated by an electron-withdrawing moiety to it. The latter is achieved by use of compounds called coupling reagents or by one or more chemical reactions. Peptide-bond formation is slower when the activated residue is valine, isoleucine, or threonine protected on the side chain; these are known as hindered residues. During the process of aminolysis, a competing reaction may occur (Fig. 1B) because the carbonyl on the amino group of the activated residue is also nucleophilic. The result is the formation of a five-membered cyclic structure, which is referred to as a 5(4H)-oxazolone that contains nitrogen and oxygen atoms and a carbonyl group, with substituents on carbons 2 and 4. The substituent at C-4 is the side chain of the amino acid; the substituent at C-2 originates from the acyl group of the activated residue. When the acyl group is another amino-acid residue, that is, if the activated component is a peptide, then the structure is a 2-alkyl-5(4H)-oxazolone that is chirally unstable. It readily tautomerizes (Fig. 1F) to give a mixture of the two enantiomers of the oxazolone. When the acyl group is that of the protector of an amino acid, the oxazolone that is formed is a 2-alkoxy-5(4H)-oxazolone, which is chirally stable and does not isomerize (3). These alkoxy-carbonyl protectors (ROCO, often referred to as “urethane” protectors, which is incorrect; urethane = ROCONH₂), namely benzoyloxycarbonyl (Cbz or Z; benzyl carbonyl), tert-butyloxycarbonyl (Boc) cleaved by aci-dolysis, and 9-fluorenylmethoxycarbonyl (Fmoc) cleaved by beta-elimination have been designed, in that order, so that the activated residue does not isomerize during couplings. The effect of the tautomerization of the 2-alkyl-5(4H)-oxazolone is the production (Fig. 1E) of some other isomer of the peptide product because the oxazolone also undergoes aminolysis. Hence, coupling of an N-alkoxycarbonylamino acid occurs without enantiomerization of the residue because even if the 2-alkoxy-5(4H)-oxazolone is formed, it does not isomerize. However, there is a danger to produce epimeric peptides during the coupling of an N-protected peptide because of the severe tendency of the activated peptide to form the 2-alkyl-5(4H)-oxazolone that isomerizes readily and is also a precursor of the product. In all cases, the product of isomerization is a mixture of epimeric peptides. When the activated residue is isoleucine or threonine that have two stereogenic centers and when the activated component is a peptide, which does
Coupling Methods for Peptide Synthesis

Activated Forms

A multitude of activated forms of amino acids and peptides is available (Fig. 2), which have varying stabilities. Some are shelf-stable reagents; some can be prepared and manipulated, but they are not stable enough to be stored; and others are intermediates that have been postulated to explain the results. In fact, the most popular coupling reagents generate activated intermediates that have been postulated to explain the results. When the activated component bears an N-alkoxycarbonyl group and the amino acid is other than glycine, isoleucine, or threonine, the process is an enantiomerization (4), but it is also referred to as racemization. The process of isomerization of an activated peptide is also often referred to as racemization, but it is careless usage of the term and an unwarranted extension of the meaning of the word.

Auxiliary Nucleophiles

Peptide-bond formation involves reaction of the nucleophilic amino group of one residue with the activated carboxyl group of a second residue (Fig. 3). In many cases, the activated residue reacts first with another molecule to produce a compound that is activated by another moiety; the latter then undergoes aminolysis to give the peptide. These molecules are referred to as auxiliary nucleophiles. The original ones were incorporated into coupling reactions when it was realized that their presence had beneficial effects. The common auxiliary nucleophiles (Fig. 3) (6-9), which are also known as additives, appear in Fig. 4. Each possesses a hydroxyl group; all except the HOSu are slightly acidic with pKs in the 3.4-4.6 range. They participate in coupling reactions in one of two forms. Either they are added in their original form HOWN, or they are present by virtue of the fact that the coupling reagent employed is composed of the deprotonated moiety -OW attached to a positively charged atom that is a strong electrophile (Fig. 5). HOSu serves primarily as a nucleophile for the preparation of activated esters (10); that is, it is added to an activated amino acid without the simultaneous presence or subsequent addition of an amino-bearing residue. The other auxiliary nucleophiles are added to carboximide-mediated reactions to suppress the side reactions of isomerization (7) (Fig. 1) and N-acylurea formation (Fig. 4). They protonate the intermediates that are formed, namely the O-acylisoureas (Fig. 4) and the oxazolone (Fig. 1), which prevents their rearrangement and tautomerization, respectively, that lead to undesired consequences. The protonated intermediates are strong nucleophiles that readily yield the peptide product. Hence, the additives trap activated intermediates that prevent the side reactions. When the additives are incorporated into coupling reagents, they are released as the anion -OW as soon as the reaction is initiated. The anion is a strong nucleophile that can react with or trap any activated intermediates that are formed. The compounds generated are activated esters that undergo aminolysis to give the peptide. The traditional additive has been HOBt (7); HOAt (HOABt) (9) is of more recent development. HOObt (8) was not employed in the past because of the known side reaction of aminolysis at the carbonyl of the additive. However, this reaction occurs in negligible amount, and the additive has received renewed interest. Although it was dormant for many years, HOBt (7) has surfaced with a vengeance as it is being employed in multikilogram amounts in the synthesis of cephalosporin and the fusion inhibitor Fuzeon. It and HOAt

Figure 2: A multitude of activated forms of amino acids and peptides is available, which have varying stabilities. Some are shelf-stable reagents; some can be prepared and manipulated, but they are not stable enough to be stored; and others are intermediates that have been postulated to explain the results. In fact, the most popular coupling reagents generate activated intermediates that have been postulated to explain the results.

Figure 3: Activated forms of amino acids and peptides are often referred to as auxiliary nucleophiles. The original ones were incorporated into coupling reactions when it was realized that their presence had beneficial effects. The common auxiliary nucleophiles, which are also known as additives, appear in Fig. 4. Each possesses a hydroxyl group; all except the HOSu are slightly acidic with pKs in the 3.4-4.6 range. They participate in coupling reactions in one of two forms. Either they are added in their original form HOWN, or they are present by virtue of the fact that the coupling reagent employed is composed of the deprotonated moiety -OW attached to a positively charged atom that is a strong electrophile.

Figure 4: Each possesses a hydroxyl group; all except the HOSu are slightly acidic with pKs in the 3.4-4.6 range. They participate in coupling reactions in one of two forms. Either they are added in their original form HOWN, or they are present by virtue of the fact that the coupling reagent employed is composed of the deprotonated moiety -OW attached to a positively charged atom that is a strong electrophile.
Coupling Methods for Peptide Synthesis

Activated forms of N-alkoxycarbonylamino acids (RCO = R1OCONHCHR2CO) and Nα-protected peptides. Some are aminolyzed (A) to give the peptide, and some may generate (B), a second activated form that is also a source of the product. Experience indicates that acyl halides and symmetrical anhydrides of peptides do not exist.

Figure 3 Auxiliary nucleophiles (additives) that suppress side reactions by reacting with activated intermediates, and the two acylated forms of benzotriazole that are produced.

are often superior to HOBt as additives; their hydroxyl groups are slightly more acidic, which provides better leaving groups in the esters. HOcbt is less expensive. The addition of HOBt to a carbodiimide-mediated reaction minimizes the dehydration of side-chain amides that occurs in its absence.

Coupling Methods

Peptide-bond formation is achieved in two ways. In the first approach, a reagent is added to a mixture that contains a component with a free carboxyl group and one with a free amino group. Such coupling reagents include carbodiimides, EEDQ (Fig. 4), andonium-salt based reagents (Fig. 5). The other approach involves preparing the activated component after which the amino-bearing component is added. Here, one can distinguish two separate cases. In one case, the activated form is a shelf-stable compound such as an activated ester or an N-carboxyanhydride (NCA). In the other case, the activated form is employed as soon as it is prepared. Examples include MxAn as well as SyAn and azides.

Dialkylcarbodiimides

The compounds implicated in carbodiimide-mediated reactions (11, 12) appear in Fig. 4A. The reaction is postulated to be initiated by protonation of one of the slightly basic nitrogen atoms by the acid followed by attack at the central carbon atom by the acid anion that generates the 0-acylisourea (OAU). Aminolysis the OAU gives the peptide with liberation of N,N′-dialkylurea. A side reaction of N-acyl-N,N′-dialkylurea (NAU) formation by 0- to N-acyl shift of the OAU can occur. The latter is avoided by adding an auxiliary nucleophile that protonates intermediates. The acyl shift is promoted by heat and tertiary amine that assures a deprotonated nitrogen atom, and delay in the aminolysis caused by bulkiness in the reacting residues. The three common carbodiimides are dicyclohexylcarbodiimide (R4 = R5 = C6H11, DCC) (11) giving urea DCU, disopropylcarbodiimide (R4 = R5 = C(CH3)2, Dic) giving urea
DIU, and ethyl(dimethylamino)propyl)carbodiimide hydrochloride \( R^2 = C_2H_5, R^3 = (HCl) (CH_3)_2NCH_2CH_2CH_2, EDC \) giving urea EDU (12). All carbodiimides are allergic, which cause skin irritation after contact with DCC. The acid displaces the ethoxy group after which spontaneous rearrangement to the MxaN occurs with expulsion of quinoline. Aminolysis of the MxaN is the source of the peptide. Unavoidable aminolysis at the other carbonyl gives a small amount of irreversibly substituted amino component as impurity.

Symmetrical anhydrides (Fig. 2)

In a variant (15) of the carbodiimide method, two moles of an N-alkoxycarbonylamine acid are reacted with one mole of DCC in dichloromethane (DCM). After 30 minutes, the DCU is removed by filtration and the solution is mixed with the aminolysing component. The DCM is sometimes replaced with dimethylformamide (DMF) for the aminolysis step. SyA/N are less reactive and hence more selective but gives clean reactions. Some SyA/N have been isolated and stored (16, 17), but the practice is no longer recommended for synthesis in solution because EDU and the NAU are sol-uble in water. It is available as crystals that are expensive. Re- actions are usually effected by adding the reagent to a mixture of the reacting components and the auxiliary nucleophile if desired. In a variant, the acid is allowed to react with itself and possibly an additive after which the amino-containing component is added. In this case, the N-alkoxycarbonylamine acid is converted into the SyA/N and the activated ester, respectively, prior to the aminolysis. The efficiency of a carbodiimide-mediated coupling of a peptide that is effected in the presence of an aux-iliary nucleophile is enhanced by the presence of cupric ion (13). Cupric ion prevents a 2,4-dialkyl-5-OH l-oxazalone from isomerizing (13).

1-Ethoxycarbonyl–2-ethoxy-1,2-dihydroquinoline (EDEQ)

This variant (14) (Fig. 3B) is employed in the same way as carbodiimides and occasionally by operators who have devel-oped a skin irritation after contact with DCC. The acid displaces the ethoxy group after which spontaneous rearrangement to the MxaN occurs with expulsion of quinoline. Aminolysis of the MxaN is the source of the peptide. Unavoidable aminolysis at the other carbonyl gives a small amount of irreversibly substituted amino component as impurity.

Mixed anhydrides (Fig. 2)

The MxaN method is popular for synthesis in solution; it is inexpensive and simple to effect. Chloroformate \( R^2OCOCl \) \( R^2 = \text{ethyl, } C_2H_5 \text{, or isobutyl, } \text{CH}(CH_3)_2 \) is added to a small excess of the acid at \(-5^\circ C\) in the presence of tertiary amine and after 30–60 seconds, the aminolysing component is added (38). The reaction is complete within 30–60 minutes. Traditional solvents have been tetrahydrofuran and DMF; the favored ter-tiary amine is N-methylmorpholine (NMM). NMM is believed to act as an acceptor of the acyl group that forms the acylmor-pholinium ion (RCO-N'N'MM), which is the acylating reagent. It was thought for many years that halo-containing solvents were unsuitable; however, DCMM is a good solvent provided the ter-tiary amine is not triethylamine as this combination retards the formation of the MxaN (19). Strictly anhydrous conditions are also not essential because a MxaN can be purified by washing it with aqueous solutions (19). MxaN have not been employed in solid-phase synthesis because a side reaction of aminolysis occurs at the wrong carbonyl (Fig. 4B) in the range of 1–10% that is unavoidable. The larger amounts develop when the acti-vated residue is hindered. The urethane produced is stable to all conditions used for deprotection; hence, the impurity can be readily eliminated. Isospropyl chloroformate seems to be a super-ior reagent for MxaN reactions (20). MxaN can be employed for acylation of an amino acid or peptide with an unprotected carboxyl group in polar or partially aqueous solvent.
Coupling Methods for Peptide Synthesis

Figure 4  (A) Compounds implicated in carbodiimide-mediated reactions. The acid is postulated to add to the carbodiimide generating the O-acylisourea. Aminolysis occurs at the point indicated by the dashed arrow. (B) Compounds implicated in EEDQ-mediated reactions. The acid is postulated to displace the ethoxy of the reagent generating \( \text{MxAn} \) which rearranges to the \( \text{MxAn} \). Aminolysis (a) gives the desired peptide and (b) the ethoxycarbonylated nucleophile as impurity.

Figure 5  Structures and examples of nomenclature for phosphonium and carbonium salt-based reagents. Rules of nomenclature dictate that the latter be named as modified ureas; see text for revised structures; “ino” indicates a ring linked at the nitrogen atom.

reagents are more attractive for solid-phase synthesis. The chemistry of the uronium-based reagents is complex. In fact, the BTU of HBTU and TBTU does not have the structure given in Fig. 5, but it is \( 1-(\text{bis}(\text{dimethylamino})\text{methylene}-1-H\text{-benzotriazolylinium} \, 3\text{-oxide, that is, the tetramethylcarbenium moiety} \, \text{is linked to the other nitrogen atom of the ring, analogous to structure 5 (Fig 3) (26). Acylated HOBt had originally been shown to exist in solution as two forms in equilibrium, the ester (RCO}_2\text{Bt}) \, \text{and the N-acyl N'-oxide (RCO}_2\text{Bt} \to \text{O}) (6). In model experiments, the uronium forms of HBTU and HATU reacted more quickly and led to less epimerization than the guanidinium forms (27). Other reagents are continuously being developed, typical examples being Bsc-\, or Bsc-1 (28) and PyCOP (Fig. 5) (29). Reagents based on HODhbt have also been described. All have their unique properties. The azabenzotriazolyl-containing reagents seem to be particularly effective for performing difficult couplings (30). Acyl halides (Fig. 2)

A straightforward method of coupling involves conversion of an Fmoc-amino acid (Fmoc-Xaa-OH) to the chloride Fmoc-Xaa-Cl, using thionyl chloride \((\text{SO}_2\text{Cl})\) in DCM (31, 32) or triphosgene \((\text{O}=\text{C(OCCl}_3)\), which is a source of phosgene \((\text{COCl}_2\)) (33). Only the latter allows preparation of derivatives with acid-sensitive side-chain protectors \( \text{Bzl} \) but not \( \text{Boc} \). Also available are \( \text{Bzl} \), \( \text{Boc} \), and Fmoc-amino-acid fluorides obtained from the acids using cyanuric fluoride \((\text{C}_3\text{N}_3\text{F}_3)\), diethylaminosulfur trifluoride \((\text{DAST, Et}_2\text{NSF}_3)\), or tetramethylfluoroformamidium hexafluorophosphate \((\text{TFHH, Me}_4\text{N}_2\text{C}+\text{HF}_6^-\text{which is activated in DMF by tertiary amine. The amino-containing component is added after removal of excess reagent and side products. TFFH is employed for generating the fluoride in the presence of the incoming nucleophile. The Fmoc-Xaa-Cl are highly activated reagents that generate the strong acid HCl on aminolysis. Neutralization by tertiary amines causes formation of the 5(4H)-oxazolone that is aminalized more slowly, so it is best to avoid this structure by use of KOBt for neutralization or by destruction (reduction) of the acid with powdered zinc (34). For this and other reasons, Fmoc-Xaa-Cl are not used routinely, but they are particularly suited for couplings between hindered residues, and for esterifying a hydroxymethyl-linker-resin. Acyl fluorides are less reactive, hence, are more stable to oxygen nucleophiles such as water or methanol; they have a lower tendency to cyclize to the 5(4H)-oxazolone and can be employed in the absence of base. Their properties resemble those of activated esters. They are also suited for coupling hindered residues, but the possible deleterious effects of prolonged exposure to base because of their lesser reactivity must not be disregarded. Acyl azides (Fig. 2)

The preparation of acyl azides involves conversion of acid \( \text{RCO}_2\text{H} \) to ester \( \text{RCO}_2\text{Me} \) and subsequent aminolysis with
performed with chains. For coupling peptides, activation of the hydrazide is necessary to create products that are enantiomerically pure. Caution: Azides are highly reactive species because of the phenomenon of anion-p binding of two chains because of the tendency of functional groups to form hydrogen bonds (41). Succinimido esters (10) are unique because of the reaction of aqueous AcOH that contains NaNO₂ or tert-butyl nitrite. The acyl azide (RCON ₃) that is formed is extracted into an organic solvent, which is dried and added to the aminalizing component. An acyl azide can be produced directly from the amino acid in the presence of the aminalizing component by use of diphenyl phosphorazidate [(PhO)₂PON₃] (36), which is a reagent that is especially good for converting a linear peptide to a cyclic peptide. The acyl azide reaction is generally applicable to Boc- or Z-Xaa-OH, which includes instances when Xaa = Ser, Thr, or His with unprotected side chains. For coupling peptides, activation of the hydrazide is performed with tert-butyl nitrite (tBuONO₂) and nitrosyl chloride (NO₂Cl) (37). Because hydrazine destroys protectors such as trifluoroacetyl (CF₃CO₂H) and nitrilotriaceticacid (NTA) by converting the hydrazide to an activated ester that allows aminalization to occur in 4 hours instead of 2-3 days (38). Acyl azides are sensitive to tertiary amines; however, they do not form oxazolones, hence, proper manipulation of the coupling species just about guarantees products that are enantiomerically pure. Caution: Azides tend to explode under certain conditions.

Activated esters (Fig. 2)

Two types of activated esters exist, those in which the activating moiety is a substituted phenol and those in which the activating moiety is a substituted hydroxylamine (Fig. 2). Common examples of the former are p-nitrophenyl, pentachlorophenyl, and pentafluorophenyl; common examples of the latter are succinimido, benzotriazolyl, 7-azabenzotriazolyl, and 4-oxo-1,4-dicyclohexylbenzotriazolyl (Fig. 3). All (RCO₂W) react with an amino group (NH₂R) to produce tetrahedral intermediate R¹'O(C=O)W-N⁺H₂R⁻ that collapses to the product R¹-C(ONSu) which is a superactive triazine ester from acids (43). A unique form of activation of an amino acid but not of a peptide is the N-carboxyaminohydride (NCA) (Fig. 2), which is the cyclic anhydride formed between the carboxyl group and a carbonyl group bound to the amino group. These anhydrides are obtained by reacting the amino acid with phosgene in tetrahydrofuran (44). An addition of Xbb-NCA to H-Xaa– gives peptide H-Xbb-Xaa– after acidification to release CO₂ from the amino-terminus (45). It is possible that the NCA oligomerizes, so the reaction must be effected under very controlled conditions of pH and temperature. For this reason, NCAs are not employed routinely. An option that eliminates the dimerization is use of the N-alkoxycarbonyl- or "urethane-" protected amino-acid NCA, Boc, Cbz or Fmoc serving as useful substituents (46). These structures also are not employed routinely but have proved advantageous in certain cases. They decompose in the presence of tertiary amine.

Tactics for Avoiding Aggregation

The coupling of a protected residue to a peptide chain is inefficient or occasionally fails once the chain has been extended to five or more residues (47). Based on the observation that replacement of the part that would normally be hindered by the binding of two chains because of the tendency of functional groups to form hydrogen bonds (41), it has been postulated that the problem occurs because of the chain folding on itself or the binding of two chains because of the tendency of functional groups to form hydrogen bonds (41). The resulting aggregation interferes with the merging of the reacting groups. Tactics to prevent the formation of hydrogen bonds and hence aggregation have been developed. These tactics involve elimination of the hydrogen atom on a peptide bond at the fifth or sixth residue of a chain and the use of microwave energy. Peptides that contain serine and are insoluble in aqueous medium can be administered to an organism in the soluble O-acyl form H-Ser(peptide ₁)-peptide ₂-OH that immediately undergoes O-to-N acyl shift to H-peptide ₁-Ser-peptide ₂-OH at above neutral pH (49).
N-Alkylation of the Peptide Bond

Elimination of the hydrogen atom at a peptide residue to prevent hydrogen bond formation involves replacing it with a functional group that can be removed at the end of the synthesis. The protector must be stable to all the conditions employed during assembly of the chain, be easy enough to introduce onto the amino group, and must allow efficient coupling of the residue bearing the two N-protectors as well as the subsequent aminolysis reaction by the N-alkylamino-acid derivative. Producing such a protector has not been an easy task. The protector of choice for Fmoc/tBu chemistry is 2-hydroxy-4-methoxybenzyl (Hmb) (50) (see Figure 6). Coupling is achieved by HOBt-assisted aminolysis by H-Xaa- of the pentafluorophenyl ester of the N,O-disubstituted-derivative 8 giving 9. The two protectors are removed, the next residue is introduced as the SyAn [(Fmoc-Xcc)2O], which acylates the 2-hydroxy group of the Hmb giving 10. The residue then migrates spontaneously to the amino group via the O-to-N acyl shift giving N-substituted peptide 11. The protector of choice for Boc/Bzl chemistry is 2-hydroxybenzyl.

A different approach that involves N-alkylated residues to prevent hydrogen bonding is the use of cyclic structures called pseudo-prolines (51). These structures are compounds formed by reaction of serine, threonine, or cysteine residues with formaldehyde, acetaldehyde, or an acetone equivalent. The products are oxazolidines or thiazolidines that bear a carboxyl group and one or more methyl groups on the other carbon atoms of the rings (Figure 7). The pseudo-prolines are employed in synthesis as the Fmoc-dipeptides, with the cyclic structure at the carboxy-terminus. The Fmoc-amino acid is combined with the terminal residue either before or after formation of the ring. The pseudo-proline-containing dipeptide is coupled to the amino group of the peptide that is being assembled. At the end of the synthesis, the pseudo-proline is converted back to its precursor amino acid by acidolysis (Figure 7). The strength of acid necessary is dictated by the nature of substituents R1, R2 that originate from the carbonyl compound and the hetero atom O or S of the ring. Methyl groups render the cyclic structures more sensitive to acid; the thiazolidine ring is much more stable to acid than the oxazolidine ring.

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
Microwave-assisted couplings

The new technology of facilitating chemical reactions by use of microwave energy has been applied to the synthesis of peptides (52, 53). The theory is that the energy excites the growing peptide chain, which prevents aggregation during chain extension. Coupling times are substantially reduced, and products of enhanced purity have been obtained.

Other Peptide-Bond Forming Approaches

A new approach to the synthesis of activated esters of peptides is through alkyl thio esters $\text{R(O-S-CH}_2\text{)}$ spacer-$\text{H}$ is released by acidolysis with HF at the bond in-}

\text{\text{nuller-resin is assembled using Boc/Bzl chemistry ex-}

\text{cept for the last residue, thio ester $\text{Fmoc-Xcc-Xxx-n-Xaa-S-CH}_2\text{ spacer-H}$ is released by acidolysis with HF at the bond in-}

\text{icated by the arrow and converted into activated peptide $\text{Fmoc-Xcc-Xxx-x-Xaa-S-CH}_2\text{ O-dithiotetracyclo[2.2.1.0$-linker-resin is assembled u-}

\text{sing Boc/Bzl chemistry ex-}

\text{cept for the last residue, thio ester $\text{Fmoc-Xcc-Xxx-n-Xaa-S-CH}_2\text{ spacer-H}$ is released by acidolysis with HF at the bond in-}

\text{icated by the arrow and converted into activated peptide $\text{Fmoc-Xcc-Xxx-x-Xaa-S-CH}_2\text{ O-dithio}$.}

References

1. V. Glaser, Bioprocessing: market growing for custom-made pep-


2. C. Johansson, L. Blomberg, W. Hlebowicz, H. Nicklasson, B.

Nilsson, L. Andersson, Industrial production of an aspartic


3. N. L. Benoiton and F. M. F. Chen, 2-Alkoxy-5(4-H$\text{N}$)-

$\text{O}$-oxazolones from N-alkoxycarbonylamino acids and their implica-

tion in carbodiimide-mediated reactions in peptide synthesis. Can. J.


4. N. L. Benoiton and F. M. F. Chen. Sometimes it is neither a

racemisation nor an epimerisation but an enantiomerisation. A

plea for preciseness in the use of terms describing denomina-


5. N. L. Benoiton and F. M. F. Chen. Not the alkoxyacyanoborato-


343–345.


7. W. König and R. Geiger. A new method for the synthe-

sis of peptides: Activation of the carbonyl group with dicy-


8. W. König and R. Geiger. A new method for the synthesis of

peptides: Activation of the carbonyl group with dicyclohexylcar-


9. L. A. Carpino, L-Hydroxy-7-azabenzotriazole. An efficient pep-


10. Anderson JW, Zimmerman JE, Callahan FM. The use of esters of


12. J. Palecek, Carbodiimides. In: M. Goodman, F. Arthur, L. Mo-

roder, and C. Tonolo (eds.), Synthesis of Peptides and Pep-

tidomimetics, Houben-Weyl Methods of Organic Chemistry, E


Simultaneous use of 1-hydrobenzotriazoles and copper(I) chlo-

ride as additives for racemization-free and efficient synthesis by


308–314.


15. H. Hagenmierl and H. Frank. Increased coupling yields in

solid phase peptide synthesis with a modified carbodiimide cou-

pling procedure. Hoppe Seyler’s Z. Physiol. Chem. 1972;353:


16. F. M. F. Chen, K. Kuruda, and N. L. Benoiton. A simple prepara-

tion of asymmetrical anhydrides of N-alkoxycarbonylamino


17. E. J. Heimer, C. Chang, T. Lambros, and J. Meienhofer. Sta-

ble isolated symmetrical anhydrides of $\text{N}$-$\text{9}$-fluorenylmethylox-

yl carbonyl amino acids in solid-phase peptide synthesis. Int. J.


18. G. W. Anderson, J. W. Zimmerman, and F. M. F. Chen. A rein-

vestigation of the mixed carbonic anhydride method of peptide


19. F. M. F. Chen and N. L. Benoiton. The preparation and reactions

of mixed anhydrides of $\text{N}$-alkoxy carbonyl amino acids. Can.


20. N. L. Benoiton, Y. Lee, and F. M. F. Chen. Isopropyl chlo-

roformate as a superior reagent for mixed amide hydroxide gene-


reagents IV (1) - Benzotriazole N-oxytrisdimethylamine phos-


1219–1222.

22. V. Dourtoglou, Ziegler J.-C., and B. Gross. L’hexafluoro-

phosphate de $\text{O}$-benzotriazolyl-$\text{N}$, $\text{N}$-tetramethyluronium hexa-

fluorophosphate: a new and efficient peptide coupling reagent. (In


24. J. Coste, M.-N. Dufour, D. Le Nguyen, and B. Castro, BOP and

congeners: Present status and new developments. In: J. Rivier, R.

Garland, and R. Marshall (eds.), Peptides Chemistry, Structure,


25. J. Coste, J. Kandasamy, B. Castro, and L. Carpio, Oxygenated

peptide coupling reagents for $\text{N}$-methylated amino acids.


26. A. Baldeney, F. Alliberia, L. A. Carpino, B. M. Foxman, and

S. A. Kates. Structural studies of reagents for peptide bond


Further Reading


See Also

Amine Acids, Chemical Properties of Click Peptides, Design and Application of Peptide Combinatorial Libraries, Solid-Phase Synthesis of Biomolecules
In this review, an overview is given for the synthesis of lipidated peptides such as those belonging to the Rab, Ras, eNOS, and hedgehog proteins. The different approaches for the synthesis of specific lipidated peptides (palmitoylated, prenylated, and myristoylated) are discussed with special emphasis on solid-phase synthesis methods, because these methods have turned out to be the preferred synthesis method for most required peptides. Different solid-phase methods are described that are individually suited for different types of lipidated peptides, differing, for example, in lipidation pattern or amino acid side-chain functionality and in their protein ligation functionality. For the described solution approaches, the block coupling strategies followed for the different lipidated peptides are highlighted. A special section in this review discusses the different synthetic tools for the incorporation of the lipid functionalities in the peptides. Although a generally validated strategy for the synthesis of lipidated peptides does not exist, because of the large number of different functionalities, this review aids in the conceptual design of synthetic protocols for lipidated peptides. Finally, the methods for the ligation of lipidated peptides to proteins are shortly discussed, as these methods also significantly influence the design and synthesis of lipidated peptides.

The continued existence of cells crucially depends on their ability to receive signals from the environment and to respond to them in a suitable way. This process of signal transduction, i.e., the forwarding of signals from the extracellular matrix via the cytoplasm into the nucleus, involves, among others, proteins that are modified with structural hydrophobic features attached to the peptide backbone. Posttranslational lipidation by acylation with fatty acids, alkylation with prenyl moieties, and conjunction of more complex lipid components, for instance, the glycosylphosphatidylinositol (GPI) moiety, occurs on a wide variety of intracellular and extracellular signaling proteins. An example of proteins containing lipid moieties are membrane-attached proteins. This review provides a concise overview of the synthesis of lipidated peptides from these membrane proteins and discusses possible methods for generating intact proteins with these peptides. The family of Ras proteins is one of the most relevant examples of lipidated membrane proteins from both their abundance and their biological and pharmaceutical importance and is therefore used as a general motif in this review [1].

General Considerations For Lipidated Peptide Synthesis

Five predominant lipidation motifs can be found on proteins, resulting in increased complexity from the stable and simple myristoyl group that can be found on N-terminal glycines to the GPI anchor featuring, apart from the lipid group, a complex sugar motif and phosphate groups (Fig. 1). The three most common types of posttranslational lipid modifications! encoun-
tered in the Ras superfamily, for example, are N-myristoylation, S-palmitoylation, and S-isoprenylation (2–4), of which palmitoylation is the only reversible one (5). These lipid modifications are important for the correct biological function of Ras proteins, as they require the localization to the inner leaflet of the plasma membrane and are generally found at the C-terminus of the protein. As for other lipidated proteins, until recently, most ex-
periments with these proteins, such as structure determination by X-ray diffraction, nuclear magnetic resonance (NMR), and bio-
chemical characterization, were performed on the soluble part of the proteins, i.e., the protein without the C-terminus. The results obtained via these approaches can, however, at best only give
an approximation of the real situation, especially because diversity between the individual members is often encountered in the lipidated part of the protein. The biochemical generation of fully functionalized and modified lipidated proteins is, however, difficult and time-consuming. In the case of S-palmitoylation, it leads to heterogeneous mixtures and is, therefore, in most cases, not practical or applicable. Therefore, in recent years, chemical biological approaches have been developed that give access to fully functional lipidated peptides and proteins, together with additional non-natural modifications, which was achieved via two techniques that were developed more or less in parallel: 1) the progress in the field of protein ligation and chemical synthesis of proteins (6) and 2) the progress in the field of lipidated peptide synthesis, both in solution and on solid support. The development of these synthetic methods with special attention to the solid-phase methods are the focus of this review. Synthetic strategies giving access to lipidated peptides have been developed during the last 10 years and have been reviewed both for solution and for solid support approaches (7, 8). This review is therefore not an all-inclusive review, but it highlights how different types of lipidated peptides can be most effectively synthesized.

The synthesis of lipidated peptides can follow different strategies such as the use of solid-phase versus solution-phase techniques, the use of lipidated building blocks versus peptide lipidation, and the application of the Boc versus Fmoc strategy as well as enzymatic (9-12) and noble metal (13) sensitive methods. The decision regarding which strategy to follow will generally depend on the types and positions of the lipid groups in the peptides, the length of the peptide, the C-terminal functionality, the presence of additional functional groups such as fluorescent markers or spin labels, and the purification strategy. Particularly regarding the lipids and functional groups, the synthetic strategy has to be fully adapted to the reactivities of these modifications. For solid-phase chemistry, the choice of the linker to the resin is an additional criterion to consider in the synthetic planning (Fig. 2).

For lipidated peptides, some general guidelines have emerged: 1) (strong) Acid-labile protecting groups or linkers cannot be used in combination with prenyl groups. 2) Prenyl groups are not compatible with protecting groups that have to be removed under hydrogenolytic conditions. 3) Thiuesters such as palmitoyl groups attached to cysteines are liable toward nucleophilies. 4) Non-natural functional groups such as fluorophores, photoactivatable groups, and tags for ligation to proteins generally impose synthetic limitations. 5) Prenylated dipeptide esters undergo rapid diketopiperazine formation upon N-terminal deprotection. 6) The palmitoyl group on an N-terminally unmasked cysteine rapidly undergoes an undesired S,N-acyl shift.

An additional issue that requires attention during the synthetic process is the purification of the lipidated peptides. Both final products and intermediates are often badly soluble in both water and organic solvent or feature detergent-like characteristics, which limits purification via both silica gel chromatography and reversed-phase high-performance liquid chromatography (HPLC). Purification can therefore be time-consuming and can result in low yields. However, the strongly dissimilar nature of the coupling partners, for instance, with respect to hydrophobicity, can also be used to facilitate the separation of product and starting compounds or the separation of double-lipidated from single-lipidated peptide.
Considering these features and limitations, it is not surprising that a general strategy for the synthesis of lipidated peptides, such as the Fmoc- or Boc-strategy for normal peptides, does not exist. However, several tools have been developed such as protecting groups, resin linkers, and synthetic approaches that have allowed some general strategies to emerge with which most lipidated peptides can be synthesized. This review tries to guide the reader to some generally validated approaches that meet the restraining demands imposed by the peptides. In our experience, solid-phase approaches offer fast and flexible entries to multiply lipidated peptides; therefore, the main focus will be on these solid-phase approaches.

### Introduction of Lipids Into Peptides

Introduction of lipid functionalities into peptides can follow two general strategies, independent of solid-phase or solution-phase chemistry. Either the lipid groups are introduced via coupling of prelipidated amino acids to the peptide (Fig. 3a), or they are introduced via selective lipidation of a peptide (Fig. 3b-e). Both the use of prelipidated building blocks (14–16) and of resin lipidation (17–25) have been investigated for the synthesis of lipidated peptides on solid support.

#### On-bead lipidation

Lipidation on the solid support requires an orthogonal protecting group strategy that allows the stepwise introduction of different lipid groups on the same peptide such as a farnesyl group and a palmitoyl group (Fig. 3c) (20, 21). This approach generally uses a large excess of the lipid groups, which becomes problematic if the lipid groups need to be synthesized (e.g., in modified form) and requires an extensive protecting group strategy. A possible solution to overcome these problems might be the application of the recently developed reversed approach of S-farnesylation and S-palmitoylation to solid-phase chemistry (26). In this approach, lipid groups are introduced via the S\textsubscript{2} displacement of a bromide by reaction of a thiol containing lipid as nucleophile with bromoalaine-containing peptides as electrophile in solution (Fig. 3a). Another elegant approach to introduce a prenyl thioether into peptides could be the use of a chemoselective conjugate addition of an thiolate nucleophile (e.g., farnesylthiolate) to a dehydroalanine (Fig. 3b,27). The absence of stereoselective control in this conjugate addition can be solved using aziridine-2-carboxylic acid-containing peptides (Fig. 3e) (28).

These newly developed methods for the incorporation of lipidated amino acids might prove beneficial in future solid-phase lipidated peptide synthesis.

#### Prelipidated amino acid building blocks

An approach to overcome the issues associated with on-resin prenylation for the synthesis of lipidated peptides relies on the use of prelipidated cysteine building blocks. These prelipidated building blocks can be handled as normal amino acids in the coupling sequence and can be inserted anywhere in a sequence without the need to modify the protecting group strategy (Fig. 3e). This approach thus provides a highly flexible and generic entry to lipidated peptides (15).

### Synthesis of Lipidated Peptides

Lipidated peptides can be synthesized both in solution and on solid support. The recent focus on lipidated peptide synthesis has shifted to the preparation of lipidated peptides on solid supports as it intrinsically offers faster and more flexible entries to the target peptides (25). Issues that should be considered for solid-phase lipidated peptides synthesis are, for example, the screening of different linkers and cleavage conditions as well as evaluation of synthesis routes for the introduction of the lipid groups with a suitable orthogonal protecting group strategy. By analogy to the solution-phase approaches, protecting groups must be orthogonal to the lipids and eventually other functional groups in the peptides. Additionally, the linker to the resin should be cleavable under conditions not interfering with the peptide functionalities. In an ideal setup, the cleavage of the linker would also allow for the introduction of different types of functional groups at the C-terminal carboxyl function.

Here, both solution-phase approaches and solid-phase approaches undertaken for the synthesis of specific lipidated peptides are reviewed (Fig. 4). For the solution-phase approaches, the focus is on the coupling strategy applied, i.e., which block couplings were pursued. The solid-phase approaches are reviewed with the focus on the type of linker to be used for specific types of lipidated peptides.

#### Small lipidated model peptides for biophysical investigations

Small cysteine-containing peptides similar to sequences of S-acylated proteins were generally synthesized in solution. The
Lipidated Peptide Synthesis

Figure 3 Different approaches for the introduction of lipid functionalities, here exemplified via the farnesyl group, into peptides. (A) Lipidated amino acid building blocks. (B) Substitution of bromoalanine with a nucleophile. (C) Alkylation or acylation of a free thiol functionality of a cysteine. (D) Conjugate addition of a nucleophile (e.g., farnesylthiolate) to a dehydroalanine. (E) Conjugate addition of a nucleophile to aziridine-2-carboxylic acid containing peptides.

tetrapeptide model Bimane-SC(S^tBu)RC(Far)OMe representative for the carboxy terminus of H-Ras and featuring the Bimane fluorophore was prepared in solution using Fmoc-chemistry, introducing the farnesyl group at the stage of the Fmoc-protected dipeptide (29). The acylated peptides (Myr)GCX-Bimane (X = G, L, R, T, V), which are found in certain nonreceptor tyrosine kinases and \( \alpha \)-subunits of several heterotrimeric G-proteins, were synthesized in solution using common solution-phase peptide synthesis with \( N \)-myristoylglycine as the building block (29). These model peptides were acylated with Palmitoyl-CoA in phospholipid vesicles at physiological pH. For such uncatalyzed spontaneous reactions, only a modest molar excess of acyl donor species (2.5:1) was necessary. Unprotected side chains of threonine or serine do not interfere with this \( S \)-acylation.

Lipidated peptides incorporating the C(GerGer)XC(GerGer)-OMe motif found in several Rab and homologous proteins were also synthesized in solution via Fmoc-chemistry following cysteine deprotection and geranylgeranylation (30).

C-terminal lipidated peptide of the influenza virus hemagglutinin A

The influenza virus hemagglutinin A contains a lipidated peptide fragment at the C-terminus featuring two palmitoylated cysteines and two amino acids with a polar side chain. This peptide has been synthesized in solution and the strategy for this synthesis was based on the fragment condensation of the lipidated tetrapeptide TIC(Pal)I that was coupled N-terminal to the palmitoylated dipeptide RC(Pal) and, after deprotection, elongated with NBD-aca-labeled methionine (31). For this block coupling strategy, a set of three orthogonal protecting groups was required, whereby the use of base/nucleophile-labile and hydrogenolytically removable protecting groups was not permitted. The use of the acid labile Boc group for the N-terminus, the enzymatically removable PAOB ester for the C-terminus, and the Pd0-sensitive Alloc group for the arginine side-chain function turned out to be a successful combination.

C-terminal lipidated peptides of H-/N-Ras

A large body of work has been devoted to the synthesis of the C-terminal lipidated peptides of the small GTPTases H-/N-Ras. As such, these peptides have been synthesized in solution, via combined solution and solid support approaches and completely on solid phase.

For the synthesis of a small library of palmitoylated and isoprenylated N-Ras peptides in solution, a modular strategy was adopted, with the tetrapeptide MGLP as the key intermediate. This tetrapeptide intermediate allowed for further elongation at its C-terminus with lipidated or nonlipidated cysteine methyl esters, as well as the addition of various \( N \)-terminal maleimidoacryloyl (MICA)-labeled dipeptides, consisting of different GC lipidated units (32, 33). The synthesis was performed under common conditions using the Fmoc, Boc, and Alloc protecting group strategy. Using this methodology, a number of N-Ras derivatives containing natural and nonnatural lipid residues were...
Lipidated Peptide Synthesis

Figure 4
Sequences of lipidated peptides discussed in this review. The frizzled lines indicate the building blocks used for the block coupling strategy of the solution-phase synthesis (when applicable). Abbreviations: BioAcA: N-(+)-Biotinyl-6-aminocaproyl; MantAca: N-methylanthranyl-6-aminocaproyl; MIC: 6-Maleimidocaproyl; NBDAca: N-(4-nitrobenz-2-oxa-1,3-diazol-7-yl)-6-aminocaproyl.

produced, and the technique was extended to also include a number of fluorescent derivatives.

In a reversed approach, N-Ras lipidated peptides were synthesized in solution via an $\text{SN}_2$ displacement of bromoalanine containing hexapeptides with thiopalmitic acid and farnesylierane as nucleophiles (26). The synthetic route started with the dipeptides and tetrapeptides A(Br)M and GLPA(Br), both incorporating bromoalanine. A farnesylation of the C-terminal tetrapeptide, it was coupled to the N-Alloc protected dipeptide. This farnesylated hexapeptide was then treated with thiopalmitic acid for final palmitoylation.

The key feature of the solution synthesis of the C-terminal octapeptide of H-Ras containing one farnesyl thioether and two palmitoyl thioester moieties is the orchestration of the acid-labile tert-butyl ester as carboxy protecting group, the Pd0-sensitive Alloc function as amino-blocking group, and the reduction-labile tert-butyl disulfide as masking of thiol groups (13). In addition, a serine hydroxyl and, in particular, a lysine $\varepsilon$-amino group are located in the vicinity of the thioester groups, increasing the danger of $\text{S-O}$ and $\text{S-N}$ acyl migrations in the course of the synthesis. The assembly in solution was achieved by dividing the triply lipidated peptide into the two selectively C- and N-terminal unmasked palmitoylated tripeptide building blocks PG(C-Pal) and M SC(Pal), the N-terminally unmasked lysine derivative, and an S-farnesylated cysteine methyl ester. In a combined solution-phase and solid-phase approach the N-Ras lipidated peptide was synthesized by coupling the C-terminal farnesylated cysteine methyl ester to the rest of the peptide that had been assembled on solid support (34). The second cysteine was protected as a tert-butyl disulfide, thus allowing reductive cleavage under physiological conditions, and the N-terminal amine of the peptide was connected to a maleimido group for protein ligation purposes. The use of these protecting and functional groups made the final segment condensation a relatively straightforward task. An important issue that has to be kept under scrutiny when performing such segment condensations is the possibility of racemization, the selection of validated coupling techniques that avoid the use of basic conditions and polar solvents, however, has usually proved sufficient to avoid or minimize this racemization. Typically, the amino acid (4 equiv) is treated with DIC (4 equiv) and an excess of HOBt (6 equiv) in DMF/CH$_2$Cl$_2$ mixtures.

For the complete solid-phase synthesis of H- and N-Ras peptides, the hydrazide linker turned out to be the linker of choice (14). This linker is cleaved by oxidation to an acyldiazene that is then attacked by an acylazide that is then attacked by a suitable nucleophile. The linker is orthogonal to classic amide protecting groups such as Boc, Fmoc, and...
Important precaution was taken, however, to avoid an
solid-phase peptide synthesis (SPPS) with other linkers. One
Coupling conditions of the amino acids were similar to normal
palmitoylated cysteine. It was found that the
shift of the palmitoyl group after the Fmoc deprotection of the
performed using a solution of 1% DBU in DMF for 2
peptides remained stable when the Fmoc deprotection was
vide supra (50 eq.), methanol (215 eq.) CH2Cl2,O2.

Figure 5

Solid-phase synthesis of an N-Ras lipidated peptide on hydrazide resin. (A) Fmoc-Cys(Far)-OH, HBTU, HOBt, DIPA, DMF. (B) piperidine/DMF (1:4). (C) Fmoc-AA-OH, HBTU, HOBt, DIPA, DMF. (D) piperidine/DMF (1:4). (E) Fmoc-Cys(Pal)-OH, HBTU, HOBt, TMP, CH2Cl2/DMF (1:1). (F) DBU (1 %) in DMF. (G) Fmoc-Gly-OH, HATU (5 eq.), DIPEA (20 eq.) CH2Cl2/DMF (7:1). (H) Cu(OAc)2 (0.5 eq.), pyridine (30 eq.), acetic acid

Solid-phase synthesis of an N-Ras lipidated peptide on hydrazide resin. (A) Fmoc-Cys(Far)-OH, HBTU, HOBt, DIPA, DMF. (B) piperidine/DMF (1:4). (C) Fmoc-Cys(Pal)-OH, HBTU, HOBt, DIPA, DMF. (D) piperidine/DMF (1:4). (E) Fmoc-Cys(Pal)-OH, HBTU, HOBt, DIPA, DMF. (F) DBU (1 %) in DMF. (G) Fmoc-Gly-OH, HATU (5 eq.), DIPEA (20 eq.) CH2Cl2/DMF (7:1). (H) Cu(OAc)2 (0.5 eq.), pyridine (30 eq.), acetic acid

Alicoc, and racemization do not occur on cleavage. Typical nu-
clerophiles that can be used for the cleavage are amines, water,
and alcohols (35). The oxidation sensitivity of the linker does
require the coupling reactions, and especially Fmoc deprotecting
reactions, to be performed under exclusion of oxygen.
The hydrazide linker allowed, for example, the synthesis of
a completely lipidated N-Ras peptide sequence, including
an additional fluorophore attached to a lysine side chain or
N-terminal glycine and a C-terminal methyl ester (Fig. 5 (14).
Coupling conditions of the amino acids were similar to normal
do phase peptide synthesis (SPPS) with other linkers. One
important precaution was taken, however, to avoid an S-N acyl/
shift of the palmitoyl group after the Fmoc deprotection of the
palmitoylated cysteine. It was found that the S-palmitoylated
peptides remained stable when the Fmoc deprotection was
performed using a solution of 1% DBU in DMF for 2 × 30
seconds. The next coupling was then performed immediately
with preactivated amino acid and HATU as a coupling reagent.
Using similar approaches, many different peptides of the Ras
family were synthesized on the hydrazide linker, including
those containing farnesyl, geranylgeranyl, fluorescent labeled
geranyl, and palmitoyl lipids. Fluorescent markers such as NBD
and photoactivatable groups like the benzophenone group have
been introduced and Ras superfamily peptides with C-terminal
acid, ester, or amide functionalities have been cleaved from
the peptide, leading to cyclic peptides or oligomers. The more
stERICally demanding and less nuclophilic alcohol functions of,
for example, the serine side-chains feature these problems to a
much lesser extent and can generally be deprotected on the resin.
Having the amine functionalities protected after cleavage thus
requires an additional deprotection step in solution. In general,
the synthesis of lipidated peptides on the hydrazide linker is
successful for peptides with up to 30 amino acids. As such, the
hydrazide linker is one of the preferable systems for a broad
range of solid-phase lipidated peptide synthesis.

Polybasic prenylated C-terminal lipidated peptides

The synthesis of polybasic isoprenylated lipidated peptides, such as the C-termini of K-Ras4B, D-Ral, and Rho A, have been
most successful on the solid support. The solution strategies
to the peptides, generally through fragment condensation, were
difficult because of, for example, different polarities and solu-
bilities of the fragments.
The solution synthesis of the polybasic C-terminus of Rho
A required coupling of a polybasic KKK tripeptide with the
SGC(GeGor) tripeptide containing the geranylgeranylated cy-
sine methyl ester and an N-terminal coupling with a fluorescent
labeled glycine (37). For this approach, the orthogonality of the
Fmoc, Alicoc, and OBlu protecting group was applied. The same
methodology for the synthesis of the polybasic K-Ras4B peptide
failed, however, because of the low coupling yields resulting
from the different solubility properties of the fragments.
The methods of choice for the synthesis of polybasic lipi-
dated peptides such as those corresponding to the C-termini of
RhoA, K-Ras4B, and D-Ral are solid-phase approaches using
the triyl linker or the chlorotriyl linker (37, 38). The mildly
acidic cleavage conditions of these resins are still orthogonal to
the farnesyl moiety and often also to the geranylgeranyl moiety.
Attachment of the peptide via the C-terminus would limit the
C-terminal functionality to a free carboxylic acid. Therefore, the peptide can generally be attached via a lysine side chain near the C-terminus (37, 38), which avoids troublesome block condensations on the solid support with concomitant risk of racemization. The C-terminal prenylated cysteine of the pep-
tide is introduced by coupling a prelipitated cysteine to the selectively deprotected carboxylic function of the anchored lysine. After obtaining such a dipeptide on the resin, the additional steps of the synthesis can be performed in the N-terminal direc-
tion, using standard protocols and, for example, an Fmoc/Aloc strategy. The Aloc protected side chains of lysine and argi-
nine can be liberated on the solid support via cleavage with palladium(0) and an appropriate nucleophile. Treatment of the
resin with only 1% TFA and a scavenger subsequently releases the peptide while leaving the prenyl/geranylated geranyl group intact (38). A critical requirement for the use of the trityl linker in the above-described method is the presence of an anchor group in the C-terminal part of the peptide sequence that allows cleavage from the resin with 1% TFA only. Another current lim-
itation to the method is the incompatibility with thioesters, for example, palmitoylated cysteines. The deprotection of the A-
loc groups with piperidine as scavenger is not compatible with the nucleophile-labile thioester. Optimization of the type of nu-
cleophile might provide a solution, but deprotection of a large number of Aloc groups on solid support is not always a trivial matter. Finally, the Aloc deprotection is incompatible with the maleimide functionality as phosphines, used as ligands for the palladium catalyst, undergo nucleophilic addition to this func-
tionality. Nevertheless, for polybasic prenylated peptides, the use of the trityl linker is a reliable solid-phase approach.

C-terminal lipidated rab peptides

The synthesis of double geranylated peptides representing the Rab C-termini, for example the Rab7 C-terminus, has also been performed both in solution and on solid support. In general, the bivalent character of these peptides, featuring both highly hydrophobic lipid groups and polar amino acids, makes them difficult to handle. The synthesis in solution does not allow assembly of the hexapeptide by N-terminal linear elongation, for example, because of undesired diketopiperazine formation at the dipeptide stage (16). Therefore, a block-coupling method was chosen, with the dipeptide Fmoc-SC(GerGer) as a general building block (15). The dipeptide was elongated by C-terminal coupling with serine allyl ester. After deprotection of the allyl ester, the tripeptide was coupled to a geranylated cysteine methyl ester. N-Terminal elongation of the tetrapeptide to the desired hexapeptide was performed again via a block coupling, this time with the dipeptide Fmoc-CC(SBu)(E/Fm), to introduce a non-native cysteine that allows for ligation of the peptide to the protein. By employment of the Fmoc protecting group for the N-terminal amino function and the Fm protecting group for the side-chain carboxylic acid functionality of glutamic acid, a system was generated that allowed simultaneous cleavage of both protecting groups under basic conditions in a final step. The solid-phase synthesis of lipidated Rab peptides has re-
lied on the hydrazide linker, similar to the lipidated Ras pep-
tides (15). By employing geranylated/geranyl or fluorescent la-
bolated lipidated cysteines as building blocks, a highly flexible straightforward strategy with no extravagant expenses or com-
plex protecting group chemistry was developed. For the peptide assembly standard Fmoc chemistry was used. The side-chain protecting groups O-Trt and COOAll were subsequently re-
moved on the resin to leave only the N-terminal amine and thiol functionalities masked. The usage of different nucleophiles for the cleavage of the peptide from the hydrazide linker al-
lowed the introduction of a variety of functional groups at the C-terminus.

Lipidated peptides of endothelial
NO-synthase

The double palmitoylated and N-terminal myristoylated 29-mer lipidated peptide of eNOS has been synthesized both via a com-
bined solution/solid-phase approach and via a solid-phase peptide chemistry alone. For the combined technique, a combina-
tion of enzyme-labile, acid sensitive, and noble metal sensitive protect-
ing groups was chosen for the solid-phase synthesis, combined with fragment condensation in solution (39). The eNOS pep-
tide was divided into the N-myristoylated decapeptide MYRG-
LKSVGQP, the S-palmitoylated pentapeptide GPPC(Pal)IG, the octapeptide (LG)4L, and the S-palmitoylated hexapeptide
LC(Pal)IGKOG. For the synthesis and selective deprotection of the S-palmitoylated building blocks, the enzyme labile para phenyl-acetoxybenzyloxycarbonyl urethane group and the Pd(0)/ligand allyl ester were chosen as temporary N-terminal and C-terminal protecting functions. The side chains of Aas, Lys, Ser, Gin, and Glu were masked with acid-labile protecting groups to be cleaved off simultaneously in the final step of the synthesis. The entire 29-mer was finally assembled in solution by appropriate fragment condensation. The eNOS lipidated peptide was also assembled on solid sup-
port, for which the Ellman sulfonamide linker was used, which also showed good results for lipidated Ras peptides (36). The linker is stable toward acid and base, activated on alkylation, but orthogonal to the methionine thioether. The target compound is released after attack of a nucleophile (40). These cleavage con-
ditions make it orthogonal to the lipid functionalities and a good tool for the synthesis of lipidated peptides. In some cases, the Ellman sulfonamide linker seems to give even better results than the hydrazide linker, especially for long peptides with multiple lipid motifs. The double palmitoylated and N-terminal myris-
toylated 29-mer eNOS peptide is such a case. Using the prelipitated building blocks the peptide could be easily assembled and ob-
tained after cleavage. The example of the eNOS peptide shows that with good design, both the combined solution/solid-phase approach and the solid-phase approach enable excess to difficult lipidated peptides. The solid-phase approach, however, has the advantage of being significantly faster and easier.

Lipidated peptides of hedgehog proteins

The C-terminal steroid modified heptapeptides, corresponding to the C-terminus of Hedgehog proteins, were synthesized in a combined solid-phase/solution approach (41). The synthesis was built on the dipeptide Fmoc-SG-OA II that was connected to the trityl resin via the free hydroxyl functionality of the serine. The C-terminal allyl ester was cleaved by metal-mediated
Lipidated Peptide Synthesis

Model transmembrane peptides

A palmitoylated model transmembrane peptide consisting of a tryptophane-flanked poly-Leu-Ala peptide sequence similar to gramicidin A has been synthesized on solid support (22). It is an interesting example displaying the possible optimization protocol that has to be undertaken with lipidated peptide synthesis. According to the experience of reported lipidated peptide synthesis, a building block strategy was chosen. The peptide was synthesized on a Rink amide linker. In the final step of the solid-phase peptide synthesis, the coupling of the palmitoylated N-acetyl cysteine building block was not feasible. Aminolysis of the thioester was much faster than the carbodiimide-mediated peptide coupling. Introduction of the palmitoyl group via on-bead lipiddation of the cysteine required deprotection of the N-t-butylsulfanyl protected side chain. The standard deprotection reagents P(But)3 in NMP/H2O, however, led to desulfurization of the cysteine, resulting in an alanine-containing peptide instead of a palmitoylated N-acetyl cysteine. The explanation for this nearly quantitative reaction might reside in the hydrophobicity of the peptide. The phosphoneum-s-cysteine moiety is thus not able to react with H2O and p-elimination mediated by the t-butyl thiolate is the favored step. The S-(4-methoxythiophenyl)cysteine derivative proved to be the building block of choice, as it could be deprotected under mild acidic conditions, orthogonal to the other protecting groups in the peptide, which allowed for the lipiddation of the cysteine and the synthesis of palmitoylated transmembrane model peptide.

Synthesis of Lipidated Proteins Using Lipidated Peptides

Production of recombinant proteins is a key technology in the life sciences. However, progress in generation of authentic or engineered polypeptides has been more difficult for proteins that undergo posttranslational modifications, such as lipiddation. This imbalance is primarily a result of the intricacy of protein modification pathways as well as the lack of methodologies for their manipulation (42). Studies of lipidated proteins require methods that provide quantitative amounts of protein, both with natural and with new functionalities such as fluorescence, photoreactivity, spin-labeled groups, or lipid groups at non-native positions. The recently developed protein ligation methods provide the necessary platform for combining large recombinant protein scaffolds with peptides generated by organic synthesis (43–46).

Progress in the area of protein ligation methods has given access to a broad spectrum of methods for the semisynthetic production of posttranslationally modified proteins. These methods yield either native bonds (e.g., prior thiol capture, native chemical ligation, or expressed protein ligation) or non-native bonds (e.g., imine capture ligation, oxime ligation, or maleimido-diacrylic acid ligation). A variety of methods have been developed for the introduction of modified lipids and lipidated peptides into proteins. Most work in this field has been performed on Ras GTPases, for which two important approaches have been developed (Fig. 6). The first approach that incorporates synthetic lipidated peptides is based on the use of the MIC-controlled ligation (Fig. 6a) (8, 33, 34, 47, 48). This ligation requires an accessible, for example, C-terminal free thiol group on the protein, usually of a cysteine, to connect the N-terminally MIC-modified peptide. The reaction yields a non-native connection, which is, for example, in the case of Ras not problematic, because it occurs in a flexible unstructured region. Care has to be taken with this reaction that there are no other free cysteine side chains available for ligation. The second approach, Expressed Protein Ligation (EPL), connects a lipidated peptide with an N-terminal cysteine to the C-terminus of a thioester-tagged protein via a native peptide bond (Fig. 6b) (15, 16, 38, 49, 50). This reaction is in general highly selective and yields at most a cysteine mutation in the protein as non-native element. An orthogonal approach to introduce specifically prenyl moieties into proteins not relying on lipidated peptide is via an in vitro attachment of the prenyl lipids using the corresponding prenyltransferases (51). A recently explored reaction for the incorporation of lipid motifs in peptides and proteins is the copper(I)-catalyzed Huisgen’s 1,3-dipolarcycloaddition. A GPI anchor could be attached to a peptide successfully, resulting in fast uptake of the peptide into cells (52).

The protein ligation methods have been applied for the synthesis of a library of lipidated Ras GTPases with both natural modifications and non-natural modifications (Fig. 6). The first approach that incorporates synthetic lipidated peptides is based on the use of the MIC-controlled ligation (Fig. 6a) (8, 33, 34, 47, 48). This ligation requires an accessible, for example, C-terminal free thiol group on the protein, usually of a cysteine, to connect the N-terminally MIC-modified peptide. The reaction yields a non-native connection, which is, for example, in the case of Ras not problematic, because it occurs in a flexible unstructured region. Care has to be taken with this reaction that there are no other free cysteine side chains available for ligation. The second approach, Expressed Protein Ligation (EPL), connects a lipidated peptide with an N-terminal cysteine to the C-terminus of a thioester-tagged protein via a native peptide bond (Fig. 6b) (15, 16, 38, 49, 50). This reaction is in general highly selective and yields at most a cysteine mutation in the protein as non-native element. An orthogonal approach to introduce specifically prenyl moieties into proteins not relying on lipidated peptide is via an in vitro attachment of the prenyl lipids using the corresponding prenyltransferases (51). A recently explored reaction for the incorporation of lipid motifs in peptides and proteins is the copper(I)-catalyzed Huisgen’s 1,3-dipolarcycloaddition. A GPI anchor could be attached to a peptide successfully, resulting in fast uptake of the peptide into cells (52).

The protein ligation methods have been applied for the synthesis of a library of lipidated Ras GTPases with both natural modifications and non-natural modifications (Fig. 6). These lipidated Ras GTPases have served to elucidate a variety of biological questions concerning these proteins. The specific localization of Ras protein on membranes was investigated using a fluorescently labeled N-Ras protein, for example, and found to have a preference for the liquid-disordered phase over the liquid-ordered and solid-ordered phase of membranes using fluorescence studies (53). Via AFM studies, it was additionally shown that this protein is actually located in the boundary region of the domains in mixed-phase liquid-ordered/liquid-disordered bilayers. These results suggest that for N-Ras, the membrane localization is probably completely governed by the lipidated C-terminus. As such, the conformation of the C-terminus of Ras proteins bound to the membrane was investigated using NMR (54). The membrane binding of the polybasic K-Ras4B protein was also studied after protein semisynthesis (38). A photoactivatable Ras construct has been made that can be activated to selectively react with protein interaction partners (34). The functionality of these semisynthetic Ras proteins in living cells was also shown (33), enabling advanced cellular characterization of, for example, the prenylation and the palmitoylation process of the protein, using selectively functionalized proteins (55-57).
Figure 6  Schematic representation of protein ligation methods applied for the introduction of a lipidated peptide on proteins and an overview of the synthesized libraries of lipidated proteins as exemplified via Ras GTPases. (A) expressed protein ligation (EPL); (B) Maleimidocaproyl (MIC).

Enzymatic studies using selectively addressable fluorescent Rab proteins have elucidated the geranylgeranylation process (16). Structural studies on selectively prenylated Rab proteins have elucidated the exact interaction mechanism of Ras proteins with their transporter protein GDP dissociation inhibitor (GDI) and have led to a membrane extraction model for Rab proteins (58, 59).

Conclusions

Chemical biology is the application of approaches originating from chemistry and the subsequent combination of chemistry and biology to study biological phenomena previously not accessible via chemical or biological approaches alone. Lipidated peptides alone or ligated to proteins offer such entries for lipi-
Lipidated Peptide Synthesis

dated proteins. For such an integrated research activity, differently modified peptides and proteins with a variety of modifications whose structure can be changed at will through synthesis are invaluable tools. They enable experiments yielding answers in precisely molecular detail hardly accessible with biological techniques alone. Therefore, the synthesis of the lipidated peptides is an important theme and is reviewed here. The reviewed examples are used to present the synthesis of new lipidated peptides and their purification cannot always be generalized. However, with the tools currently developed and appropriate design, most types of lipidated peptides can be synthesized and obtained pure. Recently developed solid-phase synthesis methods delineate the preferred strategy to access the majority of these required lipidated peptides. With the development of these solid-phase lipidated peptide synthesis techniques, the chemical tools to obtain the majority of the required proteins are now accessible.

References


Further Reading
Peptide libraries, which consist of a few thousands to tens of millions of distinct peptides, can be generated combinatorially through biologic or synthetic approaches. The phage-display peptide library method is simple and economical, and peptides with various natural protein folds can be generated by this approach; but this biologic method generally is limited to peptides that contain only L-amino acids. Synthetic combinatorial libraries, however, can accommodate D-amino acids, unnatural amino acids, and even organic moieties, which makes these approaches highly versatile.

These synthetic methods include the spatially addressable parallel library method, the combinatorial library method that requires deconvolution, the one-bead one-compound (OBOC) combinatorial library method, the self-assembled peptide nucleic acid (PNA) encoded chemical microarrays, and the synthetic library method that requires chromatography selection. Various methods to screen these combinatorial libraries have been developed. For example, phage-display peptide libraries can be screened for the ability of phages to bind to unique immobilized target proteins or whole cells, peptide microarrays can be screened for their ability to interact with fluorescent proteins, combinatorial library methods that require deconvolution can be screened by standard solution-phase assays, and OBOC libraries can be screened by on-bead binding or functional assays or several solution-phase cell-based or biochemical assays. Combinatorial peptide library methods are enabling technologies that have proven to be very useful in basic research and drug discovery.

A combinatorial peptide library is a collection of a few thousands to hundreds of millions of different peptides. Geysen et al. (1) opened the door of the combinatorial library when they reported the first combinatorial peptide library (96 peptides) that was generated with the multipin technology in 1984. Since then, several different combinatorial approaches have been developed to generate libraries that contain many peptides with discrete sequences from which bioactive peptides can be isolated and/or identified. In the early reports in this field, synthetic peptide libraries were generated and screened against antibody molecules or streptavidin (1–4). A bout the same time, the phage-display peptide library method was created (5–7). Since 1990, several totally different combinatorial library methods have been introduced. These methods enable one to generate and screen easily and rapidly libraries with millions to billions of random peptides. The screening of combinatorial peptide libraries allows for the rapid discovery of peptide ligands or substrates against a variety of biologic targets. The principle for selecting new ligands by means of combinatorial peptide libraries exploits the natural diversity of protein-protein and protein-ligand interactions. It has emerged as a powerful tool in the study of many biologic systems. The amino acid sequence of an “active” peptide then is determined. Biologic peptide libraries, such as those that involve filamentous phage-display peptides, typically can accommodate only the 20 natural amino acids. In contrast, synthetic peptide libraries have the potential to incorporate D-amino acids and other unnatural amino acids, as well as specific secondary structures or scaffolding structures that may enhance biologic activity. In addition to amino acids, biologic building blocks such as monosaccharides, nucleotides, lipids, or even small organic moieties can be used. The advent of these peptide library methods not only facilitates the...
Peptide Combinatorial Libraries

Peptide libraries are increasingly being utilized in drug discovery processes. They are also important for the fundamental understanding of molecular recognition. In this mini-review, we will focus on the preparation, screening, structure elucidation, and application of various peptide library methods.

**Peptide Library Methods**

In essence, only two approaches exist to generate combinatorial peptide libraries: biologic and synthetic library approaches. According to the different techniques used, the synthetic library approach can be divided additionally into five methods: 1) the spatially addressable parallel library method (1, 4, 8), 2) the combinatorial library method that requires deconvolution (3, 9), 3) the one-bead one-compound (OBOC) combinatorial library method (2, 10), 4) the self-assembled peptide nucleic acid (PNA) encoded chemical microarrays (11), and 5) the synthetic library method that requires chromatography selection (12). Although the focus of this review is on peptide libraries, it is important to point out that the above synthetic methods can be applied readily to the generation of small molecule, peptidomimetic, and macrocyclic libraries.

Each peptide library method has its own advantages and disadvantages. A comparison of these methods with respect to cost, resources, and uses is shown in **Table 1**.

<table>
<thead>
<tr>
<th>Combinatorial library methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biologic library methods</td>
<td>Phage-display</td>
<td>Provides up to $10^9$ peptide entities; longer peptides (e.g., 30 amino acids) with tertiary folds possible; molecular biology and DNA sequencing technique available in many laboratories; inexpensive; available commercially and from many research laboratories; and library can be expanded, aliquoted, and stored frozen.</td>
</tr>
<tr>
<td>Spatially addressable parallel library method</td>
<td>Multi-pin technology</td>
<td>Structure determination not needed; pin library can be screened directly for binding; releasable peptides from higher loading lanterns can be screened by multiple-solution phase assays; moderately expensive pins and crowns can be used a few times without obvious loss of activity; and commercially available.</td>
</tr>
<tr>
<td>SPOT synthesis</td>
<td>Structure determination not needed; convenient on-paper assay; moderately expensive; spot synthesis equipment and custom-spot peptide membrane are commercially available.</td>
<td>Peptide library is relatively small; limited to binding and simple functional assay if bound peptides are used for the assay; releasable peptides possible, but in each in small quantity; and peptide spot membrane generally not recyclable for subsequent use.</td>
</tr>
<tr>
<td>Nanokan technology</td>
<td>Easy readout with radiofrequency tags; applicable to standard solution-phase assay; solution peptides can be used for multiple assays; and commercially available.</td>
<td>Peptide library is relatively small; inefficient unless split-mix synthesis used; peptide requires cleavage from resin for subsequent solution-phase screening; and equipment and supplies are very expensive.</td>
</tr>
<tr>
<td>The 96-deep well plate system</td>
<td>Applicable to standard solution-phase assay; solution peptides can be used for multiple assays; and robotics for automated synthesis are commercially available.</td>
<td>Inefficient synthesis unless fully automatic; equipment and supplies are expensive.</td>
</tr>
</tbody>
</table>
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Combinatorial library methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide microarray</td>
<td>Microassay possible, saving expensive and precious assay reagents; moderately expensive; replicates of peptide chips can be made; and peptide chips are available commercially or can be custom-made.</td>
<td>Linker effect; in situ synthesis not available widely; spotting technique is rapid but requires synthesis of individual compound separately; limited to on-chip binding and some functional assays; and peptide chip generally not recyclable for subsequent use.</td>
</tr>
<tr>
<td>Synthetic library methods that require deconvolution</td>
<td>Many peptides can be synthesized and analyzed rapidly; can be applied to standard solution-phase assay; inexpensive; and can be synthesized easily by experienced peptide chemist.</td>
<td>Peptide mixture; requires multiple separate iterative synthesis; not as efficient as the positional scanning method (below); and library not commercially available.</td>
</tr>
<tr>
<td>Positional scanning</td>
<td>Same as iterative approach (above) except it is more efficient and less synthesis is required; libraries can be aliquoted and used for multiple assays.</td>
<td>Peptide mixture; results may be ambiguous if multiple hits with different motifs are present in the mixture; and library not commercially available.</td>
</tr>
<tr>
<td>One-bead one-compound library method (OBOC)</td>
<td>Highly efficient synthesis and screening; each peptide is spatially separable, therefore multiple different motifs can be identified; applicable to both binding and functional assays; inexpensive; and can be synthesized easily by experienced peptide chemist.</td>
<td>Linker effect unpredictable until tested; chemical structure of positive beads has to be analyzed; and library not commercially available.</td>
</tr>
<tr>
<td>Releasable assay</td>
<td>Highly efficient synthesis; screening applicable to both binding and functional assay.</td>
<td>Releasable chemistry more involved but is still much more efficient than many other library techniques; new solid support that easily can release compounds needs to be developed; and library can not be reused and is not commercially available.</td>
</tr>
<tr>
<td>Self-assembled PNA encoded chemical microarrays</td>
<td>Library decoding on DNA chip is highly efficient; mix-split synthesis possible.</td>
<td>Synthesis of PNA tag is cumbersome; requires special DNA chip for decoding; limited to binding and simple functional assay; and not commercially available.</td>
</tr>
<tr>
<td>Synthetic library method that uses affinity chromatography selection</td>
<td>Library synthesis is easy; inexpensive; and can be synthesized easily by experienced peptide chemist.</td>
<td>High background because of nonspecific binding; only predominant motif can be identified; limited application; and library not commercially available.</td>
</tr>
</tbody>
</table>

### Biologic Library Method

The first biologic peptide library was reported by Parmley and Smith in 1988 [13]. They developed the "fusion phage" by inserting foreign DNA fragments into the encoding gene of the pil protein. This development led to the expression of an L-amino acid-containing peptide on the virion surface that did not affect virus infectivity. Each phage particle has five copies of the peptide on one end. The M13 phage is the most widely used phage-display system because of its high capacity for replication and its ability to receive large DNA inserts into its genome. Trinucleotide sequences that encode specific amino acids can be constructed and inserted into the phage genome either to increase diversity or to introduce a bias into the clone population. This very powerful method enables the researcher to routinely generate $10^8$–$10^9$ different phages. Such
Peptide Combinatorial Libraries

Peptide libraries can be panned against a target molecule by standard protocols. Other biologic peptide libraries include plasmid libraries, polynucleotide libraries, fimbria libraries, yeast-display libraries, ribosome-display libraries, and other more recently developed in vitro selection methods. For a review on biologic libraries please see Reference 14.

Biologic library methods enjoy the following advantages over synthetic peptide library methods: 1) Many peptide entities can be generated easily and economically, 2) the biologic library method can take advantage of known protein folds (e.g., immunoglobulin fold, zinc-finger fold, or conotoxin fold) by grafting random oligopeptides onto such tertiary folds, and 3) less restriction exists on the size of the peptide (e.g., 30 amino acids) one can generate. However, the biologic approaches suffer from three major disadvantages: 1) With the current systems, only the natural L-amino acids (20 eukaryotic amino acids) can be incorporated into these libraries, and incorporation of unnatural amino acids or other organic subunits into these libraries generally is not feasible, 2) although simple disulfide cyclization is feasible and longer peptides with specific protein folds (such as immunoglobulin folds) may be used, complicated bicyclic, compact scaffolding, branched structures, or molecules with a special chemistry of cyclization are impossible, and 3) the screening assays of the biologic libraries generally are limited to the binding assays (e.g., panning) and some functional assays such as protease substrate determination (see below). These methods generally are not applicable to most standard drug screens that use solution-phase assays. Synthetic peptide libraries, however, can overcome many of these limitations.

Synthetic Library Methods

Spatially addressable parallel library method

Peptide libraries can be synthesized on solid-phase support by using spatially addressable high-throughput synthetic methods (manual or automated). The amino acid sequence of each of these peptides is predetermined. The reported techniques based on this strategy include multipin technology, SPOT synthesis, NanoKan, peptide microarray, multisyringe system, and the 96 deep-well plate system. These methods are labor-intensive but can be facilitated by robotics. Depending on the library method used, the peptide library is screened either by a direct solid-phase binding assay or by a solution-phase assay. The chemical structure of the positive peptide then is decoded by location. The major drawback of this method is that only limited number of peptides can be synthesized and therefore the library generally is rather small.

Multipin technology

As mentioned earlier, the first peptide library with a limited number of peptides was generated by using multipin technology in 1984 (1). In this method, acrylic acid is irradiation-grafted onto the tip of polyethylene pins. Each pin is inserted into an adapter that fits over a 96-well polypropylene plate with a standard microtiter plate footprint. Each well of the microtiter plate then serves as a separate reaction vessel for the amino acid-coupling steps. In the original experiments, ∼50 nmol (ideally) of a single peptide was linked covalently to the spherical head of each pin. Today, the multipin system is available commercially from Mimotopes (San Diego, CA) as Pepsets. The peptide loading of each pin has been increased significantly by introducing the so-called "SynPhase™ Lantern" (e.g., 75 µmol for the SynPhase PS A-Series Lanterns) or "SynPhase™ Crown" (e.g., 8.3 µmol for the I-Series Crown) to fit into the tip of each pin.

Tea bag and nanokan technique

In 1985, Houghten (15) introduced the "tea bag" method for simultaneous multiple peptide synthesis. In this method, peptide synthesis occurs on resin beads that are sealed inside labeled, porous polypropylene packets. At the end of the synthesis, the individual peptides are liberated from the resin. In 1995, this method was improved by using minibaskets (e.g., Nanokan) to encapsulate radio-frequency microchips (RF tagging) and a sample of resin beads for peptide synthesis (16, 17). Each of these minibaskets can be scanned with an electronic reader before or right after each coupling cycle during the "split-mix" synthesis (see Fig. 1a) (2, 3, 18). Therefore, the synthetic history of each minibasket can be traced. At the end of the synthesis, compounds can be cleaved off the resin beads inside each minibasket and placed in a 96-well plate to form a spatially addressable compound library. The main advantage of this method is that it offers considerable synthetic flexibility as one may use any resin bead in the synthesis. A few thousand to ten thousand individual peptides can be generated easily. Furthermore, peptides can be produced in sufficient quantity (500 µmol) for purification and complete characterization if desired. The IRORI Nanokan® system is available commercially from Discovery Partners International (San Diego, CA). However, the IRORI system is very expensive. A less expensive encoding system, called En-core (Torviq Inc., Granger, IN), uses "necklace" color tags and multicolor reaction vessels to encode the synthetic history. This system also uses the "split-mix" synthesis strategy to generate compound libraries with a limited number of compounds.

SPOT synthesis technique

A peptide library can be synthesized using the SPOT synthesis technique to form a low-density peptide spot array (e.g., 25 spots/cm²). In this method, different peptides are synthesized in situ as low-density arrays on cellulose membrane or paper (B). The volume of 20–25 nmol of Boc-amino acids and coupling reagents dispensed creates a specific SPOT size that determines both the scale of reaction and the absolute number of peptides that can be arranged on an area of a membrane. Cotton (another form of cellulose) and polysyntene-grafted polyethylene film segments also have been used as solid supports. Recently, polymeric membranes that are chemically, mechanically, and thermally more stable have been developed, which include hydroxy-functionalized PEG acrylate polypropylene membranes and an amino-functionalized ester-free PEG...
Peptide Combinatorial Libraries

Resin beads

1) Mix

2) Split

PET

H₂N

NH₂

P

E

T

H₂N

NH₂

Building blocks

20 natural amino acids

Reaction cycles

3

4

5

6

Library entities

2²⁰ = 8,000

2²⁰ = 160,000

2²⁰ = 3,200,000

2²⁰ = 64,000,000

(a)

Figure 1 The "split-mix synthesis" method to generate a one-bead one-compound combinatorial library (a) and a number of permutations for random peptide libraries (b). P, E, and T are building blocks (in this case amino acids).

(b)

Peptide microarrays

Peptide microarrays are prepared by immobilizing many peptide molecules on the surface of a solid support in a small area in an addressable fashion. The immobilization can be achieved via in situ synthesis or chemical ligation through a covalent bond. A hydrophilic linker between the solid surface and the peptide usually is added to minimize steric hinderance caused by the solid support. The most commonly used solid support for microarray printing is a standard microscope glass slide. Other solid supports also have been used such as polystyrene, nitrocellulose membranes, PVDF membranes, Hybond ECL membranes, gold surfaces, and chemical vapor deposited diamond films.

Peptide arrays on paper or cellulose membranes generated by SPOT synthesis generally are low density, even with the commercially available automatic SPOT synthesizer. Such a low-density peptide chip (several thousand peptides) now is available commercially, for example, PepChip™ microarray from Mimotopes. Foder et al. (4) first described the high-density peptide microarray using the photolithographic light-directed parallel synthesis method. The disadvantage of this method is that it requires building blocks (in this case, amino acids) protected with photolabile protecting groups (e.g., 6-nitroveratryloxycarbonyl) that are not available yet commercially. To address this problem, Gao et al. (19, 20) developed digital photochemistry for parallel synthesis of peptides to prepare a peptide microarray. In this method, they combined light-directed synthesis (controlled by computer) with microfluidics so that photo-generated acid (e.g., H+·SbF6−) can be generated in situ (with light) to remove the protecting groups of standard commercially available building blocks such as Boc-protected amino acids. More recently, the same group has developed a novel photogenerated base (PGB) applicable to Fmoc-chemistry for the parallel synthesis of a peptide on a microarray. When fully optimized, this versatile in situ, high-throughput parallel synthesis of peptide microarrays would offer unprecedented opportunities for creating various high-throughput detecting and sensing devices, which would enable a broad range of biochemical and biomedical applications (21).

We reported site-specific ligation of peptides to a carrier protein or polymer before printing onto the glass slide with an automatic arrayer to form a peptide microarray (22). Recently, Dikmans et al. (23) reported a novel process for manufacturing
multifunctional high-density peptide microarrays termed “Spotting compound-support conjugates” (SSC). In this method, a trifluoroacetic acid (TFA) cocktail containing > 80% TFA plus scavengers is used to solubilize the peptide-cellulose of each SPOT. The peptide-support conjugates then are precipitated, redissolved in dimethyl sulfoxide (DMSO), and then printed on glass slides.

Synthetic Library Methods that Require Deconvolution

The term “deconvolution” has been used to describe the process whereby the active molecule or molecules in a library are identified, usually by the iterative testing of mixtures of compounds for a specific biologic property. Using the results from biologic assays, the identity of the active component may be deduced without the need to determine directly or indirectly its chemical structure. The two main deconvolution approaches are the iterative process (24) and the positional scanning (25). In the iterative approach first reported by Geysen et al. (24) in the multipin system, a progressive selection is performed by choosing one amino acid at a time for each position. Sublibraries are generated based on the result of the previous one; therefore, the sequence is obtained step by step. In the positional scanning method reported by Pinilla et al., sublibraries of peptides with an amino acid fixed at one designated position but randomized in other positions are prepared and tested for biologic activities. Based on the biologic assay results, the amino acid sequence of the active peptide can be deduced at the end of the process. This method assumes that the contribution of each amino acid residue to the biologic activity is independent of each other. It works very well only if one predominant motif exists for the target protein. An interpretation of results from targets with multiple binding motifs likely will be difficult.

OBOC Library Method

In 1991, we first reported the OBOC concept to synthesize peptide libraries using a “split-mix synthesis” method (Fig. 1a and 2). The peptide is synthesized on resin beads such as a 90 µm diameter TentaGel resin (Rapp Polymere, Tubingen, Germany). Because each bead is in contact with only one amino acid at a time during each coupling cycle and the reaction is driven to completion, each bead expresses a single peptide entity (Fig. 1d) and carries about 100 pmol peptide. Because the peptide beads in an OBOC combinatorial library are spatially separate, an OBOC library can be considered as a huge chemical microarray that is not addressable.

The main advantages of the OBOC method include: 1) A large number (e.g., 109 – 1010) of peptides can be synthesized (Fig. 1a) and screened rapidly and simultaneously by using either one or a combination of both on-bead and solution-phase assays, and 2) multiple peptide ligands with completely different motifs often can be identified in a single screening. This occurrence is the opposite of the (convergent) iterative approach (see above) in which multistep synthesis and screening result in the emergence of only one, but not necessarily the best, solution motif.

One major disadvantage of the OBOC method using the on-bead screening method is that each library compound is tethered to the solid support via a linker such as polyethylene glycol and may result in steric hindrance between the cellular receptor and the library compound. However, in some instances the linker may be beneficial, for example, the linker can be used as a convenient handle to link the cancer-targeting ligand to the therapeutic payload.

Self-Assembled PNA-Encoded Chemical Microarrays

In this method, peptides or small molecules are prepared by the “split-mix synthesis” method and cleaved from the resin to form an encoded solution-phase library such that each library compound is tethered to a PNA code via a hydrophilic linker (11). The library then is mixed with the target protein and later exposed to planar oligonucleotide microarrays of predetermined sequences. Alternatively, the encoded soluble library can be hybridized to the oligonucleotide microarrays before incubation with the target protein.

Synthetic Library Method that Requires Chromatography Selection

In this method, the peptide library is an equimolar mixture of random peptides in solution phase from which ligands can be isolated with affinity chromatography (12). It usually is synthesized on solid support with a “split-mix synthesis” method. The peptides then are cleaved from the resins, and the solution-phase peptide library is loaded onto an affinity column with an immobilized receptor. After thorough washing, the bound peptides are eluted and microsequenced. Major concerns about this method include nonspecific binding and uninterpretable results if more than one predominant motif is present in the mixture. Additionally, enough purified peptide must be retrieved for accurate microsequencing or mass spectrometry analysis. In general, because of the high background from nonspecific binding, the affinity selection method, with rare exceptions, can be applied only to a relatively small peptide library (e.g., <10,000 peptides).

Peptide Library Screening

Peptide library screening can be divided into two categories: solid-phase screening, in which peptides still are tethered to the solid-phase support, and solution-phase screening in which the peptides are released from the solid support and tested in solution. Solid-phase screening is generally easier with higher
throughout than in solution-phase assay. However, the linker that tethers the peptide to the solid support may interfere with the interaction between the peptide and its receptor. In addition, the multivalent binding between the immobilized synthetic and phage-display peptide gives little information about the real binding affinity between the receptor and an univalent soluble peptide. Different assays such as ELISA, cell-based cytotoxicity assay, antimicrobial assay, affinity chromatography, and radiometric and fluorescence-based assays can be used to screen peptide libraries. Surface plasmon resonance spectroscopy (e.g., BLACore) is a powerful technique to screen soluble and unlabeled peptides, which allows the real-time measurement of soluble peptide binding to different targets of interest immobilized on a chip such as proteins, sugars, fatty acids, nucleic acids, cell membranes, and even whole cells. However, commercially available equipment (e.g., BLACore) that uses label-free detection equipment is generally low throughput. In the last few years, several reports have applied label-free optical detection methods to screen peptide or chemical microarrays (26, 27).

To screen large libraries optimally, carefully selected high-throughput assays should be employed. The choice of the assay system largely depends on the combinatorial library method used to construct the library, the availability of reagents such as enzymes, antibodies, and radiolabeled ligands, and of course the biologic target itself. To isolate the few active peptides from the library successfully, it is essential to have a robust and accurate screening method that can screen large libraries rapidly. One may use several assay systems serially first to narrow down the potential ligands and then retest them for biologic effect and crossreaction with other unwanted targets.

**Phage-Display Library Screening**

The phage-display peptide library can be screened easily by panning (e.g., binding to an acceptor-coated petri dish). The phages that bind specifically to an acceptor molecule will be isolated and enriched via several cycles of panning and amplification in E. coli (E. coli). By the end of the panning, the bound phagemids are eluted, cloned, and processed for subsequent identification via DNA sequencing. The phage-display peptide library also has been screened with intact cells in cell culture to identify ligands that bind cell surface receptors. Peptides with a propensity to bind and then enter intact cells can be selected by eluting the phagemids off the cell surface before lysing the cell for phage recovery. Ligands identified through in vitro protein or cell panning need to be validated and evaluated for in vivo targeting. Phage-display peptide libraries also have been screened in vivo in animal models and humans. For example, different libraries have been screened for binding to the vasculature of various organs and tumors by injecting the phage library intravenously into live animals. Organ- or tumor-bound phage then are amplified in E. coli and screened a second and third time before sequence analysis. In a recent study by Krag et al. (28), patients with late-stage melanoma, breast, and pancreatic cancer were infused with random phage-display peptide libraries or phage-display short chain Fv antibodies in either single or multiple panning experiments. For more information about using the phage-display peptide library approach to identify cancer-targeting ligands, please see our recent review (29).

In some cases, functional assays have been used to screen phage-display peptide libraries. For example, a phage-display hexapeptide library was constructed with an epitope tag distal (N-terminal) to screen for peptide substrates of a specific protease. All the phages were captured by an immobilized anti-epitope antibody. After incubation with a tissue plasminogen activator (tPA), phages that expressed a peptide substrate for tPA were released for subsequent rounds of selection. A similar approach was applied to discover peptide substrates for HIV-1 protease.

**Spatially Addressable Parallel Library Screening**

The spatially addressable parallel libraries can be screened using either a solid- or a solution-phase assay depending on the method used to generate the library. In solid-phase assays, the peptide still is linked covalently to the solid support, such as a pin, SPOT-membrane, or glass surface. The molecular target can be a purified protein, a protein mixture, an enzyme, an intact cell, or a whole organism such as bacteria or virus. For example, enzyme-linked immunosorbant assay (ELISA) assay has been used for a multipin or spot-synthesis peptide library. Binding assays can be used to determine binding specificities, such as for the identification of cell-specific surface markers. One unique feature of chemical microarrays for cell adhesion analysis is that in addition to obtaining binding profiles of different cells to many ligands, one also can examine the signaling response of each binding event. Such signaling profiles can be determined with many fluorescent-labeled antibodies in conjunction with confocal microscopy. Functional properties also could be examined by using other assays to measure specific biologic effects that the ligands have on the target. In the solution-phase assay, the peptides are cleaved from the solid support, which enables the soluble ligands to interact freely with the target molecule or cell. The peptides are put into 96-, 384-, or 1536-well plates for high-throughput biologic analysis with the aid of robotics. Many conventional assays can be applied to the screening including cell-based cytotoxic assays such as MTT and XTT assays, antimicrobial assays, competitive ELISA assays, radioactive assays, radioactive binding assays, fluorescent polarization assays, time-resolved fluorescence assays, fluorescent protein-based recombinant cell bioassays, scintillation proximity assays, and cell-based calcium flux assays.

**OBOC Peptide Library Screening**

OBOC peptide library can be screened on bead or in solution phase if the bound peptides are released via a cleavable linker. The on-bead assays include binding and functional assays. In the binding assays, the target of interest could be a...
purified protein, a protein complex, cell lysates, live cells, or
a whole microorganism. In the case of protein screening, pro-
tein first is incubated with a library of immobilized ligands.
Protein-ligand interaction can be visualized with an appropri-
ate reporting group such as a biotin, an enzyme, a fluorescent
probe, a color dye, or a nucleic acid. The biotinylated pro-
tein can be detected with a colorimetric assay by using a
streptavidin-alkaline phosphatase conjugate followed by color
development with a 5-bromo-4-chloro-3-indoxyl phosphate. Flu-
orescence microscopy has been used successfully to screen bead
libraries with fluorescent probes. An organic fluorescent dye or
quantum dots have been used as a fluorophor to label the target
protein. Radio-labeled probe is useful but more tedious than the
colorimetric method.

The use of intact live cells or whole microorganisms to screen
OBOC peptide libraries is very attractive because purified target
proteins are not needed in such a screening. For microorganisms
such as a virus or a bacteria, a reporter antibody or a specific
dye that stains such microorganisms may be useful. For intact
cells, reporter probes are not needed because cells that bind to
a bead can be visualized easily under a dissecting microscope.
The positive beads can be washed off the cells and sent for
sequencing. To enhance the quality of the results more, the posi-
tive beads can be retested against either the same cells or
different cells lines before being sequenced to ensure that they
are true positives. Alternatively, one may label one cell type
with calcein and leave the negative control cell line unlabeled.

Beads that bind only to fluorescent green cells but not to
colorless cells can be considered true positive beads. After
sequencing, they can be retested against fluorescent green cells
for other assays such as the functional assay described below.

Latipoo et al. (30) recently reported a new method called the
"Bead blot" for identifying ligand-protein and protein-protein
interactions using OBOC combinatorial peptide libraries. In this
method, a peptide library was synthesized on chromatography resin
beads is incubated with a starting material that contains a target
protein for which a ligand is sought. Then the protein-loaded
beads are immobilized in a porous matrix, and the proteins are eluted
directly from the beads and captured on a membrane
that is superimposed on the beads. The location of the target
protein on the membrane is determined by probing the bead
blot with various antibodies, and the beads that originally bound
the protein are identified and sequenced. The advantages of the
"Bead blot" include the ability to select ligands of unpurified
protein, including trace proteins present in complex materials,
and ligands of multiple proteins under a variety of conditions
in a single experiment.

Aggarwal et al. (31) synthesized a random one-bead one-
dimer peptide library on a polyethylene glycol acrylamide
(PEGA) resin by modifying the one-bead one-compound method
and screened the library with a prostate cancer cell line LNCaP.
One peptide (OMAAPKLRELH), was found to bind as a dimer
to LNCaP cells that had been spiked into the blood, but it did
not bind to normal hemapoietic cells.

In addition to the binding assay, functional assays have been
developed for the screening of OBOC libraries to identify
specific substrates for protein kinases and proteases. Highly
porous PEGA resin is used for the peptide library construction
because it allows the enzyme to gain access to the bead interior.

To identify peptide substrates for protein kinases, the bead
library is incubated with [γ-32P]ATP and the protein kinase.
The phosphorylated peptide bead library is washed thoroughly
and immobilized on a glass plate with agar. The 32P-labeled
beads then are detected by autoradiography, and individual
radio-labeled ("active") beads are isolated for microsequencing.

For protease substrate screening, OBOC peptide libraries with
fluorescent dye incorporated at the C-terminus and a quencher
at the N-terminus have been used. Peptide-beads susceptible
to protease cleavage will fluoresce. These fluorescent-labeled
beads were isolated for microsequencing under a fluorescent
microscope or isolated by a fluorescent-activated bead sorter
(e.g., COVA™ BIOBEAD, Union Biometrica, Inc, Somerville,
MA). Information about the substrate sequence, the cleavage
point, and the degree of cleavage can be obtained in a single
screening. Redal (32) extended the OBOC concept to gen-
erate one-bead two-compound (OBTC) peptide libraries. For
example, an OBTC library can be generated on a PEGA resin
that contains both a library of inhibitors of proteolytic enzymes
and a fluorescence-quenched substrate that competes with the
inhibitor for binding to the active site of the enzyme.

Thus far, only a few groups have reported on the release of
compounds from OBOC libraries for solution-phase assays. An
elegant and powerful approach to screen the OBOC libraries is
the in situ solution-phase releasable assay, in which the
compound-bead libraries are immobilized in a thin layer of agar
(33). Compounds from each bead then are released to the vicin-
ity of each bead for a solution-phase assay in the semi-solid.

This method works particularly well for identifying antimi-
icrobial and anticancer agents because a zone of growth inhibition
around the positive beads can be detected easily. Jayawickreme
et al. (34) reported a "cell-based lawn format" that uses an in
situ photolysis/cleavage method to release the compound. These in
situ releasable solution-phase assays have great potential but
will require more development before they can be used reliably
for drug screening. For example, special solid supports need to
be developed so that all compounds will diffuse freely out of all
beads into the surrounding media. We recently have developed
novel bilayer shell core beads for such purpose (unpublished
work). These beads consist of a polystyrene core in which the
coding tags reside and a hydrogel shell on which releasable
library compounds are attached.

An alternative approach to using solution-phase assays to
screen an OBOC library is to release compounds from an
individual or small collection of compound beads in a microtiter
plate. The released compounds then are subjected to standard
solution-phase assays or are used to print multiple replicates of
chemical microarrays. To have enough material from one single
bead for biologic assays, one may use macrobeads (250-500 µm
diameter) or bead aggregates that are prepared by cross-linking
the TentaGel resin beads with glutaraldehyde.

Peptide Combinatorial Libraries

Synthetic Libraries that Require Deconvolution

Many existing solution assays can be applied to the peptide libraries made by this approach. As mentioned above, multistep synthesis and screening that identify only one, not necessarily the best, solution motif are performed.

Self-Assembled PNA-Encoded Chemical Microarrays

In the original report, it was used for the screening of protein binding. Like the OBOC method, a whole-cell binding assay also can be applied to the encoded planar chemical microarrays, although this has not been reported.

Synthetic Libraries that use Affinity Chromatography Selection

For the affinity column selection library, a solution-phase peptide library (a mixture) is loaded onto a column with an immobilized target protein. This method has been applied successfully to retrieve an antibody-specific binding peptide from a mixture of peptides. It also was reported to identify peptide motifs for SH2 domains and kinase domains of protein tyrosine kinases.

Elucidation of the Chemical Structure of Active Peptides

A n important goal in screening combinatorial libraries is the identification of novel structures that interact with receptor targets of biologic interest. Depending on the methods used to generate the library, the procedures for the structure determination of active peptides identified from screening vary.

Phage-display peptide library

Standard DNA sequencing is used to elucidate the structure of an active peptide. From the DNA sequence, one then can determine the amino acid sequence of the displayed peptide.

Spatially addressable library

The structure of each peptide in the library is known already; therefore, structure determination of the positive lead is not needed.

OBOC combinatorial library

In peptide libraries composed of α-amino acids (including many unnatural α-amino acids), automatic microsequencing by Edman degradation is the method of choice. However, the Edman sequencing method is expensive, is relatively slow, and requires that the N-terminus of the peptide be free; plus, the peptide consists of α-amino acids only. For peptides without a free N-terminus (e.g., cyclic peptides that use the N-terminal amino group for cyclization), branched peptides, and peptides with one or more nonsequenceable building blocks (e.g., β and γ-amino acids), the OBOC library synthesis requires different synthetic and chemical encoding strategies. Based on the chemistry used, decoding can be achieved by either Edman microsequencing, mass spectrometry (MS), or gas chromatography. To eliminate the interference of coding tags with the screening, the coding tags should be in the bead interior (Fig. 2a). We recently have developed two simple, yet highly robust, methods to prepare topologically segregated bilayer beads (Fig. 2b) (35, 36). We recently reported a new "ladder-synthesis" approach that combines both "ladder-synthesis" and bilayer bead concepts to encode OBOC nonsequenceable peptide and pep- tidomimetic libraries (36). An add-in benefit of the new "ladder-synthesis" is that the library beads generated by this method are amenable to both MS and Edman microsequencing if the compound is a sequenceable peptide (Fig. 2c). More recently, Joo et al. (37) used a similar approach for a high-throughput sequence determination of OBOC cyclic peptide library members using partial Edman degradation/MS.

Synthetic library methods that require deconvolution

Structure determination by physico-chemical methods for deconvolution libraries is not necessary because the chemical

![Figure 2](attachment:figure2.png)

Figure 2. (a) Spatial separation of a testing peptide and coding tag. A, B, and C are sequenceable or nonsequenceable amino acids; X, Y, and Z are sequenceable coding units; and X encodes A, Y encodes B, and Z encodes C. (b) Photomicrograph of topographically segregated bilayer beads. Free amines at the inner core of each bead (which is the site for a coding tag) were reacted with bromophenol blue to show blue color. (c) MS decoding of an OBOC library generated by the new "ladder-synthesis" approach.
Peptide Combinatorial Libraries

structure of the active compound can be deduced from the synthetic history and analysis results of the compound mixtures.

Self-assembled PNA-encoded chemical microarrays

The identity of the positive library compound that interacts with the target protein can be determined by knowing the nucleic acid sequences of the oligonucleotide microarrays.

Affinity column selection

For the synthetic library method that uses the affinity chromatography selection approach, the bound peptides can be eluted and microsequenced by Edman degradation. Concurrent microsequencing of the retrieved peptide mixture can be performed rather than sequencing individual peptides. Sequence motifs then can be defined in a fast and efficient way. However, the amino acid sequence obtained will be the result of the summation of the peptide mixture. Unless a predominant, distinct motif and an alignment of one or more of the critical residues exists within the peptide sequence of the library, with a fixed residue at a specific position, the result could be very difficult if not impossible to interpret.

Practical Applications of Combinatorial Peptide Libraries

Combinatorial library technology can be employed to study almost any biologic target, and results from these studies can increase our fundamental understanding of cellular function and signaling pathways. It also can be used for the discovery and optimization of drug leads. Random peptide libraries displayed on phages have been used successfully for a variety of biologic applications. These applications include the discovery of peptide ligands to target receptors, specific ligands for DNA sequences, enzyme inhibitors, peptides that mimic carbohydrate structures, protein-protein interfaces, and receptor binding sites. Others include peptide ligands for cancer cells and cancer-associated proteins. In vivo screening of phage-displayed peptide libraries has resulted in the discovery of peptide ligands that bind to the endothelium of tumor blood vessel and lymphatic vessels of specific organs, tissues, or cancers.

Over the last decade, the synthetic combinatorial library approach has been applied successfully to various biologic systems. These systems include the identification of ligands against antibodies (both continuous and discontinuous epitopes), streptavidin, avirin, opioid receptors, melanin-diminulating hormone (MSH) receptors, signal transduction adapter molecules such as SH2 and SH3 domains, adhesion molecules such as gplllbIlla, specific metal ions, double-stranded DNA, and organic dye molecules. In addition, peptides with antiprotease (including HIV protease), anti-inflammatory, antibacterial, and catalytic activities have been discovered. Furthermore, substrate motifs for various protein kinases, proteases, deacetylases, and other posttranslational modification enzymes have been elucidated. Peptide-based enzyme mimics (artificial enzymes) have been identified through screening combinatorial peptide libraries. We reported the use of a $^{32}$P or $^{33}$P phosphorylation assay and an autoradiographic method to identify specific and efficient peptide substrates for protein kinases. We also have described the use of whole-cell binding assays in which bead libraries are mixed with live cells to identify cell surface-binding peptide ligands specific for prostate cancer, nonsmall cell lung cancer, and lymphoma cells. Table 2 summarizes some published applications of combinatorial library methods.

Combinatorial chemistry also has been applied successfully to the field of material science (38, 39). This application includes the discovery of polymeric structures with specific physical, chemical, electrochemical, photochemical, or photoelectric properties.

Future Directions

Combinatorial chemistry has been playing a central role in the field of chemical biology. Huge libraries of peptides, peptidomimetics, small molecules, natural product-like molecules, and macrocycles can be generated on solid support or in solution phase. Many methods have been developed to screen such libraries. Numerous reports detail the successful application of these technologies to basic research and to the development of pharmacetics, diagnostics, imaging agents, new materials, catalysts, and biosensors. Biologic libraries such as phage-display peptide libraries, although very powerful and useful, are limited primarily to the standard 20 L-amino acids. With existing technology, the incorporation of a few selected unnatural amino acids into such libraries is, in principle, feasible but far from a routine (40). A need exists to develop highly robust biologic combinatorial chemistry systems that enable one to incorporate many unnatural amino acids, including some that can be modified subsequently through site-specific ligation or derivatization. Peptide and chemical microarrays (planar or bead-based), with or without microfluidics, have been evolving rapidly in the last decade. Many new technologies to generate and screen such microarrays have been developed. One major challenge is to develop a robust label-free optical detector that can measure the real-time binding kinetics of a pure target protein or complex protein mixtures against high-density microarrays (e.g., 10,000-20,000 discrete peptide or chemical spots per three square inches, a size of a microscope glass slide). A further challenge is to develop an efficient in situ light-directed synthesis of chemical molecules other than peptides and nucleic acids in a high-density microarray format.

The OBOC combinatorial library method is highly versatile and economical. It also is a form of chemical microarrays. Many investigators successfully have applied the on-bead screening methods in their research. The solution-phase and cell-based assays for OBOC libraries, however, are much less developed and have been applied successfully in only a few laboratories. A need exists to develop robust methods that allow investigators to screen routinely huge OBOC releasable peptide or chemical libraries (e.g., 200,000 compounds) with multiparametric...
## Peptide Combinatorial Libraries

### Table 2: Combinatorial library methods and their applications

<table>
<thead>
<tr>
<th>Peptide library methods</th>
<th>Applications (examples)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biologic library methods</strong></td>
<td></td>
</tr>
<tr>
<td>Phage-display</td>
<td>Cell surface ligands (integrins, adhesion molecules or surface receptors, vasculature, tumor lymphatics, normal organs/cells/body fluids); B-cell epitope mapping (gp120, neoptactetetraosyleramide); MHC class II molecule binding peptides; peptide substrate for IRA; peptide substrates for HIV-1 protease; and SH2 and SH3 domain-binding peptides.</td>
</tr>
<tr>
<td>Plasmid-display</td>
<td></td>
</tr>
<tr>
<td>Polysome-display</td>
<td></td>
</tr>
<tr>
<td>Fimbria-display</td>
<td></td>
</tr>
<tr>
<td>Yeast-display</td>
<td></td>
</tr>
<tr>
<td>Bacterial-display</td>
<td></td>
</tr>
<tr>
<td>mRNA-display</td>
<td></td>
</tr>
<tr>
<td>Ribosome-display</td>
<td></td>
</tr>
<tr>
<td>In vitro selection methods</td>
<td></td>
</tr>
<tr>
<td><strong>Synthetic library methods</strong></td>
<td>B-Cell epitope mapping (foot-and-mouth disease, TMV antigen, shrimp allergen); T-Cell epitope mapping (type 1 diabetes, alloreactive T-cell); protein kinase substrate/inhibitor (protein kinase I); MHC class I molecule binding peptides; and epitope mapping for monoclonal and polyclonal antibodies.</td>
</tr>
<tr>
<td>Spatially addressable parallel library</td>
<td></td>
</tr>
<tr>
<td>- Multipin technology</td>
<td></td>
</tr>
<tr>
<td>- SPOT synthesis</td>
<td></td>
</tr>
<tr>
<td>- Nanokan technology</td>
<td></td>
</tr>
<tr>
<td>- Peptide microarray in situ synthesis spotting of finished peptides</td>
<td></td>
</tr>
<tr>
<td>- Multisyringe system</td>
<td></td>
</tr>
<tr>
<td>- 96-well plate library method</td>
<td></td>
</tr>
<tr>
<td><strong>Synthetic library methods that require deconvolution</strong></td>
<td>T-Cell epitope mapping (multiple sclerosis, human papillomavirus); B-Cell epitope mapping (various antigens); opioid receptor agonists and antagonists; MHC class I and II molecules binding peptides; protease inhibitors; DNA-binding peptides; mimotope/epitopes for monoclonal or polyclonal antibodies; Tab-2-binding peptides; antibacterial agents; CAMP-dependent protein kinase motif; and molecules with catalytic and hemolytic activities.</td>
</tr>
<tr>
<td>- Iterative approach</td>
<td></td>
</tr>
<tr>
<td>- Positional scanning</td>
<td></td>
</tr>
<tr>
<td>- Recursive deconvolution approach</td>
<td></td>
</tr>
<tr>
<td>- Orthogonal partition approach</td>
<td></td>
</tr>
<tr>
<td>- Dual recursive deconvolution</td>
<td></td>
</tr>
<tr>
<td><strong>One-bead one-compound library method</strong></td>
<td>B-Cell epitope mapping (insulin); T-Cell epitope mapping (partially cleaved OBOC for diabetes/Tb reactive T-cells); imotopes/epitopes for monoclonal antibodies; Cell surface ligand identification (surface idiotype, integrins); MHC class I binding motif; Protein kinase substrate/inhibitor (CAMP-Dependent protein kinase, c-src, c-abl, Etk, metalloproteinase); Protein-binding ligands (SH2 and SH3 domains, avidin, streptavidin, Bombesin-receptor, GpII bIIIa, anti-β-endorphin antibody, artificial receptors); Factor Xa inhibitor; Protease substrates and inhibitors; Antibacterial agent; Peptides with catalytic activity</td>
</tr>
<tr>
<td><strong>Self-assembled PNA-encoded chemical microarrays</strong></td>
<td>Protease substrates (serine protease thrombin, cysteine protease caspase-3)</td>
</tr>
<tr>
<td><strong>Synthetic library method that requires affinity chromatography selection</strong></td>
<td>Protein kinase substrate/inhibitor (Cak, 3bp2, Fps Fes, Grb-2, Hpc, Shc, Syk, Vav, Zap-70); Epitope mapping for monoclonal antibodies; SH2 domain-binding peptides; and kinase domains of protein tyrosine kinases</td>
</tr>
</tbody>
</table>

© 2008, John Wiley & Sons, Inc.
solution-phase cell-based assays for compounds that affect specific cell signaling pathways or target proteins in one single 10 cm Petri dish. In the areas of catalysis and material sciences, one would expect that new materials would continue to be developed combinatorially. We will not be surprised if one day novel chlorophyll mimics that consist of peptides and organic chromophores and are developed through combinatorial chemistry become a major component of future photocells. The future of combinatorial chemistry is bright, and it will continue to serve as an indispensable research tool for many investigators to solve difficult problems across many different scientific disciplines.

Acknowledgments

The authors would like to thank Dr. Joseph Kappel for editorial assistance. Supported by NIH grants R33 CA89706, NCCDOG U54CA113296, RO1CA-098116, RO1CA15483, and NSF CHE-0320122.

References


Further Reading

Prepare peptide microarrays using digital chemistry: http://gaolab.chem.uh.edu/res1.htm

See Also
Chemical Ligation: Peptide Synthesis
Peptide Synthesis
Peptides, Chemistry of
Phage Display
Synthetic Peptides and Proteins to Elucidate Biological Function
Peptidomimetics
Claudio Toniolo and Fernando Formaggio, Institute of Biomolecular Chemistry, Padova Unit, CNR, Department of Chemistry, University of Padova, Padova, Italy

doi: 10.1002/9780470048672.wecb439

This review article provides an introduction to folded and extended conformational motifs of synthetic peptidomimetics as templates for chemical biology applications. Several variants of these 3-dimensional structures have been assessed, which comprise versatile scaffolds for presentation of side-chain and main-chain peptide groups for molecular recognition. Specifically, relevant parameters and stereochemical consequences are discussed for helices of peptidomimetics based on: 1) \( \alpha \)-amino acids alkylated (e.g., methylated or ethylated) at the \( \text{C}^\text{\lambda} \)-atom; 2) \( \omega \)-didehydroamino acids, in particular \( \omega \)-didehydrophenylalanine, with a carbon–carbon double bond between the \( \alpha \)- and \( \beta \)-positions; 3) N-alkylated \( \alpha \)-amino acids, whereby the side chains typical of coded residues are transferred from the \( \text{C}^\text{\lambda} \)-atom; 4) \( \beta \)-amino acids, the subclass of residues investigated most extensively in which the amino and carboxyl functionalities are separated by more than one carbon atom; and 5) short-range, backbone-to-backbone cyclizations that generate small-ring, monolactam or dilactam building blocks.

Introduction

Peptidomimetic is a molecule that bears identifiable resemblance to a peptide that, as a ligand of a biologic receptor, can imitate or inhibit the effect of a natural peptide. Numerous peptidomimetics have been developed over the past 25 years by bioorganic and medicinal chemists. Some molecules (i.e., those with modifications in the backbone amide group) also have been termed pseudopeptides or peptide bond surrogates (1). These synthetic analogs of peptides have a variety of applications, but most expanded interest in this area focuses on their potential for developing metabolically stabilized and/or more potent and selective bioactive compounds. The complex synthetic aspects (2) and relevant biological data (1, 3, 4) of peptidomimetics have been summarized extensively in review articles and in book chapters. In this introductory review, we focus on the 3-dimensional (3-D) structural properties of peptidomimetics characterized heavily by a few, well-studied amino acids with modifications in the main chain and/or in the side chain. An effort was made through appropriate citations to expose the nonexpert and the beginning graduate students to the remarkable advances made in this area of research. The field of peptidomimetic chemistry is entering new decades of exploding activity. This time is exciting as new techniques and methodologies have now made it possible to answer questions that were not possible to address even few years ago. Many developments ushered in the new era of peptidomimetic chemistry, which include advancements in solution and in solid-phase peptide synthesis, organic synthesis, nuclear magnetic resonance and related biophysical techniques, and mass spectrometry. These advancements take advantage of an increasing understanding of protein folding. Numerous examples now exist of low molecular weight peptides and peptidomimetics that adopt folded or extended structures, and in some cases they serve as exceptional substitutes for their much larger protein counterparts in life processes. Chemists in both academia and industry are making important strides to push the frontiers in these areas. With the ability to incorporate unnatural amino acids that have strong conformational or nucleating bias, it will be possible to force much smaller peptides and peptidomimetics to adopt appropriate 3-D structures and functions. The future for the young scientists who ultimately want to pursue this line of research is indeed very promising. Rather than looking at peptidomimetics as a mature field, it should be argued that we are only in our infancy in terms of sophistication. Every reason exists to believe that a combination of approaches, which include biophysical and functional characterizations, will propel this area forward in the next years. It is especially important for those who are considering future research areas to analyze seriously the potential of such studies.

Among the various types of known ordered secondary structures adopted by peptides formed exclusively by \( \omega \)-amino acid building blocks, in this review article we shall discuss the classic...
Peptides Based on Cα-Alkylated α-Amino Acids

Peptides rich in the uncoded Cα-methylated α-amino acids (1) are biased strongly to fold tightly into the intramolecularly H-bonded, α-helical, or the strictly related 310-helical conformation (5–9), the latter generated by sequences of helical p-turns (10,11). The same 3-D structural propensities are shared by the α-helix coiled coil (14–16) (Fig. 1a), which has an integer number of amino acids per turn (n = 3.3) and consequently its smallest repeat (i.e., the shortest main-chain length that brings two side chains exactly one on top of the other) is a heptad (7 residues); and 2) in the 310-helix (Fig. 1b), which has an integer number of amino acids per turn (n = 3.0), a triplet of residues selected carefully will produce the expected amphiphilicity.

The amphiphilic helical structure of most antibacterial peptides is a prerequisite for their propensity to form channels across the double-layered biological membranes [12]. In aqueous solution, positions a and d of the α-helical heptad repeat (a, b, c, d, e, f, g) (13) require hydrophobic residues for the onset of the widespread antiparallel dimer (or multimer) superstructure (α-helix coiled coil) (14–16) (Fig. 2). The hydrophilic positions e and g, immediately on the back, reinforce the dimer stability via ionic interactions. Deep knowledge of the factors operative in helix dimer formation may help our understanding greatly of protein-protein interactions in chemical biology.

The preferred conformation of peptides based heavily on Cα-ethylated α-amino acids (3) is different. In the resulting 2.05-helical conformation, intramolecular H-bonding takes place between the N-H and C=O groups of the same amino acid building block (17, 18). To achieve the pentagonal disposition, the H-bonding parameters are distorted remarkably and the sp2 N-C=C-C= bond angle is compressed severely (by ∼6.5°). An example of this fully extended (all-trans) conformation, based on Cα-Cα-diethylglycine (3, R = CH₂CH₃) (17), is illustrated in Fig. 3. This twofold structure is extremely rare in globular proteins and is known only for a Gly-rich sequence (18).

Only few Cα-alkylated α-amino acids occur naturally (in peptide antibiotics) (20, 21), which include Aib (α-aminoisobutyric acid or Cα-amino-α-methyl glycine) (1, R = CH₂CH₃), Iva (isovaline or Cα-ethyl, Cα-methylglycine) (1, R = CH₂CH₃), and (αε)Val (Cα-ethyl norvaline) (3, R = CH₂CH₃). (αε)Val

Table 1 Relevant average parameters for peptide helices based on L-configured α-amino acids discussed in this review article

<table>
<thead>
<tr>
<th>Parameter</th>
<th>α-Helix</th>
<th>31-1 Helix</th>
<th>2.05 Helix</th>
<th>Poly(Pro)ₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ψ (°)</td>
<td>−63</td>
<td>−57</td>
<td>180</td>
<td>−70</td>
</tr>
<tr>
<td>ϕ (°)</td>
<td>−42</td>
<td>−30</td>
<td>180</td>
<td>360</td>
</tr>
<tr>
<td>θ (°)</td>
<td>3.63</td>
<td>3.24</td>
<td>2.00</td>
<td>−3.30</td>
</tr>
<tr>
<td>d (Å)</td>
<td>1.56</td>
<td>1.94</td>
<td>3.70</td>
<td>2.22</td>
</tr>
<tr>
<td>p (Å)</td>
<td>5.67</td>
<td>6.29</td>
<td>7.40</td>
<td>7.33</td>
</tr>
</tbody>
</table>

1 In the poly(Pro)ₙ helix, all α-Cα-Cα- (N₁-O₁-C₁₁) torsion angles are in the unusual cis conformation (0°).
2 The (Cα–Cα)=C–C– bond angle.
3 The (N=C–C=N)=C–C bond angle.
4 Number of amino acids per helical turn (positive values refer to right-handed helices, whereas negative values refer to left-handed helices).
5 A axial translation (per residue).
6 Pitch or axial translation per helical turn.

Peptides Based on Cα,β-Didehydro α-Amino Acids

Cα,β-Didehydro α-amino acids (∆AA-α) are found in a variety of naturally occurring microbial and fungal metabolites, in a limited number of globular proteins, and in polycyclic peptide antibiotics (22). Bioactive peptide molecules that contain these residues are less prone to enzymatic degradation.

The double bond between the sp2 Cα and Cβ atoms of ∆AA-α induces higher lipophilicity and geometric alterations in the bond distances and angles and restricts the conformational

Image
Peptidomimetics

Figure 1 Right-handed (a) α-helical and (b) 310-helical peptide models from L-configurated, Cα-methylated α-amino acids viewed along the helix axis, highlighting their amphiphilic properties. In these models, the largest black (oxygen) and white (nitrogen) atoms on the right side are those of the hydrophilic side chains of Ser, Asp, Glu, and Lys.

Figure 2 Antiparallel dimer formation generated by an amphiphilic α-helix.

Figure 3 The X-ray diffraction structure of the 2.05-helical homo-pentapeptide from Cα,α-diethylglycine with five, consecutive C=O···H–N intramolecular H-bonds (17).

The flexibilities of the peptide backbone and the side chain as compared with those of protein residues. Most 3-D structural studies in this subclass of peptidomimetics have exploited ΔPhe mainly because of its convenient chemical synthesis and interesting conformational properties. ΔPhe exists in two diastereomeric forms: the Z-configurational isomer (ΔZPhe, 4), where the N–H group is in the cis disposition with respect to the benzyl moiety, and the E-isomer (ΔEPhe, 5), where it is in the trans disposition. Almost all conformational studies described so far have been performed on the Z-isomer (22–24) essentially because a large part of the synthetic routes results in this stereoisomer.

From the investigations on ΔPhe-based peptides, it is clear that the 3-D disposition largely preferred by this residue is the helical conformation (23, 24). Therefore, most corner positions of the various types of β-turns and all positions of 310/α-helices are accessible easily to this residue. Fig. (4) shows the right-handed 310-helix generated by a double repetition of the model triplet -ΔZPhe-ΔZPhe-L-Ala (a). Aromatic amino acid residues, which include ΔZPhe (25), are known to...
Peptidomimetics

(a)

(b)

Figure 4 Right-handed 310-helical conformation formed by a double repetition of the \(\Delta\text{ZPhe-}\Delta\text{ZPhe-L-Ala}\) model peptide: views orthogonal to (a) and down (b) the helix axis.

Another useful property of \(\Delta\text{Phe}\) is its strong absorption in the near-UV region (\(\sim 280\) nm). This characteristic has been exploited in several studies to probe peptide conformational transitions and helix screw sense changes of chiral switches (26-28).

Interestingly, a series of homo-peptides derived from the \(\text{C}^{\alpha,\beta}\)-didehydro \(\alpha\)-amino acid with the shortest side chain (\(\text{C}^{\alpha,\beta}\)-didehydroalanine or \(\Delta\text{Ala,}6\) was shown unambiguously to overwhelmingly adopt a multiple, consecutive, fully extended conformation (29). These peptide molecules are essentially flat, including the amino acid side chains, and form completely planar sheets (Fig. 5). This uncommon peptide 3-D structure is stabilized by two types of intramolecular H-bonds, \(\text{N-H} \cdots \text{O} = \text{C}\) (typical of the 2.05-helix) and \(\text{C}^{\beta}{i+1} \cdots \text{H} \cdots \text{O} = \text{C}\) (characteristic of \(\Delta\text{Ala}\) peptides). The molecules are isolated and pack in layers without any significant contribution from intermolecular \(\text{N-H} \cdots \text{O}\) H-bonds. For this exceptionally flat ("sole-like") peptide structure, it is reasonable to foresee a bright future as a potentially active bridge in charge transfer systems.

Peptides Based on N-Substituted \(\alpha\)-Amino Acids (Peptoids)

In 1992, Simon et al. (30) introduced a new concept in the search for potentially biactive peptide molecules: The side chain of the \(\alpha\)-amino acid \(\alpha\)-carbon is shifted by one atom along the backbone to the next nitrogen to generate a peptoid (i.e., an N-substituted oligoglycine). Peptoids have been shown to be protease resistant (only tertiary amide groups are present in the main chain), oral bioavailable, and easy to synthesize in oligomeric forms with complex sequences (31). Peptoids have been also exploited for the construction of many biologically active compounds in many different fields of pharmaceutical research.

Because of this principal structural divergence from coded peptides, peptoids lack amide protons. This property precludes the formation of the intramolecular H-bonds that contribute largely to the stabilization of the most common helical structures of \(\alpha\)-peptides. Unlike that of \(\alpha\)-peptides, the peptoid backbone is inherently achiral (as it is based on Gly residues). However, it has been reported that sufficient bias to form stable helices of a specific screw sense can be provided by side chains with an \(\alpha\)-chiral carbon atom (that is linked directly to nitrogen) (32, 33). The extraordinary resistance of these helices to loss of their ordered secondary structure is another intriguing property of peptoid molecules.

The only published X-ray diffraction structure for any peptoid oligomer, the \(\text{N-rich}\) homo-pentamer, shows clearly that the molecule is folded in a left-handed helical conformation with...
Peptides Based on β-Amino Acids

ω-Amino acids (that have more than one carbon atom intervening between the amino and carboxyl functions) possess a larger degree of conformational variability than α-amino acids. Initial conformational investigations of oligomers of ω-amino acids were driven by the extensive interest in the various forms of nylon. More recently, a large body of 3-D structural investigations on β-amino acid oligomers have been discovered, but the conformational preferences of peptides based on ω- and l-amino acids and of “hybrid” peptides (e.g., those formed by ω- and β- or ω- and γ-amino acids) are also the subject of numerous spectroscopic and X-ray diffraction analyses (37-39). In this section, we summarize the most relevant conformational results extracted from the studies of peptides rich in the β-amino acids listed below.

In p-Ala [30], the simplest amino acid of this class, the central, characteristic C2=C3 bond is unconstrained and can adopt values for the related n (also called µ in some publications) torsion angle of approximately ±60° (gauche) and 180° (trans). Detailed analyses of the rotational flexibility about this bond in linear p-Ala peptides revealed that it is mostly trans oriented (40, 41). Also, these studies provided almost exclusive evidence for the formation of extended peptide chains, which usually self-aggregate into sheet-like structures.

More recently, introduction of (one or multiple) substituents and chiralities on the C2 and C3 atoms (32-32, 15, 16) and incorporation of C2=C3 (14, N=C2 (17), and N=C3 (18) cyclized motifs seemed to facilitate folding of β-peptides into a new type of helices (42-47). A variety of these ordered structures has been authenticated. If the nomenclature is used that terms peptide helices according to the number of atoms involved in the pseudocyclic structure generated by the intramolecular C=O···H-N H-bond, those found most frequently are the 14-, 12-, and 12/10-helices. Interestingly, although the direction of the H-bond for the 12-helix is the same as that typical of α-peptides (i.e., from a C-terminal N-H group to an N-terminal C=O group; in this specific case i → i + 1), the direction is reversed (i → i + 1 and i → i + 2, respectively) for the 10- and 14-helical, β-amino acids in the extended conformation may also produce parallel or antiparallel pleated-sheet structures.

The 14-helix emerged as the best documented structure among the β-peptide folded structures. The ω, ω torsion angles are extended (±130° - 140°). The n value is constrained (e.g., by the cyclohexyl ring of the cyclic β13AA (34) (n = 2), close to ±60°. Both the pitch and the diameter of this termary helix are ±5.0A.

When the substituents on the C2 and C3 atoms of the β-amino acid are part of a five-membered ring (as in the trans cyclic β13AA, 34, with n = 1), the accessible n torsion angle is restricted to about ±95°. The resulting 12-helix has a set of ω, ω torsion angles near ±90°. The helix repeats approximately every 2.5 residues, with a pitch of ±5.5A, and a diameter slightly less than 4.0A.

The 12/10-helix was found in alternating β12 (33) and β11 (12) monosubstituted residues. The dipeptide unit that contains a CONH-bond with no adjacent R-NH2 substituents generates the 12-membered ring, whereas that with the two adjacent substituents is part of the 10-membered ring. This ordered structure consists of consecutive narrower 10-membered and wider 12-membered H-bonded rings. Consequently, the two successive H-bonded rings exhibit opposite directionality. In this conformationally mixed helix, the peptide planes of the 10-atom rings are approximately perpendicular to the helix axis, whereas the peptide planes of the 12-atom rings are aligned nearly with the helix axis. As in the 14-helix, the n value is close to ±60°. The number of amino acids per turn is 2.7. The
overall helix macrodipole is much smaller compared with those of the other helical structures. The $\text{C}=\text{O}$ (and the $\text{N}$–$\text{H}$) bonds show an alternating up/down direction with respect to the helix axis.

Two (parallel and perpendicular to the helix axis) views for each of the three β-peptide helices discussed in this section are presented in Fig. 7. Recent applications of β-peptide helices to relevant biological issues (antibacterial activity, transport across cell membranes) (48–51) will certainly stimulate additional work in this promising area.

Lactam-Based Peptides

Several review articles and books have dealt extensively with the synthetic problems encountered and have discussed in detail the biological data of lactam-based peptides. However, insufficient attention has been paid to the conformational implications induced in the resulting constrained analogs (52–56). Here, we will mention briefly only those lactam-based conformational restrictions in peptides that are achieved by short-range cyclizations that involve the backbone $\text{N}$ and $\text{C}^\alpha$ atoms (from residue $i$ to residue $i$ or $i+1$). Therefore, conformational restrictions that involve the $\text{C}^\gamma$ (carbonyl) atom and those obtained by medium- and long-range cyclizations, although of great theoretic and practical interest, have not been considered. In addition, in all peptides discussed in this section, the repeating $\text{N}$–$\text{C}^\alpha$–$\text{C}^\gamma$ sequence is maintained.

As a part of a program that evaluates N-acylated γ-lactams as conformationally constrained building blocks of pseudopeptide foldamers, homo-oligomers of L-pyroglutamic acid (L-pGlu) (19) to the tetramer level were synthesized (57). The preferred conformation of this pseudopeptide series with $\text{N}_i \leftrightarrow \text{C}_{i+1}$ ring restriction in structure-supporting solvents was assessed by various spectroscopic techniques. In addition, the crystal-state structure of the $\text{N}^\alpha$-protected dimer was established by X-ray diffraction. A high-level DFT computational modeling was also performed based on the crystallographic parameters. In this analysis, it was demonstrated that an $\alpha\text{C}^\gamma$-$\text{H} \cdots \text{O}$=C intramolecular H-bond is responsible for the stabilization of the s-trans-L-pGlu-L-pGlu conformation by 1.4 kcal mol$^{-1}$. This effect can be detected easily by $^1\text{H}$ NMR because of the anomalous chemical shifts of the $\alpha\text{CH}$ protons present in all oligomers. In summary, a new polyimide-based, foldameric structure (Fig. 8) was developed that, if functionalized appropriately, holds promise as a rigid scaffold for novel functions and applications.
Figure 7: Two orthogonal representations for each of the β-peptide 14-, 12-, and 12/10-helices (a, b, and c, respectively) formed by a β^2 AA or/and a β^3 AA, where R^2/R^3 are methyl groups. The H-atoms, except the -CONH- H-atoms, are omitted for clarity.
Peptidomimetics

Figure 8
Computer model of the ternary helical structure of the pseudopeptide \(- (L-pGlu)_{4} - \) segment generated by the repeating \( \psi_{1}, \omega_{1}, \) and \( \psi_{2} \) backbone torsion angles obtained from the X-ray diffraction structure of Boc-(L-pGlu)_{2}-OH (Boc, tert-butyloxycarbonyl) (57).

Figure 9
Computer models of: (a) the cyclic correlate based on 12 monomeric units of \( (2S, 4R)-4\text{-amino-5-oxopyrrolidine-2-carboxylic acid} \). The idealized backbone \( \phi, \psi \) torsion angles are \( 120^\circ, 120^\circ \), respectively. (b) The nanotubular structure resulting from vertical self-assembling of the cyclic correlate through intermolecular N-H···O = C=H-bonds that involve the trans amides (62). The black and white atoms are oxygens and hydrogens, respectively.

Energy calculations for \( (S)-3\text{-amino-2-pyrrolidone (\( \gamma \)-lactam) and (S)-3-amino-2-piperidone (\( \delta \)-lactam) derivatives indicate that the most stable conformations have \( \psi \) values restricted to the range \(-125 \pm 10^\circ\). As a consequence, these chiral mono-lactams can be accommodated readily at the left corner of a type II' \( \beta \)-turn \( (\psi = -60^\circ, \phi = -120^\circ) \) (10, 11), a position occupied typically by \( \alpha \)-amino acids of \( \alpha \) (D) chirality, but obviously not at the same position of a type-II \( \beta \)-turn \( (\phi = -60^\circ, \psi = 120^\circ) \). X-ray diffraction, \( ^{1}H \) NMR, CD, and computer graphics analyses on a variety of analogs of bioactive peptides are in excellent agreement with the theoretic findings, in particular that underlie the strict requirement of reversing the chirality of the residue at the left corner of the \( \beta \)-turn to keep the same type (II or II') of chain folding when a 5(or 6)-membered ring lactam replacement is involved (59). Despite the gross conformational similarities between the 5- and 6-membered ring lactams, distinct biological properties usually are noted for peptides that contain such rings. More subtle differences in conformation and/or in bulkiness between the two lactams might be responsible for the observed diverging biological activities.

The effect of the incorporation of a succinimide \( (22) \) moiety into a peptide chain (another type of \( C_{\alpha} \leftrightarrow C_{\alpha}^{+1} \) ring restriction) has been examined carefully (60, 61). The five-membered succinimide annular system is an intermediate in the repair of selective degradations of side-chain deamidated proteins. It has been shown both theoretically and experimentally that L-Asu-Xxx- (Asu, \( \alpha \)-aminosuccinyl) dipeptide sequences strongly favor the type-II' \( \beta \)-turn, a behavior similar to that of the \( \gamma \)-lactam modification discussed above.

The conformational preference of the terminally blocked homo-trimer from \( (2S, 4R)-4\text{-amino-5-oxopyrrolidine-2-carboxylic acid (22a), characterized by C}_{2}^{n} \leftrightarrow C_{2}^{n+1} \) ring restriction} and an unique, alternating cis-trans secondaryamide sequence was analyzed by X-ray diffraction (62). Using computer modeling, it was also shown that the rigid 3D structure of the
trimer can be exploited as a template to construct novel linear oligopeptide foldamers and large ring cyclic correlates with self-recognizing properties (for the latter, see Fig. 9).

The most stable conformation of peptides that contain 3-amino-2-piperidone-6-carboxylic acid (22b, also an amino acid with a C^α ↔ C^α+1 ring restriction, has been determined by ^1H NMR (63). A chair conformation is a feature that characterizes the piperidone ring. Also, an intramolecularly H-bonded β-turn form is the most plausible explanation for the experimental finding on these peptides.

The synthesis of a novel macrocyclic (10-membered ring) dilactam (23, n = 2), an additional template with a C^α ↔ C^α+1 ring restriction that may promote formation of type-II β-turns, has been reported (64). By analysis of the ^1H NMR spectrum, it was established that a chair-like conformation is preferred for this ring structure and that the two amide groups are trans with roughly parallel planes. It seems, however, that the presence of the two macrocyclic secondary amide functions perpendicular to the plane of the turn would facilitate intermolecular association that leads to poor solubility and gel formation. A similar (11-membered) ring structure has been synthesized to generate a relatively rigid model of the bioactive peptide thymopentin. Peptides that contain two consecutive macrocyclic dilactams, separated by a C^α-C′-N-C^α region that retains flexibility about the ψ_i+2 and ω_i+2 torsion angles, have been prepared (65). This swivel-like effect can allow the peptides to adopt low-energy folded conformations. From potential energy calculations, it was established that the best candidates for β-turn models are the 12-membered ring structures that have the configurational sequence L,L,D,D (in these compounds all amide groups are nearly trans planar). The 10- and 11-membered dilactams were shown to be higher in energy. The results of a CO-H NMR study are consistent with the presence of a β-turn in the -L-Lys-L-Glu-D-Lys-D-Glu-bis-dilactam-bridged tetrapeptide.

In search for enkephalinamides with improved analgesic activity, conformationally restricted analogs, where the Tyr, Gly, Phe, and Leu/Met residues are cyclized to form a piperazin-2-one (24, an n = 2, → N_i+1 ring restriction), have been prepared. Analogs with either D-Tyr¹ or L-Leu⁵(L-Met⁵) are active. The absolute configuration of the piperazin-2-one derived from the intermediate dipeptide amide H-L-Phe-L-Leu-NH₂ has been assessed by X-ray diffraction. The enantioface-differentiating abilities of cyclic peptides that contain N,N'-ethylene-bridged dipeptides have been examined. The X-ray diffraction structure of an L-Ala-L-Ala piperazin-2-one derivative has been described (66). The ψ_i torsion angle is fixed at about −35°, whereas ϕ_i and ω_i, in part external to the ring structure, are trans.
Summary

In this introductory review we have presented the most relevant 3-D structural characteristics of a selected set of peptidomimetics that, hopefully, will be of help to the nonspecialized scientists for potential use in the currently expanding field of chemical biology. More specifically, we have discussed examples of peptide surrogates with main-chain and side-chain modifications either at the α-carbons, or at the β-carbons, or with different types of cyclizations (N=N, N=C, and C=C). For journal issues or books devoted to these and other types of peptidomimetics, the reader is referred to References 67–73.

References


Further Reading
Peptidomimetics


See Also

Amino Acids, Chemistry of
Natural and Unnatural Amino Acids, Synthesis of
Peptide Synthesis
Peptides, Chemistry of
Synthetic Peptides to Define Structure-Function Relationships
Amide bonds are found ubiquitously in natural or synthetic molecules of biologic interest. Since the early days of synthetic organic chemistry, methods for the formation of amides have been described. More recently, with the development of solid-phase chemistry and automated peptide synthesis, new strategies and reagents have been devised to overcome typical problems such as low conversion and racemization. This article provides an overview of the methodology that is available today. Depending on the nature of the synthetic target and the associated synthetic challenges, different approaches can be envisaged. Methods range from the rather straightforward use of acyl halides, anhydrides, and carbodiimides, to the more elaborate, low-racemization inducing methods that use phosphonium/uronium-based reagents. New amide bond-mediated ligation methodologies now offer new convergent strategies for the synthesis of highly functionalized molecules of biologic interest.

Amide bonds are found commonly in small or complex synthetic or natural molecules of biologic interest. Amide bonds are, for example, the structural backbone of proteins that play a crucial role in almost every biologic process. In nature, amides are formed via complex enzymatic pathways that ensure selectivity and specificity of the formed molecule. An analysis of the Comprehensive Medicinal Chemistry database revealed that more than 25% of known drugs bear amide functionality (1). For example, Taxol (paclitaxel; Bristol-Myers Squibb, New York) is a highly functionalized diterpenoid, which is extracted originally from the bark of the pacific yew tree. It is used to treat ovarian and breast cancers (2). Fuzeon (enfuvirtide; Roche Pharmaceuticals, Nutley, NJ) is a synthetic biomimetic peptide and is the first of a novel class of fusion inhibitor antiretroviral drugs used to treat HIV-1 infection (3). Lipitor (atorvastatin hemicalcium salt; Pfizer, Inc., New York) is the best-selling drug in the world, and it is used to treat high cholesterol (4). Sprycel (dasatinib; Bristol-Myers Squibb) is an oral dual BCR/ABL and Src family tyrosine kinase inhibitor, which is approved for use in patients with chronic myelogenous leukemia (5) (See Fig. 1).

Since the early 1840s (6), amide bond formation has been a very active field of research in organic chemistry. A multitude of synthetic methods have been elaborated and optimized (7). Relevant examples of these methods available to the bioorganic, organic, medicinal, or combinatorial chemist are reported in this article.

Amide Bond Formation: Overview

Amide bond formation by direct condensation between an acid and an amine is not obvious and must overcome adverse thermodynamics (8). This dehydrative process can be achieved under forcing conditions such as high temperatures (160–180 °C), which are usually incompatible with the presence of other functionalities. Contrary to the ester formation between an acid and an alcohol, which is an equilibrium, the acid and the amine undergo first an acido–basic reaction that yields the stable salt. Therefore, the acid must be activated by the attachment of a leaving group before being reacted with the amine (see Fig. 2). These two steps can be carried out separately with intermediate isolation of the activated species or by a one-pot synthesis with late introduction of the amine. More recent conditions allow coupling to occur with both the acid and the amine present in the reaction mixture. Often, the choice of the methodology for one specific amide formation is not only governed by the yield. Avoiding racemization at neighboring chiral centers, improving difficult isolation...
from reaction by-products or reducing the costs of the reagents, might be the key elements for decision.

Amide Bond Formations: Methods and Strategies

Acyl halides

Acyl or acid chlorides are used frequently in amide formation as activated forms of the corresponding carboxylic acid. A wide selection of acyl chlorides is available commercially. Otherwise, they can be prepared readily from the corresponding carboxylic acid in the presence of reagents such as thionyl chloride (9), oxalyl chloride (10), phosphorus trichloride (11), and phosphorus pentachloride (12). Reactions that use oxalyl chloride or thionyl chloride are promoted by the addition of a catalytic amount of DMF (13) (see Fig. 3).

Despite the high reactivity and low cost, these chlorinating reagents generate hydrogen chloride in situ and are not suitable for amide formation in the presence of acid labile groups. Hence, alternative basic conditions have been studied. For example, cyanuric chloride is used commonly to generate acyl chloride in nonacidic conditions (14). The presence of an organic or inorganic base maintains the basic pH conditions throughout the reaction. The cyanuric chloride method is also used in large-scale synthesis (15).

Neutral conditions (i.e., non-acid generating) that use triphenylphosphine and a source of chloride as carbon tetrachloride (16), hexachloroacetone (17), trichloroacetimide (18), or trichloroisocyanuric acid (19) have been developed. Alternatively, 1-chloro-N,N,2-trimethylpropenylamine (20) converts acids into acyl chlorides readily without HCl formation.

The coupling reaction with the amine usually requires an additional acid scavenger (often a base like triethylamine, DIEA, or NMM) to trap the formed HCl. The reaction can also be accelerated in the presence of a catalytic amount of DMAP (21), pyridine, or metallic zinc (22).

Nevertheless, acyl chlorides have some limitations. Acyl chlorides that bear α-hydrogens can undergo ketene formation under basic conditions (23). The subsequent amine addition occurs with potential loss of chirality and side reactions (24). Oxazolone-mediated racemization is encountered in peptide chemistry. N-protected peptidoyl chlorides yield the corresponding oxazolones spontaneously. These transient species react readily with nucleophiles; but, under the standard aminolysis conditions, the α-proton is acidic enough to enable acido–basic equilibrium, which compromises the chirality of the α-center. Peptides are, therefore, grown usually at their N-terminus, thus avoiding the oxazoline formation. Furthermore, the acyl chloride activation of N-urethane-protected amino acid (e.g., Boc, Fmoc, or Cbz) is unadvisable, as they react with the carbonyl of the neighboring urethane to yield the corresponding NCA (25). NCA functionalities are notoriously reactive and promote many side reactions (see Fig. 4). Similar problems can also be observed with other activation methods.

Alternatively, acid fluorides are used to activate the acid. Acyl fluorides are less sensitive to moisture and are more reactive toward primary and secondary amines than the corresponding acyl chloride. Furthermore, they are compatible with basic (Fmoc and Cbz) or even acid- (Boc) labile amine protecting groups and less prone to promote racemization than their chlorinated homologs (26). Thus, the acid fluoride method is often used in peptide synthesis (27). Cyanuric fluoride (28), TFFH (29), DAST (30), and deoxofluor (31) are used commonly as fluorinating reagents (see Fig. 4).
Acyl chloride formation using oxalyl chloride and DMF as catalyst:

\[
\text{Cl}_2\text{C}=\text{Cl} + \text{H}_2\text{N} \rightarrow \text{Cl}_2\text{CO} + \text{HCl}
\]

Racemisation via ketene formation:

\[
\text{R}^-\text{CH} = \text{O} \rightarrow \text{R}^-\text{C} = \text{O} \quad \text{addition}
\]

Racemisation via oxazolone formation:

\[
\text{R}^-\text{CH} = \text{O} \rightarrow \text{R}^-\text{N} = \text{C} = \text{O} \quad \text{amination}
\]

NCA mediated side reactions:

\[
\text{RCOOH} + \text{Nu} \rightarrow \text{R}^-\text{N} = \text{C} = \text{O} + \text{CO}_2
\]

Figure 4  Acyl chloride–mediated amide coupling and potential side reactions.

Acyl azides

The acyl azide strategy was developed for peptide synthesis in the early 1900s. The original preparation of the acyl azide from the corresponding methyl ester is a two-step synthesis. Displacement of the methyl ester with hydrazine, which generates the acyl hydrazide, is followed by the nitrosation reaction to yield the acyl azide. Isocyanate formation, also called the Curtius rearrangement, is a possible side reaction, and ureas are often observed as side products. An improvement of this method is the one-pot synthesis of the acyl azide from the carboxylic acid using DPPA. Acyl azides are, however, potential explosives and the leaving group (free azide) is toxic, which provides some limitation to this method.

Acyl imidazoles

CDI allows one-pot amide bond formation and is also used for large-scale peptide chemistry. Initially, the mechanism may involve the formation of acyl carboxy imidazole and imidazole. Both intermediates react together to lead to the activated species as the acyl imidazole. Then the amine is added to undergo aminolysis. As imidazole is generated in situ, the reaction does not need an additional base and it is usually compatible with amine HCl salts. Incidentally, the acyl imidazole intermediate can also be isolated and stored. Some simple acyl imidazoles are even available commercially. Similarly, carbonylimidazolium salts have been introduced. For example, CBMIT is described as an efficient amino acylating reagent for peptide synthesis with sterically hindered amino acids.

Anhydrides and mixed anhydrides

Anhydrides react readily with diverse nucleophiles such as amines. For example, the use of acetic anhydride was reported in the 1850s to produce acetamides. Symmetric anhydrides can be prepared by dehydrating the corresponding acid under strong acidic conditions or at high temperatures. A more practical approach, however, consists of treating the carboxylic acid with DCC. The anhydride is then subjected to aminolysis with the desired amine. Under some specific conditions, hydroxyl carboxamides can be prepared directly from the corresponding hydroxyl acid without protection of the alcohol.
Synthetic Chemistry: Formation of the Amide Bond

Acyl fluorides

\[
\begin{align*}
\text{Acyl azides} & \quad \text{Curtius rearrangement} \\
\text{Acyl midazoles} & \quad \text{Mixed pivalic anhydride}  \\
\end{align*}
\]

Figure 4  Acyl fluorides, acyl azides, and acyl midazoles.

Figure 5  Symmetric anhydride preparation and coupling reaction.

process, which practically limits the application of this method to cheap and commercially available symmetric anhydrides. Several mixed anhydride methods have, therefore, been developed. The acid is coupled to a second acid moiety that is considered to be disposable. This strategy relies fully on the regioselectivity of the aminolysis where the reactive center a is more reactive than b (see Fig. 6). The concept of mixed anhydrides has been extended to carbonic, boric, iso-ureas, phosphoric, and phosphinic acid-derived species. Iso-ureas and phosphorous-containing reagents will be discussed in separate articles.

Mixed pivalic anhydride 18 (42) is an example of selective mixed carboxylic anhydride. Selective aminolysis could be caused by the steric hindrance of the tert-butyl group. A mixed carboxylic anhydride strategy has also been studied to overcome the aminolysis selectivity problem. This method exhibited excellent selectivity when ethoxycarbonyl 19 (43) or tert-butyloxycarbonyl 20 (44) anhydrides were used. The selectivity is mainly caused by the higher electrophilicity of one of the carboxylic centers toward the amine. Ethoxycarbonyl anhydrides can be prepared conveniently using ethyl chloroformate or EDDQ 21 (45).

Acyl aryl boron species also react with amines to yield amides with mixed results. For example, catecholborane was used to generate lactams successfully (46). Aryl boronic acids with electron-withdrawing groups, such as 3,4,5-trifluorobenzoic boronic acid and 3,5-bis-(trifluoromethyl)benzenboronic acid, can act efficiently as an amidation catalyst when added to a mixture of acid and amine (47).
Synthetic Chemistry: Formation of the Amide Bond

O-acyl iso-ureas

O-acylisourea species 22 are generated easily by reacting carbodiimides with a mixture of the desired carboxylic acid and amine. Then, they undergo aminolysis readily to form the amide and the urea by-product (48). Often, formation of unreactive N-acylureas 23 by acetyl transfer and racemization are observed (see Fig. 7). This side reaction can be minimized substantially by reacting the acid with the carbodiimides at 0° C before adding the amine or by using DMAP or HOBt 24 as adjuvants (49).

DCC 17, DIC 25, EDC 26, and polymer supported PS-CC 27 are available commercially. Elimination of the resulting ureas can be achieved easily by filtration or solvent wash depending on their solubility. In the case of PS-CC, the urea by product is formed advantageously on the solid support while the amide is released in solution.

Esters

Alkyl esters

Alkyl esters (e.g., methyl, ethyl, benzyl esters) are usually stable toward amines and, thus, are used as protecting groups. Some anecdotic examples of amide bond formation with alkyl esters, however, are reported in the literature. High reaction temperature, addition of a Lewis acid (52), or use of organoaluminium species generated from DIBAL-H-H2NR (53) can enable these reactions. Saturated ammonia in methanol can also react with

<table>
<thead>
<tr>
<th>Name</th>
<th>R&quot;=&quot;N=C=N&quot;R&quot;</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCC - 17</td>
<td>Nt(CAN)</td>
<td>Resulting DCU poorly soluble. Eliminated by filtration. Solution phase chemistry.</td>
</tr>
<tr>
<td>DIC - 25</td>
<td>Nt(CAN)</td>
<td>Resulting diisopropyl urea soluble in DCM. Eliminated by DCM washes. Solid/phase chemistry.</td>
</tr>
<tr>
<td>EDC - 26</td>
<td>Nt(CAN)</td>
<td>Resulting urea is water soluble. Eliminated by aqueous work-up Solution phase chemistry.</td>
</tr>
<tr>
<td>PS-CC - 27</td>
<td>Nt(CAN)</td>
<td>Resulting urea immobilised on solid support Solution phase chemistry using polymer supported reagents</td>
</tr>
</tbody>
</table>

Figure 6 Different types of mixed anhydrides used for amid bond formation.

Figure 7 Carbodiimides.
methyl esters at room temperature to form the primary amides (54).

Recently, amidation of methyl esters with unprotected amino alcohols in the presence of a catalytic amount of IMes, a readily available carbene, was achieved in good yields (see Fig. 8). Initially, the carbene was proposed to react with the methyl ester to generate the activated C2-acylimidazolium intermediate (55). X-ray evidence, however, suggests a more complex mechanism (56).

Figure 8  Heterocyclic carbene catalyzed amidation.

Activated esters
Usually, esters of phenols are easier to hydrolyze than alkyl esters and they also react with a wide range of nucleophiles such as amines. The reactivity is increased when electron withdrawing groups are present on the phenyl ring. Activated esters can be obtained from the acid using DCC 37 mediating coupling or via the acid chloride. They can be used immediately without additional purification or they can be stored. For example, PNP 28 esters are purified easily by recrystallization in alcoholic solvents with which they do not react. Then, amidolysis is performed at room temperature (37). Other examples of alcohols used to activate acids are represented in Fig. 9.

PFP esters have been recommended for the synthesis of heterocyclic amides, when DCC 17 mediated coupling or via the acid chloride. They can be used immediately without additional purification or they can be stored. For example, PNP 28 esters are purified easily by recrystallization in alcoholic solvents with which they do not react. Then, amidolysis is performed at room temperature (37). Other examples of alcohols used to activate acids are represented in Fig. 9.

Phosphonium-based coupling reagents
80F 38 also known as Castro’s reagent, was the first of the HOtB-based oxonium reagents (73). (see Fig. 13). The deprotonated acid reacts first with BOP 39 to produce both the highly reactive (acyloxy)phosphonium 40 species and the OBt anion. The initial mechanism postulates the direct attack of the OBt anion on the (acyloxy)phosphonium 40, which generates the aminolysable activated O-Bt ester 41. This process is driven by the formation of the stable phosphoric triamide 42 (74). More recent studies suggested the existence of an additional step. The (acyloxy)phosphonium intermediate 43 reacts with the carboxylate to form the symmetric anhydride 44. The subsequent activation with HOtB 45 is relatively slow and yields the anticipated O-Bt ester while regenerating the carboxylate. The additional step can be accelerated by the addition of HOtB 45 to the reaction mixture (75). Whether it is really necessary to use HOtB 42 in conjunction with HOtB-based oxonium or phosphonium reagents is still debatable and it must be investigated on a case-by-case basis. PyBop (G.L. Biochem (Shanghai) Ltd., Shanghai, China) and PyAOP 45 (76) where the dimethyl moieties were replaced were introduced to avoid the generation of toxic HMFA (42,77). Other examples of phosphonium-coupling reagents are shown in Fig. 32.

A new field of application for active esters is solid-phase synthesis. Some polymer-supported reagents are available commercially (see Fig. 9). The acid is first immobilized on a polymer support as an active ester and the excess reagents are washed away conveniently. Finally, the amine is released by amine treatment. During the cleavage, a limited amount of amine can be used to avoid the presence of excess amine in the final mixture. The acid is loaded onto the resin using classic ester condensation methods for TFP resin 35 (66), HOtB resin 36 (67), and oxime resin 37 (68). In the case of the triazine resin 38, the acid is loaded via an aromatic nucleophile substitution in the presence of a base (69).
N-methylated amino acids. The Bt ester is believed to be too stable to react with these hindered amines, which enables degradation or racemization to occur. Some effective reagents that eliminate the need for HOBt have therefore been developed (see Fig. 12). For example, PyBrop is an efficient peptide-coupling reagent for N-methylated amino esters. It is interesting to notice that reagents such as CloP and BroP had been considered initially as poor peptide-coupling reagents as they lead to noticeable racemization with primary amino acids.

Other organo-phosphorous reagents are based on the mixed carboxylic-phosphoric or phosphinic anhydrides. Initially used to convert carboxylic acids into acyl azides, DPPA has been introduced as a one-pot coupling reagent for peptide chemistry, and it was adapted later to solid-phase chemistry. The driving force of these reactions is the formation of the phosphoric or phosphinic acids and their salts. Later DPP-Cl and FDPP were introduced. FDPP has been used successfully in macrocyclizations. Examples of phosphoric- and phosphinic-based coupling reagents are shown in Fig. 12 [DEPBT, BDP, BOP-Cl].

Uronium/guanidium-based coupling reagents

Another popular family of reagents is based on uronium species such as HBTU and TBTU. These species are the hexafluorophosphate and tetrafluoroborate salts of the same molecule, respectively. The coupling is performed the same as the phosphonium reagents, and the nature of the non-nucleophilic counter ion has no influence on the outcome. The driving force is, in this case, the formation of the stable tetra-methylurea (see Fig. 13). In some cases, the amine can react directly with the coupling reagent to form the undesired guanidine. This side reaction can be suppressed by adding HOBt to the reaction. Usually, these reagents are found in their more stable guanidinium form (N-form) (89, 90). The uronium species of HBTU and HATU, however, have been isolated and are more active than the guanidinium species. Unfortunately, under standard coupling conditions, the O-form is converted quickly to the N-form (91) (see Fig. 14).

Some examples of uronium-based coupling reagents are represented in Fig. 15 (e.g., BCC, TDBTU, TNTU, TPTU, TSTU, HAPyU, TAPipU, CIP, BTFFH, HOTT / TOTT). As for phosphonium-based coupling reagents, poor results are observed with sterically hindered amines when HOBt is present. Therefore, some alternative reagents have been designed.
**Synthetic Chemistry: Formation of the Amide Bond**

\[
\begin{align*}
\text{BOP} & \quad \text{mediated coupling.} \\
\text{Phosphonium based coupling reagents} & \\
\text{Miscellaneous organophosphorus coupling reagents} & \\
\text{Iminium-based coupling reagents} & \\
\end{align*}
\]

**Figure 11** BOP 39-mediated coupling.

**Phosphonium based coupling reagents**

**Miscellaneous organophosphorus coupling reagents**

**Iminium-based coupling reagents**

Example, HATU 58 has proved to be very efficient in some sterically hindered couplings (64, 98). The superior reactivity of the At-activated ester toward amines is discussed in the Activated Esters section. TOTU 70 enables the formation of an activated acyl oxide intermediate, and low racemization in peptide couplings has been reported (99).

Inimium reagents are inspired directly from the uronium family. They differ structurally by the replacement of the amino groups with a hydrogen, an alkyl, or an aryl group. For example, BOMI 71 was reported to be more reactive than the correspondent uronium reagents (100). The increased reactivity could stem from...
**Synthetic Chemistry: Formation of the Amide Bond**

R-OH

N

N

O

N

NEt3

BF4

PF6

OBt

R

O

O

N

TBTU - 55

HBTU - 54

HOBtR´NH2

Figure 13

HBTU 54/TBTU 55-mediated coupling.

![Figure 14: Uronium and guanidinium forms of HBTU (X = C) 54/HATU (X = N) 58.](image)

from the reduced number of mesomeric forms observed in the iminium species compared with the uronium forms. Representative examples of iminium reagents are shown in Fig. 15

**Quaternized nitrogen-based coupling reagents**

**Triazinyl esters**

Recently, DMTMM 75 has been described as an efficient, one-pot coupling reagent for ester and amide bond formation (102). This reagent first undergoes an SNAr reaction as in the case of cyanuric fluoride 9. The activated ester then undergoes amidolysis. The in situ liberation of N-methyl morpholine avoids the use of an additional base conveniently. The triazine 76 by-product is eliminated easily by an aqueous wash (see Fig. 16).

An inexpensive substitute to DMTMM is 2-chloro-4,6-dimethoxy-1,3,5-triazine in the presence of an additional organic base (103).

**Mukaiyama’s reagent**

2-Chloro-1-methylpyridinium iodide 77, also called Mukaiyama’s reagent, in the presence of a carboxylic acid and a tertiary amine yields an activated 2-acyloxy-pyridinium species 78. This intermediate reacts with a range of nucleophiles. The driving force for this reaction is the generation of a stable tautomeric pyridone. Mukaiyama’s reagent has been reported in the preparation of activated esters, the formation of amides, and the conversion of β-amino acids to the corresponding β-lactams (104, 105). The poor solubility of pyridinium iodides in conventional solvents indicates that the reaction requires refluxing in methylene chloride. To alleviate this limitation, novel reagents have been suggested: BEMT 79 (106), BEP 80, FEP 81, BEPH 82, and FEPI 83 (107) (see Fig. 16). Tetrafluoroborate and hexachloroantimonate have been adopted as counter ions. These reagents were used successfully in the synthesis of oligopeptides and in solid-phase peptide synthesis.

**Isoxazolinium Salts**

Woodward’s reagent K or NEPIS 84 is a zwitterionic isoxazolinium that reacts with N-protected amino acids in presence of triethylamine to form the activated enol ester 85 (108) (see Fig. 16). This intermediate can be reacted without additional purification with an amine to yield the desired amide and the sulfonate by-product that can be easily removed by aqueous extraction.

**Amide bonds by chemical ligation**

New ligation strategies allow the selective formation of an amide bond between two highly functionalized fragments such as unprotected peptides, glycopeptides, or other molecules of biologic interest. The convergent assembling of complex, pre-formed peptidic sequences overcomes the inevitable contamination issues observed during extended linear peptide synthesis (109).

**Native thioligation**

This original methodology (110) requires the presence of a cysteine at the N-terminal position of fragment B and the...
Synthetic Chemistry: Formation of the Amide Bond

**PF₆**

![Chemical Structures](image)

**Figure 15** Uronium/guanidinium and iminium-based coupling reagents.

introduction of a thioester at the C-terminal position of fragment A. Regiospecific coupling of the two unprotected fragments can be achieved in a two-step process described in Fig. 17. The first step is a chemoselective trans-thioesterification between the thiol functionality of the terminal cysteine and the terminal thioester to form the thioester linked intermediate. This step is called the "capture reaction." The second step is the rapid, intramolecular acyl transfer from the thio- to the amino-position of the cysteine to yield the desired amide bond. Usually, no racemization is observed (111).

Different Boc (112) or Fmoc (113) compatible solid-phase strategies have been devised to allow the preparation of peptides bearing a C-terminal thioester. Recently the introduction of a 2-(ethyldisulfanyl)phenol ester at the C-terminal position of fragment A has been used in an elegant solution phase approach (114) (see Fig. 18). The first amino acid of fragment A is coupled to 2-(ethyldisulfanyl)phenol. The resulting phenol ester is sufficiently stable to be used as a protecting group and to allow the growing fragment A to use standard Boc strategy peptide synthesis. After final deprotection, fragment A can undergo native thioligation. First, the disulfide bond is cleaved under reductive conditions or in the presence of an excess of thiol reagent. The resulting 2-mercaptophenyl ester might be in an unfavorable, yet dynamic, equilibrium with the corresponding S-2-hydroxyphenyl thioester via intramolecular O- to S-acyl transfer, which generates in situ the appropriate setup for native thioligation.

A similar type of native chemical ligation has been extended to B fragments that contain homocysteine (115), selenocysteine (116), and histidine (117) at their N-terminal positions.

**Staudinger ligation**

The Staudinger ligation (118) is a less restrictive approach that can be applied to couple unprotected peptides. First, a C-terminal phosphinomethylthioester is introduced on Fragment A, and the N-terminal amine of fragment B is converted to the corresponding azide (88). The two fragments are then reacted together via a Staudinger reaction (see Fig. 19), which yields the iminophosphorane (89). This intermediate undergoes intramolecular S- to N-acyl transfer readily to produce the corresponding hydrolysable amidophosphonium salt (90).

In principle, this methodology is independent of the nature of amino acids present at the ligation point and can be even extended to all types of acids and amines. For example, N-glycosylated amides (119) and peptides have been prepared via stereoselective Staudinger ligation (120). Challenging, medium-sized (7- to 10-membered) lactams have been ring-closed in good yield using the Staudinger ligation sequence described in Fig. 19 (121). In this example, the phosphine reagent has been protected judiciously as a borane complex. The Staudinger ligation is then triggered by the deprotection of the phosphine under basic conditions.
Synthetic Chemistry: Formation of the Amide Bond

**DMTMM 75 mediated coupling**

**Mukayama coupling**

**NEPIS 84 mediated coupling**

**Figure 16** Quaternized nitrogen-based reagents.

**Figure 17** Native thioligation.

**Enzymatic catalysis**

Enzymes such as proteases (122), subtilisin (123), acylases, peptidases, amidases, and lipases (124) are reported to catalyze amide bond formation with, in some cases, enantiospecificity of over 99%. Despite limited enzyme-substrate compatibility, specific conditions have been developed to reverse their natural reactivity, which is in favor of the hydrolysis. For example, Kyotorphin (Tyr-Arg) (125), a potent analgesic, was produced on an industrial scale using \( \alpha \)-chymotrypsin, a peptidase isolated from bovine pancreas.

**Conclusions**

This article presents an overview of the different amide bond formation methodologies that are available to the organic and biochemist. For nearly two centuries, the methods have evolved from the original symmetric anhydrides and acyl chlorides. When Fischer started to study peptidic couplings in the early 1900s, it became obvious that poor yields and racemization would present a major challenge. During the last three decades, the design and synthesis of new coupling reagents have been an area of intense investigation. The predominance of carbodiimide and active ester procedures have been replaced gradually...
by onium salts approaches. The introduction of racemization suppressants combined with the development of solid-phase chemistry and urethane-based protecting group have allowed high-throughput chemistry and automated peptide synthesis to be a reality. Emerging technologies such as ligation allow convergent synthesis by amide coupling between two highly functionalized molecules.

**Abbreviations**

- **AOP** (1H-7-azabenzotriazol-1-yloxy)tris(dimethylamino)phosphonium hexafluorophosphate
- **At** azabenzotriazole
- **BCC** benzotriazolyloxy-bis(pyrrolidino)carbonium hexafluorophosphate
- **BDMP** 5-(1H-benzotriazol-1-yl)-3,4-dihydro-1-methyl 2H-pyrrol-1-yl hexafluorophosphate
- **BDP** benzotriazol-1-yl diethyl phosphate
- **BDP** benzotriazol-1-yl diethyl phosphate
- **BEMT** 2-bromo-3-ethyl-4-methylthiazolium hexachloroantimonate N-oxide
- **BEH** 2-bromo-1-ethylpyridinium tetafluoroborate
- **BEI** 2-bromo-1-ethylpyridinium tetafluoroborate
- **Boc** tert-butylxycarbonyl
- **BOMI** N-(1H-benzotriazol-1-ylmethylene)-N-methylmethanaminium hexachloroantimonate N-oxide
### Synthetic Chemistry: Formation of the Amide Bond

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOP</td>
<td>(benzotriazol-1-yl)tris(dimethylamino)phosphonium hexafluorophosphate or Castro's reagent</td>
</tr>
<tr>
<td>BPMP</td>
<td>1-(1H-benzotriazol-1-yl)oxy)pyridinum hexafluorophosphate or ethyl-1-hydroxy-1H-1,2,3-benzotriazin-4(3H)-one</td>
</tr>
<tr>
<td>BOP-Cl</td>
<td>bis(2-oxo-3-oxazolidinyl)phosphinic chloride</td>
</tr>
<tr>
<td>Bt</td>
<td>benzotriazole</td>
</tr>
<tr>
<td>BTFFH</td>
<td>1,1,3,3-bis(tetramethylene)fluorouronium hexafluorophosphate</td>
</tr>
<tr>
<td>CBMIT</td>
<td>N,N'-carbonyl(bis(3-methylimidazolium)triflate</td>
</tr>
<tr>
<td>Cbz</td>
<td>benzoylcarbonyl</td>
</tr>
<tr>
<td>CIP</td>
<td>2-chloro-1,3-dimethylimidazolium hexafluorophosphate</td>
</tr>
<tr>
<td>CDI</td>
<td>carbonyl dimaladiazole</td>
</tr>
<tr>
<td>CloP</td>
<td>chloro-tris-(dimethylamino)phosphonium hexafluorophosphate</td>
</tr>
<tr>
<td>DAST</td>
<td>diethylaminosulfur trifluoride</td>
</tr>
<tr>
<td>DCC</td>
<td>dicyclohexylcarbodiimide</td>
</tr>
<tr>
<td>DEPB</td>
<td>3-(dithiophosphoryl)oxy)-1,1,2,2'-benzotriazin-4(3H)-one</td>
</tr>
<tr>
<td>DIC</td>
<td>diisopropyldiamidimide</td>
</tr>
<tr>
<td>DIEA</td>
<td>N,N-diisopropylamine also known as Hünig's base</td>
</tr>
<tr>
<td>DMAP</td>
<td>N,N-dimethylaminopyridine</td>
</tr>
<tr>
<td>DMTMM</td>
<td>4-(4,6-dimethoxy-(1,3,5)triazin-2-yl)-4'-methyl-morpholinumchloride</td>
</tr>
<tr>
<td>DPPA</td>
<td>diphenylphosphoryl azide</td>
</tr>
<tr>
<td>DPP-Cl</td>
<td>diphenylphosphoric chloride</td>
</tr>
<tr>
<td>DCC</td>
<td>dicyclohexylurea</td>
</tr>
<tr>
<td>DODQ</td>
<td>2-ethoxy-1-ethoxycarbonyl-1,1'-diaryldipropionilne</td>
</tr>
<tr>
<td>FDPP</td>
<td>pentafluorophenyl diphenyl phosphate</td>
</tr>
<tr>
<td>FEP</td>
<td>2-fluoro-1-ethylpyridinium tetrafluoroborate</td>
</tr>
<tr>
<td>FEPH</td>
<td>2-fluoro-1-ethylpyridinium hexachloroantimonate</td>
</tr>
<tr>
<td>Fmoc</td>
<td>9-fluorenlymethoxy carbonyl</td>
</tr>
<tr>
<td>FOMP</td>
<td>5-(pentfluorophenyl)oxy)-3,4-di-hydroxy-1-methyl-2H-pyridium hexachloroantimonate</td>
</tr>
<tr>
<td>HATU</td>
<td>1H-benzotriazol-1-yl)-1,3,3'-tetramethylthiouronium hexafluorophosphate</td>
</tr>
<tr>
<td>HAPyU</td>
<td>1-(1H-benzotriazol-1-yl)-1,3,3'-triazolio<a href="pyridin-1-y">4,5-b</a>lylmethene(pyridin-1-yl)hexafluorophosphate N-oxide</td>
</tr>
<tr>
<td>HBTU</td>
<td>0H-benzotriazol-1-yl)-1,3,3'-tetramethylthiouronium hexafluorophosphate</td>
</tr>
</tbody>
</table>

**Chemical Abbreviations:**

- **HMPA:** Hexamethylphosphoric triamide
- **HOAT:** 1-hydroxy-7-azabenzo[c]azolone
- **HOBi:** 1-hydroxybenzotriazole
- **HOCl:** Ethyl-1-hydroxy-1H-1,2,3-benzotriazin-4(3H)-one
- **HODnbt:** 3,4-dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazin-4(3H)-one
- **HONB:** N-hydroxy-3-norbornene-2,3-dicarboximide
- **HOSu:** N-hydroxysuccinimide
- **HOTT:** 5-(1-oxido-2-pyridinyl)-1,1,3,3-tetramethylthiouronium hexafluorophosphate
- **HPP:** Pentafluorophenyl
- **NCA:** N-carboxyanhydride
- **NEPIS:** N-ethyl-phenylisoxazolium-3-sulphonate, also known as Woodward’s reagent K
- **NMMP:** N-methylmorpholine
- **O-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium tetrafluoroborate**
- **O-(3,4-dihydro-4-oxo-1,2,3-benzotriazin-3-yl)uronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(3,4-dihydro-4-oxo-1,2,3-benzotriazin-3-yl)uronium tetrafluoroborate**
- **O-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
- **O-(1H-benzotriazol-1-yl)-1,1,3,3-tetramethylthiouronium tetrafluoroborate**
References


Synthetic Chemistry: Formation of the Amide Bond


Synthetic Chemistry: Formation of the Amide Bond


Further Reading


See Also

Synthetic Peptides and Proteins to Elucidate Biological Function

Christian F. W. Becker and Roger S. Goody
Max-Planck Institut für Molekulare Physiologie, Dortmund, Germany

doi: 10.1002/9780470048672.wecb583

The enormous progress made in the use of gene technological techniques over the past several decades has been the main driving force in the accumulation of our knowledge of biology at the molecular level. This progress has at times tended to push more classic approaches, such as those stemming from synthetic chemistry, into the background, and there has even been a tendency to regard contributions from this area as being superfluous. This attitude has begun to change recently, with the emergence of the field now referred to as chemical biology, and it is now appreciated that synthetic chemistry can make a unique contribution to the outstanding problems in fundamental biological and medically oriented research. The full potential of these methods is beginning to be realized in the area of peptide and protein synthesis, and this will be the topic of this article.

Synthesis of Biological Macromolecules

One of the early dreams of synthetic chemists was to achieve the total synthesis of important complex biological molecules (1). At the level of polymeric molecules, this includes proteins, nucleic acids, and polysaccharides. In all cases, early work initially involved synthesis of small fragments of the polymeric molecules (peptides, oligonucleotides, oligosaccharides) and addressed, and partially solved, the initially formidable synthetic obstacles, in particular those concerning protection and deprotection to prevent reactions occurring at unwanted positions of the molecules involved. The seminal breakthrough that led to extension of these methods to longer polymers in reasonably short periods of time was made by Merrifield (2), who was the first to show that synthesis of polymeric biological molecules could be achieved on a solid support, thus removing or at least dramatically simplifying the need for time-consuming purification and isolation of intermediates after the addition of each monomer. Merrifield introduced this principle for peptide synthesis, but in fact polynucleotide synthesis, in particular DNA synthesis, proved to work at least as well, and in terms of reaching the long-term aim of total synthesis of biological macromolecules was the first to be accomplished successfully and in relatively routine fashion (3). This is largely due to the fact that oligonucleotide synthesis of fragments of a length of ca. 50 nucleotides is relatively facile on a solid support, and that enzymes can be used to ligate such fragments in a directed fashion to achieve the goal of total gene synthesis. Although this is not the most routinely used method for generation of complete coding regions for specific proteins, there are often situations where this is the method of choice, because it allows complete control of codon usage to optimize protein expression in the organism to be used. Gene synthesis is now offered on a commercial basis and plays a significant role in modern biological research.

Progress toward total synthesis of proteins has been slower, mainly due to the lack of easy availability of an enzymatic procedure equivalent to DNA ligation that would allow coupling of peptides of a length that can be conveniently prepared by solid-phase synthesis (depending on sequence, the largest fragments that can be produced are between 50 and 100 residues long). This situation has changed significantly over the past 10–15 years with the introduction and widespread use of methods for the ligation of protein fragments together with the combination of the methods of synthetic chemistry with techniques originating in biology.

In the following, we initially discuss the advances that have been made at the technical level, and then introduce some of the many applications that exploit the new methods for the study of biologically important processes.
Chemical Methods for the Generation of Large Polypeptides

The principles of solid-state peptide synthesis have been reviewed extensively and will not be repeated here, except to remind the reader that this usually involves attachment of a suitably protected amino acid, which will become the C-terminal residue in the finished sequence, via its carboxyl group to a polymeric support. After exposing the N-terminus of this residue, this is allowed to react with the next protected and activated amino acid to form a peptide bond between the last and the penultimate amino acids in the target sequence. Repetition of this cycle allows the stepwise construction of the desired polypeptide from the direction of the C-terminus toward the N-terminus. Removal of protecting groups and cleavage from the solid support leads to the free polypeptide.

The procedure outlined here is limited to oligopeptides and polypeptides of up to ca. 50 amino acids, and thus, it limits the availability of fully synthetic proteins, because most proteins or functional domains are at least ca. 100 residues long. A solution to this problem would be to generate polypeptides with a length of several tens of amino acids and then to couple (or ligate) them to produce significantly larger proteins. In earlier work, this principle was used in a block condensation approach using fully protected polypeptides, but this did not prove to be a viable procedure in most cases. Another approach is to connect fragments of the protein using non-peptide linkers with chemistry, which obviates the need for side-chain protection. An example of this approach is given below, and it can be put to good use in certain cases. A major breakthrough was the introduction of the method known as native chemical ligation in 1994 (4). In this procedure, a peptide or polypeptide bearing an N-terminal cysteine residue (Fig. 1) The ligation reaction involves a thioester exchange reaction followed by an S→N acyl transfer to generate a native peptide bond, a reaction that had been reported much earlier (5) but that had not been considered as a ligation method.

Chemical ligation has been used for the total synthesis of a large number of proteins in recent years, as described in several reviews (6–9), and recent examples extend the size range to the order of 200 amino acids, in this case using multiple ligation steps (10, 11). Despite this progress, an attractive approach that is being used increasingly is that of a combination of synthetic and molecular biological methods in the technique referred to as expressed protein ligation, as discussed in a later section.

First Access to HIV1-PR

HIV-1 PR is a homodimer made up of 99 amino acids (per monomer) that was made accessible for the first time in 1988 by Schneider and Kent, who synthesized this protein using solid-phase peptide synthesis (SPPS) (13). An automated, rapid, and highly efficient procedure in combination with purification by size exclusion chromatography was used to generate a partially purified HIV-1 PR (14), which then also became available later on in 1988 by recombinant expression in Escherichia coli (15). Proteins generated by these two procedures had the same enzymatic properties. After the initial synthesis of HIV-1 PR, one advantage of this methodology, namely, the possibility to incorporate unnatural amino acids during chemical synthesis, was demonstrated by replacing all cysteine residues in HIV-1 PR by α-amino-n-butyric acid. The resulting enzyme was fully active and was crystallized to obtain one of the first three-dimensional structures of HIV-1 PR that formed the basis for structure-assisted design of HIV-1 PR inhibitors (16, 17). At the same time this structure confirmed that chemically synthesized proteins can fold and crystallize identically with proteins from natural sources. Three different crystal forms of chemically synthesized HIV-1 PR with bound peptide inhibitors were subsequently published and contributed to the further development of HIV-1 protease inhibitors (18–20).

Backbone Engineering of HIV-1 PR

The flexibility of chemical protein synthesis was used to introduce changes into the protein backbone that could not be incorporated by other means. This paved the way for a general protein engineering approach and at the same time introduced the possibility of joining two fully unprotected peptide segments by a chemoselective reaction that generated an unnatural thioester bond between Gly51 and Gly52 of each HIV-1 PR subunit (Fig. 2a). The thioester linkage was generated by reacting an N-terminal HIV-1 PR peptide segment (aa 1–51) carrying a C-terminal thioester with a C-terminal segment (aa 52–99) having the N-terminal glycine replaced by bromoacetic acid and all additional cysteine residues replaced by α-amino-n-butyric acid (Fig. 2a) (21). This constitutes an early example of a chemoselective ligation reaction that provided access to a medium-sized protein by linking two smaller unprotected polypeptides (easily accessible by SPPS) without the need for elaborate protection schemes as used in fragment condensation reactions.

The resulting enzyme exhibited full activity, even though the thioester bond was placed inside a flexible β-hairpin loop (flap region) of HIV-1 PR, a region that undergoes drastic conformational changes during substrate and inhibitor binding. This

HIV-1 Protease as a Paradigm for Elucidating Biological Function by Chemical Protein Synthesis

Chemical protein synthesis and semisynthesis have been used to study the molecular basis of protein function in numerous cases. One of the very early and most impressive applications of chemical synthesis to the production of functional as well as site-specifically modified enzymes concerns the protease from human immunodeficiency virus 1 (HIV-1 PR). This enzyme cleaves the gag-pol polypeptide into functional proteins during virion budding from host cells and is essential for replication of the virus (12). Inhibitors of HIV-1 PR are an important class of anti-HIV drugs, and their development is at least partially based on the availability of structural and molecular information obtained with chemically synthesized HIV-1 PR.
Native chemical ligation (NCL) between two unprotected peptide segments. The initial trans-thioesterification reaction leads to an intermediate that undergoes an S to N-acyl shift via a five-membered cyclic transition state and generates a native amide bond at the ligation site.

is due to the positioning of the two glycine residues on the outside of the flaps, away from the substrate. However, the synthesis of another HIV-1 PR analog by Kent and Baca placed the thioester bond between Gly49 and Ile50, leading to a reduction in catalytic activity by a factor of 3000 (Fig. 2b) (22). This constituted the first experimental evidence that hydrogen bonds between the backbone of the flap region and the substrate are important for catalytic activity. However, substrate specificity and affinity were not affected. These particular hydrogen bonds are “transmitted” from the protease backbone to the substrate via an internal water molecule and are believed to contribute to the distortion of the scissile bond of the substrate (23).

The applicability of the thioester-forming chemoselective ligation approach was broadened by the fact that this chemistry can be carried out under acidic conditions in the presence of sulfhydryl groups. By taking advantage of this selectivity of the alkylation reaction, two different HIV-1 PR monomers were prepared. These monomers carried a free sulfhydryl group at their N- or C-terminus, respectively, and were, subsequent to the thioester-forming ligation step, joined together by a disulfide linkage to generate tethered dimers of two distinct HIV-1 subunits (24). This tethering of the two subunits produced one of the largest functional proteins prepared by chemical synthesis at that time and allowed the preparation of HIV-1 PR molecules with asymmetrical subunits. One example of such asymmetrical HIV-1 PR analogs was constructed with one subunit having a thioester bond between Gly51 and Gly52, which did not interfere with the biological activity of the protease, and a subunit that had a thioester bond between Gly51 and Gly52 and an additional ester bond instead of an amide bond between Gly49 and Ile50 (23). By replacing an amide with an oxygen atom in a unique position, no backbone hydrogen...
Synthetic Peptides and Proteins to Elucidate Biological Function

Figure 2. Chemical synthesis of backbone engineered HIV-1 protease. The peptide segments were synthesized by SPPS and harbored a C-terminal thioacid (HIV-1 PR, aa 1-49, shown in blue) or an N-terminal bromoacetic acid modification (HIV-1 PR 51-99, shown in red). These unique functional groups led to an unnatural thioester bond either between Gly51 and 52 (Strategy A; the ligation site is shown in yellow in the cartoon representation of the HIV-1 PR dimer) or between Gly49 and Ile50 (Strategy B; the ligation site is located at the end of the N-terminal peptide segment depicted in blue). The chemoselective ligation reaction is followed by purification steps and folding of the protein into its functional conformation. Strategy A led to a fully functional HIV-1 PR, whereas strategy B led to a severe reduction in catalytic affinity. The functional dimer of the HIV-1 PR is drawn as a cartoon with only one subunit showing the modifications introduced during synthesis. The second subunit (in green) is shown unmodified for clarity. Aspartic acid 25, site-specifically labeled with $^{13}$C for NMR spectroscopy studies, is shown in magenta in one HIV-1 PR subunit.

Site-Specific Side Chain Labeling of HIV-1 PR

The incorporation of an aspartic acid residue with a $^{13}$C atom at the side-chain carboxyl function at position 25 into aspartyl protease has made this catalytically essential group visible for nuclear magnetic resonance (NMR) spectroscopy (Fig. 2 (25)). The chemical shifts of this $^{13}$C atom were observed as a function of the pH and the presence and absence of substrate or inhibitor molecules. These titration experiments provided additional evidence for the suggested working model of aspartyl proteases and confirmed that HIV-1 PR is a member of this class of enzymes (26). The two aspartyl side-chain carboxyl groups (one from each subunit) act as general base and acid, respectively, thereby leading to the breakdown of the enzyme-substrate intermediate.

The work on HIV protease demonstrates how chemical protein synthesis allowed isotope labeling of a 22-kDa protein with atomic precision and provided further insights into the chemical basis of the proteolytic cleavage reaction. Isotope labeling with atomic precision has since then been used to reveal structural features of other either chemically synthesized or semisynthetic proteins (27-28).
A Mirror Image HIV-1 PR

A characteristic feature of many biomolecules is their chirality and the stereochemical specificity that is conferred to proteins and especially enzymes by being constructed from monomers with uniform stereochromical configuration. This fact has inspired chemists and biochemists to generate mirror images of proteins (as well as other biomolecules) to test the properties and especially enzymes by being constructed from monomers 

Although the total synthesis of a protein allows complete control over the structure, including posttranslational modifications and regard their biophysical behavior, enzymatic activity, and specificity. Currently, it is still not possible to modify ribosomal protein synthesis so that all-D-polypeptides can be produced, and this would in fact be a daunting undertak

ing. However, chemical protein synthesis and its ability to link peptides produced by solid-phase peptide synthesis via chemoselective reactions to form medium-sized proteins allows the synthesis of peptides from D-amino acids. M. Ilton et al. demonstrated this capability of chemical protein synthesis by producing a mirror image HIV-1 PR using their already described thioester-forming ligation approach (30). When compared with the L-form of this enzyme (also produced by chemical protein synthesis), both proteins exhibited full catalytic activity but inverse chirality, meaning that the D-form only cleaves L-substrates and the L-form only D-substrates. A crystal structure of the D-HIV-1 PR revealed that it was the mirror image of the L-form, and in the presence of a substrate-based D-inhibitor (D-MVT101), all major interactions between enzyme and substrate were clearly visible (31). In addition, all secondary structure elements clearly exhibited mirrored relationships such as the inverse handedness of alpha-helices and twists of anti-parallel beta-sheets (6). The synthesis of D-HIV-1 PR impressively demonstrates the basic determinants of protein structure and emphasizes the freedom and power of chemical protein synthesis. So far only a few D-proteins have been prepared, but potential applications are mirror image-based screenings where one screens a large library of L-peptides (generated by phage display) against a D-protein for high affinity binders (32). Any hits out of such a screen could be translated into D-peptides that would bind to naturally occurring L-proteins and possess highly interesting properties such as high stability against proteolytic cleavage and possibly low immunogenicity.

Semisynthetic Proteins of the Ras-Superfamily

A thorough total synthesis of a protein allows complete control over the structure, including posttranslational modifications and introduction of labels at desired sites in the sequence, it is still a major undertaking for which most laboratories whose main interest is in the biology of their target proteins are not equipped. In certain cases, for example when the site of introduction of a specific chemical modification is near the C-terminus, a combination of molecular biological and chemical methods has proved to be very powerful.

With the Ras-family of guanine nucleotide binding proteins, where the C-terminus plays a critical role in location to specific membranes, two approaches have been used to solve the problem of generating a C-terminus that is either naturally or unnaturally modified. In one of these, C-terminall peptide and the reaction product (prenylated Ypt1) are insoluble

In the example described, which uses chemistry to create an unnatural linkage between the C-terminal region of Ras and the rest of the prote

Figure 3  (a) Preparation of prenylated Ypt1 (a yeast Rab-protein) by expressed protein ligation. A C-terminal thioester of the truncated Rab protein was allowed to react with a doubly geranygeranylated tricysteine peptide, leading to transesterification and an S→N acyl shift to generate a native peptide bond. (b) Interaction of the C-terminus of semisynthetic doubly geranygeranylated Ypt1 with the lower domain of yeast GDI. GDI is shown in green as a ribbon structure, the C-terminus of YPT1 in magenta, and the geranygeranyl groups in red and blue CPK representation. Several residues of the C-terminus of YPT1 were not visible in the electron density map, so that the connection to the prenyl groups is not observed directly. One prenyl group (in red) is buried deeply into the hydrophobic core of GDI, whereas the other (in blue) is more superficially bound and shows interaction with the other prenyl group. The lipid binding site is generated by an opening movement of two α-helices.

In an aqueous environment, the ligation reaction was performed in detergent solution. Using the expressed protein ligation approach, both singly and doubly prenylated Ypt1 molecules could be produced. The complexes of these proteins with their solubilizing protein, GDI (GDP-dissociation inhibitor), could be crystallized, and their three-dimensional structures were determined (46-48). This revealed for the first time the nature of the lipid interaction with a binding site in an unexpected part of the GDI molecule (Fig. 3b). In the previously determined structure of GDI without a bound Rab molecule, this binding site was not detected, because a movement of one of the α-helices of...
the lower domain of GDI has to occur to create space for lipid binding, and this seems only to occur when the lipid residues, or possibly the whole prenylated Rab molecule binds. The position of binding was essentially the same for single or double geranyl groups, and this seems only to occur when the lipid residues, or possibly the whole prenylated Rab molecule binds. The position of binding was essentially the same for single or double geranyl groups, and the structure depicted in Fig. 3b shows that the corresponding residue in yeast GDI (I100) is not in the lipid binding site but makes an important hydrophobic interaction with a conserved hydrophobic motif in the Rab C-terminal hypervariable domain.

The same technology was used to create Rab proteins bearing a variety of fluorescent groups at the C-terminus. This approach allowed introduction of such reporter groups near to the reactive SH groups, which are the site of prenylation while leaving these groups free for the prenylation reaction, a process that results in large fluorescence signal changes in certain cases. Experiments on the prenylation of such selectively modified Rab proteins allowed insights into the molecular basis of another hereditary disease, namely x-linked choroideremia, a disease caused by underprenylation of certain Rab proteins [50]. The structural determination of the complex between GDI and the Rab protein GDIα, which is highly expressed in brain, and which thus sediments to a high degree in the lipid binding site, has provided considerable information on the mechanism of action of GDI in the recycling of Rab proteins between target and donor membranes (48, 49). It also sheds light on the molecular basis of a form of x-linked non-syndromic mental retardation, in which there is an L92P mutation in GDIα, which is highly expressed in brain, and which results in a reduced ability to extract Rab proteins from membranes. It was previously thought that this residue would be in the lipid binding site, but the structure depicted in Fig. 3b shows that the corresponding residue in yeast GDI (I100) is not in the lipid binding site but makes an important hydrophobic interaction with a conserved hydrophobic motif in the Rab C-terminal hypervariable domain.

Split-Inteins for Protein Semisynthesis in vitro and in vivo

The technique of expressed protein ligation has been exploited extensively during the last couple of years to produce semisynthetic proteins with tailor-made properties [4, 38]. Examples are described above, and the method has been reviewed in detail recently [41, 51–53]. The discovery that naturally occurring inteins, protein splicing domains that can excise themselves from a given polypeptide and join the flanking domains via a peptide bond, can be split into two pieces that possess the ability to spontaneously associate and form a functional intein has further extended the utility of intein technology [54–56]. In particular, the DnaE intein system has enabled the first semisynthesis of a GFP–FLAG fusion protein in vivo [59]. To achieve this goal, the N-terminal DnaE segment was fused to GFP and expressed in CHO cells. These cells were complemented with a chemically synthesized C-terminal part of the intein together with a FLAG tag and a protein transduction domain (PTD) for efficient uptake into the cells. The GFP–FLAG fusion protein that was generated upon successful trans-splicing was unambiguously identified by GFP- and FLAG-specific antibodies. Such a system allows the in vivo incorporation of biophysical probes, as long as the chemically synthesized part can be brought into the cells of interest. Detailed insights into the mechanism of the trans-splicing reaction of the DnaE intein were provided by crystal structures of this protein after excision and of a splicing-deficient precursor protein [60].

Further applications of the DnaE split intein include the development of a tandem trans-splicing system that is based on a combination of the DnaE split intein and the engineered, inducible VMA split intein [61]. Such a system allows the segmental labeling of proteins with specific isotopes [as demonstrated by Otomo et al. with the artificial PiP-Pul and PiP-Full split inteins (62)] and fluorophores. The DnaE split intein was also used by Camarero et al. to achieve the site-specific, oriented immobilization of proteins such as maltose binding protein (MBP) and enhanced green fluorescent protein (EGFP) onto glass surfaces [63]. A covalent bond to the glass surface was established by thiol-ene formation between a maleimide group on the surface and a thiol group bearing PEG linker that also carried four amino acids, including a cysteine residue, which could act as a nucleophile in trans-splicing reactions, and the C-terminal segment of the DnaE intein (36 aa). Upon addition of a MBP- or EGFP-N-intein fusion construct that was either produced by recombinant or cell-free expression the intein evolves associated and trans-splicing occurred, leading to the immobilization of MBP or EGFP on the surface. The associated DnaE intein halves were washed away, and the proteins remained, covalently bound via a PEG spacer, on the surface. The advantage of this approach is that no purification of the expressed proteins is necessary because only intein fusion constructs undergo the highly specific immobilization reaction. Furthermore, only low concentrations are needed to achieve efficient trans-splicing reaction (dissociation constant of the DnaE split intein halves is 43 nM, and trans-splicing occurs at a rate of ca. 7 × 10^7 s^-1) [61, 64], which constitutes an advantage over immobilization techniques that rely on chemoselective reactions and strongly depend on reaction concentrations [65–68]. Thus, this approach points to a new route to produce protein chips without the need for large amounts of purified protein.

The DnaA Intein

The DnaA intein from Synechocystis spp. consists of 429 amino acids, including a formin endonuclease domain, in its native form. The removal of 275 amino acids leads to the formation of a C-terminal segment consisting of only 36 aa and is easily accessible by chemical synthesis and therefore allows the addition of specifically modified peptides to its C-terminus that are, upon trans-splicing, transferred onto the N-terminal protein that was expressed as a fusion protein with the N-terminal intein segment. This split intein system has enabled the first semisynthesis of a GFP–FLAG fusion protein in vivo [59]. To achieve this goal, the N-terminal DnaA segment was fused to GFP and expressed in CHO cells. These cells were complemented with a chemically synthesized C-terminal part of the intein together with a FLAG tag and a protein transduction domain (PTD) for efficient uptake into the cells. The GFP–FLAG fusion protein that was generated upon successful trans-splicing was unambiguously identified by GFP- and FLAG-specific antibodies. Such a system allows the in vivo incorporation of biophysical probes, as long as the chemically synthesized part can be brought into the cells of interest. Detailed insights into the mechanism of the trans-splicing reaction of the DnaA intein were provided by crystal structures of this protein after excision and of a splicing-deficient precursor protein [60].
Synthetic Peptides and Proteins to Elucidate Biological Function

Figure 4  Mechanism of trans-protein splicing. (a) Initial association of the intein halves to form a functional intein. (b) Activation of the N-terminal splice-junction via an N-S acyl shift. (c) Formation of a branched intermediate upon transthioesterification. (d) Branch resolution and intein release by succinimide formation. Spontaneous S–N acyl rearrangement yields the processed product with a native peptide backbone.

To test whether trans-splicing also occurs in vitro, Mootz et al. expressed a fusion protein consisting of MBP and the N-terminal half of the DnaB intein (104 aa) and a fusion construct of the C-terminal half (47 aa) and a hexa-histidine tag (69). Upon mixing in stoichiometric amounts, successful trans-splicing produced the MBP-His-tag fusion protein. This constituted the first case of an artificial split intein that spontaneously assembled to form the active intein and underwent trans-splicing without the need for a denaturation-renaturation step. The only other artificial split intein that does not require such a renaturation-denaturation step was the VMA intein from Saccharomyces cerevisiae. However, the N- and C-terminal segments of this intein do not assemble spontaneously to form a functional intein. They require a dimerization domain that brings both halves in close proximity to each other, which induces trans-splicing (70–72). This renders the DnaB split intein highly interesting for protein engineering approaches, and in combination with the DnaE split intein or with an inducible split intein such as the VMA intein, it provides a valuable tool to combine three protein segments with each other by two concomitant or subsequent trans-splicing reactions. An additional advantage of the DnaB split intein is the occurrence of a serine residue as the C-terminal nucleophile for the splicing reaction instead of cysteine residues. Cysteine residues might not be desirable in some cases because they can interfere with folding or labeling of the newly generated protein. Nevertheless a cysteine can replace the serine as a nucleophile at this position as demonstrated by the fact that the DnaB intein has been used to generate protein segments with N-terminal
cysteine residues. This was achieved by expressing the desired DnaB intein as a fusion construct with the target protein in inclusion bodies and by taking advantage of the pH sensitivity of the DnaB intein to prevent premature cleavage during work up (73).

To extend the utility of the DnaB split intein, Liu et al. have tested 13 different sites to split this intein into two segments of different length (58). Until this series of experiments, all known artificial split inteins had been split at the endonuclease domain. Out of 13 tested sites, 3 gave functional split inteins that would undergo trans-splicing, including 1 that consisted of only 11N-terminal amino acids. Such a short N-terminal split intein half is accessible by chemical synthesis, and the introduction of chemically modified peptides at the N-terminus via trans-splicing was recently demonstrated. Such a system nicely complements the already established C-terminal modification approach via the DnaB split intein (74).

Prospects

The work reviewed here illustrates that, in the century since Fischer formulated his vision that the synthesis of proteins should be achievable using the methods of organic chemistry (1) this prediction has been largely fulfilled. While he could not possibly have predicted was the role that molecular biological techniques would play in combination with chemical methods, although he was realistic enough to imply that chemistry would not be the method of choice if biotechnological methods were available. Future developments in the area of synthetic and semisynthetic proteins are likely to include extension of ligation methods to amino acids other than cysteine and the increased use of strategies for generating proteins with precisely engineered properties in cells, including such approaches as conditional splicing, a technique in which a specific protein activity is generated intracellularly by exposure to a small membrane-permeable molecule (70–72).

References

26. Suguna K, Padlan EA, Smith CW, Carlson WD, Davies DR. Binding of a reduced peptide inhibitor to the aspartic protease...
Synthetic Peptides and Proteins to Elucidate Biological Function


To circumvent the constraint imposed by the 20 canonical amino acids on the study of protein structure and function, various chemical and biosynthetic methods have been developed to incorporate unnatural amino acids into proteins. Unnatural amino acids now can be genetically encoded in living cells in a manner similar to that of common amino acids, which expands site-directed mutagenesis to diverse novel amino acids. The use of unnatural amino acids grants researchers a multitude of chemical and physical properties that cannot be found in the normal genetic repertoire, which significantly improves their ability to manipulate proteins and protein-involved biologic processes. Changes have been tailored into proteins to accurately dissect the contribution of hydrogen bonding, hydrophobic packing, cation-π interaction, and entropy to protein stability, as well as to precisely examine the structural and functional role of crucial residues. Unnatural amino acids also enable the introduction of new chemical reactivities, biophysical probes, mock posttranslational modifications, photoactive groups, and numerous other functionalities for the modification and regulation of protein activities. These studies not only reveal fundamental information of protein structure and function but also explore new means for generating novel protein properties and controlling biologic events.

Conventional site-directed mutagenesis of specific amino acids currently is the preferred method for investigating various structural and functional characteristics of proteins. A serious limitation to this methodology is the constraint of using the 20 canonical amino acids fixed by the universal genetic code. This constraint lies in the limited chemical and physical properties of these amino acids, which hinder the ability to make precise alterations. For instance, modification of an amino acid such as glutamine is limited because only asparagine has similar characteristics. For amino acids such as proline, no analogous amino acid exists in the genetic repertoire, which makes it difficult to investigate the role of this amino acid in specific processes without abolishing it completely. Breaking this limitation would enable in-depth investigation of the principles underlying protein structure and function as well as the engineering of novel protein properties and cellular functions. In the past decade, great progress has been made in incorporating unnatural amino acids into proteins to harness their extensive and powerful capabilities. Here, we will introduce unnatural amino acids with a brief overview of various methods for incorporating them into proteins and we will present examples that illustrate how unnatural amino acids have impacted a wide array of research that investigates biologic systems.

Unnatural Amino Acids

Common amino acids consist of an amino group, a carboxyl group, a hydrogen atom, and a side chain all attached to the Cα in the L configuration. Analogs with altered side chains, or those that deviate from other features, are generally called unnatural amino acids (Fig. 1). The most widely used group of unnatural amino acids is the group in which the Cα side chain is charged.

*These authors contributed equally to this article.
Unnatural Amino Acids to Investigate Biologic Processes

Figure 1  Different forms of unnatural amino acids.

Variation of unnatural side chains is diverse and can range from structural analogs of canonical amino acids to those with specific chemical moieties, such as reactive functionalities and reporter groups for biophysical characterization. Modification of the amino group results in changes in the peptide backbone. For example, changing the amino group into a hydroxy or sulfhydryl group converts the endogenous amide bond between two residues into an ester and thioester link, respectively. Such changes make the resultant analog no longer an “amino” acid but an α-hydroxy acid and a thio acid. An analog with aminooxy replacing the amino group also has been incorporated into protein biosynthetically (1). The amino group also can be alkylated with different moieties to form amino acids that contain secondary amines (2). Another category of unnatural amino acids is α,α-disubstituted amino acids, whose α-hydrogen is replaced by an additional side chain. Moving the amino group away from the α-carbon leads to extended β- or γ-amino acids, which can be compatible with protein biosynthetic machinery as well (3). Finally, D-amino acids, which are mirror images of the L counterparts, have been introduced selectively into functional proteins to study structural characteristics (4).

Methodology

Although this review is mainly focused on the use of unnatural amino acids in the investigation of biologic systems, a brief background of the methodology of incorporation is useful in understanding the power and application of this technology. A more comprehensive coverage of various methods can be found in Reference 5 and the references contained therein.

Chemical approaches

Global alterations to certain amino acids can be done in vitro through chemical modification of their exposed reactive side chains (6). The selectivity of chemical modification relies on the differences in chemical reactivity of amino acid functional groups. Judiciously selected chemicals will react with specific amino acids only, which allows chemical changes to be applied to that amino acid alone. Typical modification involves the thiol group of Cys, the ε-amino group of Lys, the carboxylate group of Asp and Glu, and the N-terminal amino group. The hydroxy group of Ser and Thr can be oxidized selectively when Ser and Thr are at the N-terminus of the protein (7). Side chains of Tyr and Trp can be modified selectively with transition metal catalysts (8, 9). Initial applications of this method focused on determining the functional roles of certain amino acid species for biologic activity, but they have expanded to other applications, including biophysical probe tagging, chemical cross-linking, and the conjugation of various synthetic functionalities. However, site-specific alterations are difficult using this approach because the chemical will react with all accessible target amino acids if more than one such amino acid exists in the protein.
In addition, chemical modification must be done in vitro, and affects only those side chains that are solvent-accessible. A further method to introduce unnatural amino acids into a polypeptide chain is through complete chemical synthesis (10). The predominantly used method, stepwise solid-phase peptide synthesis (SPPS), attaches the C-terminal amino acid to a solid support, and amino acids are added one at a time to the N-terminus. A clear advantage of chemical synthesis is that it enables the accurate introduction of unnatural amino acids at any site in a protein. The number of unnatural amino acids that can be introduced is limited only to the size of the chain, and chains of entirely unnatural amino acids can be produced using this method. Chemical synthesis is useful particularly for the incorporation of isotopic labels and unnatural amino acids that are toxic to cells or incompatible with the translational machinery. However, construction of a polypeptide chain using even the most advanced chemical synthesis techniques is daunting when confronted with the construction of an entire protein, as these methods currently are limited to approximately 100 amino acids (10).

Semisynthetic protein ligation methods, in which two or more protein fragments of recombinant or synthetic origin are chemically ligated to make the full-length protein (11), overcome the size limitation of SPPS. Among these methods, the native chemical ligation strategy couples peptide fragments to form a native peptide linkage, which leaves no chemical artifacts behind (12, 13). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and is then incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein. This method, like SPPS, has the power to form a native peptide linkage, which leaves no chemical artifacts behind (14). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and is then incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein fragment (14). This method, like SPPS, has the power to form a native peptide linkage, which leaves no chemical artifacts behind (12, 13). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and is then incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein fragment (14). This method, like SPPS, has the power to form a native peptide linkage, which leaves no chemical artifacts behind (12, 13). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and is then incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein fragment (14). This method, like SPPS, has the power to form a native peptide linkage, which leaves no chemical artifacts behind (12, 13). The desired unnatural amino acid is introduced in the synthetic fragment by using chemical synthesis and is then incorporated into proteins after ligation. Once this unnatural protein is folded, biochemical characterization of kinetic parameters and function can be performed. Peptide ligation in living cells is also possible. A synthetic fragment can be injected into cells to react with an endogenously produced protein fragment (14). This method, like SPPS, has the power to form a native peptide linkage, which leaves no chemical artifacts behind (12, 13).

Biochemical approaches

Methods that use the endogenous cellular machinery to introduce unnatural amino acids into proteins are not limited by protein size and will facilitate the investigation of biologic processes in vivo. A general in vitro biosynthetic method allows for the site-specific incorporation of unnatural amino acids into proteins (15). In this method, a suppressor tRNA is chemically acylated with an unnatural amino acid, and the codon of interest in the target gene is mutated to the amber stop codon, TAG. When added to cell extracts that support transcription and translation, the suppressor tRNA recognizes and selectively incorporates the attached unnatural amino acid in response to the UAG in the transcribed mRNA. Using this method, a variety of unnatural amino acids have been incorporated into proteins, regardless of position or protein size, and have been applied to a large number of problems in protein chemistry (16). Besides the amber stop codon, rare codons and extended codons also have been used to specify the unnatural amino acid (17). An extension of this method involves the microinjection of the chemically acylated tRNA and UAG-containing mutant mRNA into Xenopus oocytes (18). The endogenous oocyte protein synthesis machinery supports translation and incorporation of the unnatural amino acid. This method enables the structure-function studies of integral membrane proteins, which are generally not amenable to in vitro expression systems (19). A purified in vitro translation system that consisted of only ribosomes, initiation factors, elongation factors, mRNA, and tRNAs preloaded with desired amino acids was used to incorporate simultaneously several unnatural amino acids into peptides in response to sense codons (20). By reassigning the meaning of codons, this system ultimately may allow the synthesis of peptides and proteins that contain multiple unnatural amino acids. The drawback to these methods lies in the chemical acylation of the suppressor tRNA, which is technically demanding and can exclude certain unnatural amino acids from attachment. In addition, acylated tRNA is consumed stoichiometrically and cannot be regenerated in cells or cell extracts, which leads to low expression of the target protein.

Multiple incorporation of unnatural amino acids by using cellular machinery has been achieved in auxotrophic bacterial strains (21) and in mammalian cells (22). This method relies on the idea that aminoacyl-tRNA synthetases, although with very high substrate specificity can mischarge unnatural amino acids that are close structural analogs of the cognate amino acids. An unnatural amino acid analogous to a canonical counterpart is introduced into a bacterial strain that is incapable of producing the natural amino acid or into mammalian cells that are deprived of the natural amino acid. The translational machinery then replaces the natural amino acid with its analog in all proteins. The incorporation efficiency of unnatural amino acids can be improved by increasing the expression level of the synthetase (23) and by introducing mutations that relax the substrate specificity of the aminoacylation domain (24) or attenuate the proofreading function of the editing domain of certain synthetases (25). However, this strategy is limited because it is restricted to global replacement of one amino acid with an analog and does not allow specific single alterations with a specific protein.

It would be ideal to genetically encode an unnatural amino acid in a manner similar to that of common amino acids, enabling site-directed mutagenesis in living cells with unnatural amino acids. A general method to expand the genetic code to include unnatural amino acids was developed. It involves the generation of a new tRNA-codon-synthetase set that is specific for the unnatural amino acid and does not crosstalk with other sets for common amino acids (26). The new synthetase is evolved to charge specifically an unnatural amino acid onto the new tRNA. This tRNA recognizes a codon that does not encode any common amino acids (e.g., a stop codon or an extended codon). When expressed in cells, the new tRNA-synthetase pair enables the unnatural amino acid to be site-specifically incorporated into proteins at the unique codon with high fidelity and efficiency. This method allows the use of unnatural amino acids in the investigation of biologic systems in an in vivo setting. It may be possible to generate stable cell lines or...
transgenic animals capable of inheriting such alterations for long-term studies. However, toxic unnatural amino acids and those incompatible with the protein biosynthesis machinery cannot be incorporated using this approach.

Application of Unnatural Amino Acids

Unnatural amino acids enable the structural, chemical, and physical properties of the building blocks of proteins to be customized according to needs. Such tailored changes have contributed to our understanding of the fundamental questions of protein chemistry on the molecular and atomic level, have been used to modify and enhance protein properties, and are being exploited to control protein activities to investigate various biologic processes and to create novel biologic functions.

Protein stability

There are many factors contributing to protein stability, including hydrogen bonding, hydrophobicity, packing, and conformational entropy, among others. It is difficult to access individual contributions by using conventional mutagenesis because changing one common amino acid to another often alters several properties at a time. For example, mutagenesis to disrupt hydrogen bonds, usually by deleting one member of a hydrogen-bonded pair, will leave an unpaired hydrogen donor or acceptor and/or alter local solvation and packing interactions, all of which may lead to protein destabilization.

To determine the effect of side-chain hydrogen bonding on protein folding, Tyr27 in staphylococcal nuclease (SNase) was replaced with several isosteric, fluorinated tyrosine analogs (unnatural amino acids 1 to 3) (Fig. 2) (27). These unnatural amino acids were designed to gradually increase the strength of the Tyr27-Glu10 hydrogen bond while minimizing the steric and electronic perturbations associated with deleting one hydrogen-bonding member. The stability constants \( K \) of the corresponding mutants were found to be in the range of the hydroxyl group in the tyrosine analogs. This result provides strong evidence that intramolecular side-chain hydrogen bonds preferably stabilize the folded state of a protein relative to the unfolded state in water.

\( \alpha\)-Hydroxy acids have been used to study the contribution of the backbone hydrogen bonds to protein stability (Fig. 3). The replacement of a common amino acid with an \( \alpha\)-hydroxy acid that contains the same side-chain effectively substitutes a good hydrogen-bond acceptor (the amide carbonyl group) with a considerably weaker one (the ester carbonyl group) in a conservative manner and disrupts a potential backbone hydrogen bond because the ester linkage cannot serve as a hydrogen-bond donor as does the N-H. \( \alpha\)-Hydroxy acids were incorporated at the N-terminus, the middle, and the C-terminus of the \( \alpha\)-helix 39-50 of T4 lysozyme (28). At the N-terminus and the C-terminus, where only one hydrogen-bonding interaction is perturbed, the ester substitution destabilizes the protein by 0.9 kcal mol\(^{-1}\) and 0.7 kcal mol\(^{-1}\), respectively. In the middle of the helix, where such substitution perturbs two hydrogen bonds, the protein is destabilized by 1.7 kcal mol\(^{-1}\). In another study, Leu 14 in an antiparallel \( \beta \)-sheet of SNase was replaced with leucic acid (29). This amide-to-ester change decreases the stability by 1.5-2.5 kcal mol\(^{-1}\). Altogether, these results convincingly show that both side-chain hydrogen bonds and main-chain hydrogen bonds significantly contribute to protein stability.

To examine the importance of the packing interaction in the core of a protein, Leu133 in T4 lysozyme was replaced with a series of analogs with extended or shortened alkyl side chains (unnatural amino acids 4 to 7) (30). Leu133 lies along the edge of the largest cavity in the interior of T4 lysozyme, which makes it possible to change the bulk of the side chain with minimal conformational strain. Incorporation of (S,5)-2-amino-4-methylethanoic acid (unnatural amino acid 4) and (S)-2-amino-4-cyclopropanoic acid (unnatural amino acid 5) stabilizes T4 lysozyme by 0.6 kcal mol\(^{-1}\) and 1.24 kcal mol\(^{-1}\), respectively, which indicates that the increased bulk of buried hydrophobic residues can enhance protein stability. During protein folding, the cyclic amino acid 5 will lose less conformational entropy than does 4. That the 5-containing mutant is more stable than the 4-containing mutant suggests that there is a packing role of a residue that is misinterpreted by or remains ambiguous to conventional mutagenesis and other methods. For example, Glu43 is important for the catalytic activity of SNase.

Protein structure and function

Structural and Functional Role of Specific Residues

Unnatural amino acids can be designed to elucidate the functional role of a residue that is misinterpreted by or remains ambiguous to conventional mutagenesis and other methods. For example, Glu43 is important for the catalytic activity of SNase.
because its replacement by Asp and Gln significantly decreases the catalytic efficiency. Previous structural and mutagenesis studies suggested that Glu43 functions as a general base to activate a water molecule for hydrolyzing the phosphodiester bond of DNA. However, substitution of Glu43 with either homoglutamate (unnatural amino acid 9) or (S)-4-nitro-2-aminobutyric acid (unnatural amino acid 10) yielded mutant enzymes with kinetic constants similar to those of wide-type SNase (33). Because these two unnatural amino acids are isoelectronic and isosteric to glutamate but a much poorer base, such substitution would decrease SNase activity if Glu43 were a general base during catalysis. In addition, the X-ray crystal structure of the...
homologumate mutant showed that the carboxylate side chain of this residue occupies a position and orientation similar to that of Glu43 in the wild-type enzyme. Therefore, Glu43 may play a structural role instead and serve as a bidentate hydrogen-bond acceptor to fix the conformation of the neighboring loop.

Proline is unique among the natural amino acids in that its α-nitrogen is part of a pyrrolidine ring. The proline residue disrupts main-chain hydrogen bonding; it cannot serve as a hydrogen-bond donor because of the lack of a backbone NH moiety. Also, proline forms cis-peptide bonds at a frequency (5%) much higher than any other natural amino acids (22). In ion channels, Pro is often conserved at critical sites, such as Pro221 in the nicotinic acetylcholine receptor (nAChR) and Pro256 in the 5-hydroxytryptamine-3A receptor (5-HT3AR). To probe which feature of Pro is functionally significant, α-hydroxy acid anologs (analogs of Gly, Val, and Leu) were incorporated at these sites, which all produced mutant receptors with properties similar to the wild-type receptor (34, 35). In contrast, incorporation of canonical amino acids Gly, Ala, or Leu yielded nonfunctional receptors. Because α-hydroxy acids similarly lack the NH moiety for backbone hydrogen bonding and the nature of side chains does not affect receptor activity, these results suggest that the functional importance of the conserved Pro in both receptors is to remove backbone hydrogen bonding. A further conserved proline residue of the 5-HT3AR, Pro308, has been shown to be indispensable for channel gating using conventional mutagenesis. However, substitution of this Pro with α-hydroxy acids produced nonfunctional receptors, which suggests that the lack of backbone hydrogen bonding is not the key to the proper function of this Pro. Interestingly, proline analogs that strongly favor the cis conformer (unnatural amino acids 11 and 12) produced no gating response, but those that favor the cis conformer (unnatural amino acids 13 and 14) yielded highly sensitive channels. Moreover, a linear energy correlation was observed between the cis-trans energy gap of the proline analogs and the receptor activation (36). This study strongly suggests that the critical role of Pro308 is to provide the switch that interconverts the open and closed states of the channel through cis-trans isomerization.

Cation-π Interaction

Cation-π interaction is a noncovalent electrostatic interaction between a cation and the electrons in π orbitals, which plays an important role in protein structure, binding, and catalytic function. The energetic contribution of this interaction to proteins cannot be measured accurately with conventional mutagenesis because no positively charged natural isoctepe exist for common amino acids. To engineer a cation-π interaction in SLase, Val74, which occupies a hydrophobic pocket composed of one tyrosine side chain and two phenylalanine side chains, was replaced with the positively charged 5-methylmethionine (an unnatural amino acid) to yield SLase with homologumate mutant. Because fluorine is an electron-withdrawing group, substitution of H with F in the aromatic ring weakens the cation-π interaction. Ab initio quantum mechanics was used to predict the cation-π-binding abilities of the fluorinated tryptophanes, and the calculated binding energy has a linear relationship with receptor activation by the agonist (38). Such correlations were not observed for other aromatic residues, which suggests that the cation-π interaction indeed exists for agonist binding and pinpoints it to Trp149. This interaction was shown later as a general binding pattern between the Cys-loop superfamily of neurotransmitter receptors, such as the 5-HT3AR receptors and γ-aminobutyric acid receptors, and their cationic ligands or substrates (39).

Biophysical Probes

The site-specific introduction of biophysical probes into proteins has proven extremely powerful in revealing subtle changes of proteins with high spatial resolution. The carbon-deuterium (C-D) bond absorbs at ~2300 cm⁻¹, which is within the transparent IR window (~1800-2700 cm⁻¹) of proteins and, therefore, makes it easily observable by IR spectroscopy. The inherent fast timescale of IR spectroscopy also provides high temporal resolution. Therefore, unnatural amino acids with C-D bonds are excellent probes of protein folding and dynamics. Absorptions at different frequencies indicate the existence of multiple intermediates, and an increased line width of the absorption shows increased flexibility of the local environment. Amino acids containing C-D bonds were incorporated at different positions throughout cytochrome c (cyt c) by using semisynthetic approaches (40). By characterizing the absorption frequencies and line widths of the C-D bonds of these residues, it was found that no significant difference exists in the flexibilities of the oxidized and reduced states of cyt c. The data also show that parts of the protein exist in dynamic equilibrium with locally unfolded states and that cyt c is less stable than previous studies suggest.

A another infrared probe, p-cyano-L-phenylalanine (pCNPh), has been genetically encoded in E. coli and used to examine different ligand-bound states of the heme group in myoglobin (41). The stretching vibration of the nitrile group of pCNPh is strongly absorption and a frequency (νCN) at ~2200 cm⁻¹, which falls in the transparent window of protein IR spectra. A substitution of pCNPh was made for His64, which is at the distal face and close to the iron center of the heme group in myoglobin. In the ferric myoglobin, the Fe(III) ligand was changed from water to cyanide, νCN shifted from 2248 cm⁻¹ to 2236 cm⁻¹, which indicates a less polar active site. In the ferrous myoglobin, a νCN absorption at 2239 cm⁻¹ was observed for the linear Fe(III)CO complex, and the bent Fe(III)NO and Fe(III)O2 complexes showed a νCN absorption at 2230 cm⁻¹. These results demonstrate that the nitrile group is a sensitive probe for ligand binding and for local electronic environment. Small fluorescent probes sensitive to various environmental changes have the great potential for monitoring many biologic processes.

Unnatural Amino Acids to Investigate Biologic Processes
Unnatural Amino Acids to Investigate Biologic Processes

events as a complementary reporter for the widely used fluorescent proteins. For example, L-(7-hydroxycoumarin-4-yl)ethylglycine (CmrGly, 28) has been incorporated into holomyoglobin to study its local unfolding (42). CmrGly was incorporated at position Ser4 in helix A and at position His37 in helix C, respectively. The coumarin fluorescence intensity increases with solvent polarity. When the Ser4CmrGly mutant was unfolded with 2 M urea, its fluorescence increased 30%, which indicates that helix A is disordered. In contrast, the fluorescence intensity of the His37CmrGly mutant did not change significantly until the urea concentration was raised to 3 M. These results suggest helix C and helix A unfold at different times and concentrations of the denaturing agent.

Modification and regulation of protein activity

Green fluorescent protein (GFP), whose chromophore is autocatalytically formed by the tripeptide Ser65-Tyr66-Gly67, has become one of the most important in vivo markers for biologic studies. An aromatic amino acid at position 66 is necessary for fluorescence generation. To determine how the spectral properties of GFP could be altered by this residue, tyrosine analogs bearing different substituents at the para position of the phenyl ring (unnatural amino acids 21 to 24) were used to replace Tyr66 (43). The absorbance and fluorescence emission maxima of mutant GFPs are all blue-shifted, spanning the range from 375 to 435 nm and 428 to 498 nm, respectively. The wavelengths of the maxima increase in the order of bromo, iodo, methoxy, hydroxyl, amino, and deprotonated hydroxyl group. This shifting trend is consistent with the electron-donating ability of the substituents. In another experiment, Trp66 of the enhanced cyan fluorescent protein was replaced with \( L - 4 \)-aminotryptophan (unnatural amino acid 25) (44). The electron-donating amino group significantly red-shifts the fluorescence emission by 69 nm, which changes the color from cyan to gold.

Comparison of the \( p \)-methoxy-Phe (unnatural amino acid 22) mutant GFP with wild-type GFP also provides direct evidence for the peak assignment of GFP. Wild-type GFP has two absorbance maxima at 397 nm and 475 nm, which are believed to correspond to a neutral chromophore (phenol of Tyr66) and an anionic chromophore (phenolate anion of Tyr66), respectively. Excitation at either absorbance peak leads to a single fluorescence emission centered at 506 nm, which corresponds to the anionic chromophore in the excited state (45). Picosecond spectroscopy revealed that the excited neutral chromophore should emit at 460 nm (46). The absence of 460 nm emission in wild-type GFP suggests that an excited state proton transfer process is involved. Substitution of the hydroxyl group of Tyr with a methoxy group removes the possibility of deprotonation and proton transfer. Indeed, when Tyr66 is replaced with \( p \)-methoxy-Phe, only one absorbance maximum at 394 nm is observed, which is close to the absorbance maximum of the neutral chromophore of wild-type GFP. Moreover, only one emission maximum at 460 nm is detected for this mutant, which corroborates the ultrafast spectroscopic results (43).

The specificity of nucleic acid-binding proteins relies greatly on the hydrogen bonding between protein polar atoms and nucleic acid bases. Unnatural amino acids that can change isosterically the hydrogen-bonding pattern have been exploited to alter the substrate specificity. The \( \lambda \)-repressor recognizes the C:G pair at position 6 in the operator site OL1, and Lys4 of the \( \lambda \)-repressor is crucial for this recognition. The \( \epsilon - \text{NH}_2 \) group of Lys4 forms hydrogen bonds with the carbonyl group of Asn55 and the 6-oxo group of the guanine, functioning as two hydrogen bond donors. Substitution of Lys4 with isosteric \( S -(2 \)-hydroxyethyl)-cysteine changes the \( \epsilon - \text{NH}_2 \) to the \( \text{–OH} \) group, which now should accept hydrogen bonding from the amino group of adenine while preserving the hydroxyl group of Lys4 with Asn55 as a donor (Fig. 4). In fact, after the unnatural amino acid was introduced into the \( \lambda \)-repressor through site-directed mutagenesis and chemical modification, the binding specificity was switched from the C:G to T:A base pair (47).

The chirality of \( D \)-amino acids has been harnessed for pharmaceutical purposes. \( D \)-peptide ligands should be resistant to proteolytic degradation and thus are more desirable as...
Unnatural Amino Acids to Investigate Biologic Processes

Drugs. However, large libraries of D conformers cannot be encoded genetically and expressed for selection. A method termed mirror-image display solved this problem in an intriguing way (48). An L-peptide library is encoded genetically and displayed on the phage surface, and peptides of this library are selected by the target protein that is synthesized using all D-amino acids. The identified L-peptide then is resynthesized using D-amino acids, which should interact with the target protein of the natural handedness for reasons of symmetry. This approach has been used successfully to identify D-peptides that bind the Src homology 3 domain of c-Src and the HIV-1 gp41 protein (48, 49).

Unnatural amino acids that mimic posttranslational modifications can be used to control protein functions. For example, protein phosphorylation regulates many signal transduction pathways and is a reversible process catalyzed by various phosphatases and kinases. The dynamic change of the phosphorylation status of a protein makes it difficult to study the effect of this modification in detail. The generation of metabolically stable phosphoproteins would be useful to dissect the function and to direct signal transduction. Unnatural amino acid p-carboxymethyl-L-phenylalanine (pCMF, 26) is a non-hydrolyzable analog of phosphotyrosine and was found capable of mimicking the phosphorylated state of Tyr. This capability was demonstrated in a model phosphoprotein, the human signal transducer and activator of transcription-1 (STAT1). STAT1 has only a weak affinity for DNA, but during phosphorylation of Tyr701, STAT1 forms a homodimer and strongly binds a DNA duplex that contains M67 sites. The mutant STAT1 with Tyr701 substituted with pCMF also bound the M67-containing DNA duplex tightly, which suggests that pCMF could replace phosphotyrosine in the generation of constitutively active phosphoproteins (50).

The development of photoactive amino acids provides researchers with an extremely useful tool not only to probe biologic function but also to control spatially and temporally a variety of biologic processes. One strategy is to attach a suitable photoremovable protecting group to the amino acid, which renders the amino acid inactive. Photolysis releases the caging group and converts the amino acid to an active form, which generates abrupt or localized changes to the target protein. The 2-nitrobenzyl derivative is the most prevalent form for caged compounds. For example, the conserved Ser1802 at the upstream splice junction of the self-splicing DNA polymerase of Thermococcosis litoralis was substituted with o-(2-nitrobenzyl)serine (Fig. 5a). The full-length precursor protein underwent photolysis only when the unnatural residue was reverted back to wild-type Ser during photolysis (51). In other examples, o-nitrobenzyltyrosine (Fig. 5b) was used to replace Tyr93 or Tyr198 in the α subunit of the nAChR. These two Tyr residues are highly conserved for agonist binding. Millisecond flashes of light at 300–350 nm decaged the protected tyrosines and produced abrupt increments of currents that were conducted by the ion channel (52). Also, o-nitrobenzyltyrosine has been incorporated at the essential Tyr93 site of β-galactosidase to activate its enzymatic activity by using light both in vitro and in E. coli (53). Mutation of the active-site cysteine residue in the proapoptotic protease caspase 3 to o-nitrobenzylcysteine led to a catalytically inactive enzyme, whose activity could be restored by photocleavage (54). In addition to caging the active side chains, the 2-nitrobenzyl group

Figure 5 Photolysis of 2-nitrobenzyl caged serine (a) and tyrosine (b) restores the wild-type residues. Photolysis of 2-nitrophenyl glycine (c) cleaves the protein backbone.
Unnatural Amino Acids to Investigate Biologic Processes

Figure 6 (a) The geometrical change resultant from the cis-trans isomerization of azobenzene moves an ion channel blocker in and out of the ion channel to close and open the ion channel, respectively. Such activity was used to modulate the spontaneously firing hippocampal neurons. The firing frequency is significantly decreased when the azobenzene is in the cis-form after irradiation at 390 nm. Normal firing behavior is restored during irradiation at 500 nm (reprinted from (57), Copyright 2005, with permission from Elsevier). (b) Structure of phenylalanine-4′-azobenzene (AzoPhe) in trans-form and gel mobility shift assay to determine the binding affinity of the catabolite activator protein (CAP) to the lactose promoter DNA fragment (reprinted with permission from (56), Copyright 2006, American Chemical Society). Lane 1, DNA only. Lane 2, DNA+ wild-type CAP. Lane 3, DNA+ CAP with AzoPhe incorporated at residue 71 (after irradiation at 334 nm). Lane 4, DNA+ CAP with AzoPhe incorporated at residue 71 (before irradiation at 334 nm). Substitution of Ile71 with trans AzoPhe in CAP results in a fourfold decrease of the binding constant $K_b$ of the CAP for its promoter sequence. Photoirradiation at 334 nm partially converts the trans AzoPhe to the cis-form and decreases the $K_b$ by another fourfold. The latter affinity loss can be completely recovered after irradiation at $>420$ nm, which switches the cis-form back to the predominant trans-state.

Photolysis of a caged amino acid residue is an irreversible process. Reversible modulation can be achieved with the photocromic azobenzene compounds. Azobenzene undergoes a reversible cis-trans isomerization: The more stable trans isomer can be converted to the cis isomer upon illumination at 320–340 nm, and the cis-form can revert to trans-form either thermally or by irradiation at $>420$ nm. The resultant change in geometry and/or dipole of the compound can be used for regulating protein activity. For example, a known K$^+$ channel blocker, tetra-ethyl ammonium, was linked via an azobenzene group to a cysteine that was introduced at specific sites of a K$^+$ ion channel (Fig. 6a). When the azobenzene group isomerizes between the extended trans-form and the shorter cis-form in response to specific wavelengths of light, the structural change moves the blocker into or out of channel-blocking position and, thus, opens and closes the ion channel, respectively (58). Such photomodulation can be used to control neuronal activity non-invasively. The azobenzene group has been genetically encoded in the form of phenylalanine-4-azobenzene (AzoPhe). AzoPhe was incorporated at the Ile71 site of the E. coli catabolite activator protein (CAP), a transcriptional activator. Its binding affinity for the promoter sequence decreased fourfold after irradiation at 334 nm (Fig 6b), which converts the predominant trans AzoPhe to the cis-form. The isomerized cis AzoPhe then was switched back to the trans-state by irradiation at $>420$ nm, after which the affinity of the protein for the promoter was completely recovered (56).

Future Directions

The examples summarized here are only representative and by no means comprehensive. Many unnatural amino acids now can be incorporated, but simply have yet to be used in the investigation of biologic function. Unnatural amino acids that contain photocross-linkers, biophysical probes, chemical moieties with unique reactivities, and posttranslational modifications, among many others, have much promise in their capabilities. The use of these amino acids will expand the capabilities of probing protein structure and function as well as protein-involved biologic processes. The methodology of incorporation is advancing as well. It may be possible to genetically encode unnatural amino acids in many other cell types and organisms. The incorporation of multiple unnatural amino acids simultaneously by using

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
extended codons may enable more complex investigations to be performed. A different work using unnatural amino acids can lead to the design and synthesis of novel and diverse biologic functions. By incorporating specific chemical moieties and physical characteristics into proteins, new protein properties may be discovered and used. Such exploration can be attempted either rationally or combinatorially. Diversities of protein libraries would be increased greatly by the addition of only a few unnatural amino acids, which may enhance the probability of discovering proteins that contain novel properties and functions. It is easy to see how unnatural amino acids can be extended into the pharmaceutical industry to create more efficient therapeutics. Finally, the creation of a sustainable organism that is capable of using unnatural amino acids will enable the investigation of the evolution of the genetic code on this planet.

Acknowledgments

We thank A. R. Parrish for help editing this manuscript. J. Xu is supported in part by the Pioneer Fellowship of the Salk Institute. L. Wang is a Searle Scholar and a Beckman Young Investigator.

References

Unnatural Amino Acids to Investigate Biologic Processes


See Also

Amino Acids, Chemistry of
Chemical Lipidation: Peptide Synthesis
Natural and Unnatural Amino Acids, Synthesis of
Proteins, Chemical Modification of Proteins
Proteins: Structure, Function and Stability
Membrane Potentials in Living Systems, Tools to Measure

Paul O’Shea, Cell Biophysics Group, School of Biology, University of Nottingham, University Park, Nottingham, United Kingdom
doi: 10.1002/9780470048672.wecb310

Membrane potentials are a ubiquitous feature of all living cells and play different roles in manifold cellular processes. Three different membrane potentials are known to exist that include the transmembrane potential difference, the surface electrostatic potential and the membrane dipole potential. Each of these are discussed in terms of their different physical origins and characteristics together with representative examples of how they feature in cell biology as well as some examples in physiology. Techniques for the quantitative measurement of each of these potentials are described in both model systems as well as with living systems. As many of these practices involve the use of spectroscopic technologies (particularly fluorescence) they lend themselves to spatial (single cell) imaging applications and examples of the biological roles that spatial variation or localisation of each of these potentials are also outlined.

Biological Background

Membranes act as one of the major macromolecular assemblies of living cells (both prokaryotic and eukaryotic) but also feature as physical envelopes in viruses such as influenza and HIV. Typically cell membranes act as semi-permeable structures that compartmentalise cellular processes and behave as 2D fluids, rapidly annealing any fractures by 2D molecular flow. Membranes, however, undertake many more biological roles than these simple generalizations imply as they also act as platforms for many important biochemical processes. Similarly, it has recently been appreciated that membranes exhibit localised structures in the form of microdomains (i.e. so called rafts) as shown in Fig. 1. These act as structures upon and within which, additional physico-chemical mechanisms may operate. The following sections outline the biological features of the trinity of membrane potentials (1) and their physico-chemical origins. Practical methodologies for their respective measurements are then described that exploit these different molecular origins.

The transmembrane potential difference: roles in signaling and energy transduction

The first membrane potential to be considered resides in a membrane’s ability to act as a permeability barrier or resistance to solute (particularly electrolyte) movement. Thus ions that are actively transported across the membrane by membrane proteins are not easily able to return to their former compartment. This leads to a thermodynamic gradient that is utilised by cells in two ways. The first, takes advantage of the ionic concentration gradient of a specified ion across the membrane. The second due to a small electrical capacity of the membrane (ca. 1 µF cm⁻²) may lead to an electrical gradient across the membrane. Thus chemical or light energy that is ‘spent’ transporting the ions across the membrane is partially ‘conserved’ as the ionic gradient across the membrane (e.g. as illustrated in Fig. 2). Two enormously important biological examples featuring these processes are worth emphasising, the first example is that the conservation of light or chemical (oxidation-reduction)
energy as an ionic/electric gradient is an essential part of the process of oxidative or photosynthetic phosphorylation. The ionic gradient in these latter cases resides in a proton gradient across the mitochondrial or chloroplast membranes respectively (2). Interestingly, in the case of the mitochondria, the proton gradient manifests mostly as an electrical gradient (3) due as mentioned above, to the relatively small electrical capacity of the mitochondrial membrane with only a small pH gradient. Whereas due to the relatively easy movement of Mg\(^{2+}\) ions across the chloroplast membrane, the electrical capacity is effectively much increased and the proton gradient then manifests mostly as a pH gradient with a small electrical gradient. Both these manifestations of the electrochemical potential difference of the proton across the membrane however, are equivalent and are utilised to manufacture ATP directed through the membrane ATPases. In fact, this electrochemical gradient is utilized essentially as a power source for many additional utilities such as coupled solute transport, to drive the flagellum movement of some bacteria as well as for protein import across prokaryotic and eukaryotic membranes (4).

The second general biological example is that the electrical gradient so generated can be utilised as a signalling mechanism such as in nervous transmission in neural cells or as a driving force for ionic currents that mediate cell signaling. These signaling mechanisms are both complex and ubiquitous with ion channels activated through soluble or membrane-bound ‘ligands’ or 2nd messengers that upon opening permit e.g. calcium ion movements driven by the membrane potential that then elicits further cell signaling cascades for neurotransduction, muscle contraction and nuclear translocation of molecular switches to turn genes on or off. Typically there are spatial elements in these signaling elements and thus there are also many measurement strategies that include imaging modalities as part of the experimental goals.

Finally, in the case of the role of membrane potential in nervous conduction, a travelling wave of an electrical potential gradient (voltage) is used to relay signals over relatively long distances along the cell membrane i.e. as in axonal linkages between cells (see e.g. 5). This latter manifestation of a “membrane potential” as part of the nerve impulse is perhaps the most commonly understood membrane potential and is described very fully in Reference 6.

The membrane surface potential: roles in modulating ion channels and cell–cell interactions

The surface potential is now understood to play important roles in many physiological processes. These range from the simplest involving coulombic repulsion between adjacent membranes that prevent cells sticking to each other. Thus sialic acid as a sugar moiety to the membrane protein glycoporphin present on the cell surface of many cells and particularly red blood cells appears to prevents cell–cell aggregation. Under circumstances of uncontrolled-diabetes, however, the nature of the sugars on the erythrocyte cell surface may change with sialic acid replaced by uncharged sugars. This reduces the coulombic repulsion between the red cells allowing thrombosis to occur (7, 8).

The contribution to the cell membrane surface potential by sialic acid is also implicated in gating the voltage dependence of Na\(^{+}\) Channels (9). This kind of electrostatic effect on ion pump/channel loading/unloading (with other contributors to the surface potential such as phosphatidylserine), however, has been considered for many years. McLaughlin as long ago as 1971, for example, suggested that surface potential changes were responsible for negative or positive shifts of the current-voltage relationships of neural tissue as embodied in the Hodgkin-Huxley equations (5). This kind of process is the result of the effect of the surface potential on the activity of ions on a membrane surface and leads to changes of e.g. the pK of membrane surface-located acid-base groups (10, 11). We have exploited this latter phenomenon to develop a panel of fluorescent phospholipids as real-time measures of the membrane surface potential (12). By measuring small changes of the membrane surface potential in this manner due to the interaction of charged molecules, it proved possible to study the kinetics and thermodynamics of intermolecular interactions with membranes (13).

The membrane dipole potential: role in modulating microdomain-located membrane proteins

The presence and biological roles of cell membrane microdomains (often referred to as membrane rafts) are discussed elsewhere in the WECB although it must be conceded that presently, this subject remains slightly contentious. Nevertheless in our hands we have clearly observed these structures in both artificial model systems (14) and the membranes of living cells (1). We published the first theoretical model of the possible mechanisms of assembly and disassembly of these structures (14) and firmly believe they are a feature of living cell membranes. The accepted functions of microdomains appear to revolve around their ability to act as local platforms for endocytosis/ectocytosis or to localise reactants whether they be small ligands or proteins. In this way rafts are localised in the raft aiding their interaction or they are sequestered from the more fluid membrane preventing their interaction with potential...
partners resident in the fluid membrane. Both these processes, however, are conceptually analogous. We considered that membrane microdomains may exhibit a vastly different membrane dipole potential to that of the fluid phase membrane due to their different lipid packing and complement of sterols and lipids (etc). We later demonstrated this was indeed a fact of the various lipids present (15) and also showed that this parameter had a significant effect on membrane protein conformation (16).

This work demonstrated with representative ligand-receptor systems that this behaviour may alter the function of such receptor systems depending on whether they were resident in the rafts or in the fluid phase regions of the membrane (15, 17). Following this a number of other laboratories also observed modulation of membrane behaviour via the agency of the dipole potential (18). Thus the membrane dipole potential appears to be an additional tool for which membrane protein function can be controlled in particular localities. The explicit molecular mechanism by which this may take place is discussed in the next section.

Physical Chemistry Background of Membrane Potentials

It is axiomatic that biological systems must adhere to the physico-chemical laws of nature. Sometimes, biology when appearing to contradict such laws (eg organisms and cells representing low entropy structures, historically brought it into conflict with the 2nd law of thermodynamics), however, merely serves to illustrate the ingenuity of nature in working within the laws. Electrostatics and electrolydynamics together represent one such cornerstone of nature that biology exhibits great expertise in making use of to derive exquisite control of biological function, as briefly outlined above.

Membrane potentials manifest in a number of different ways as part of the process of biology and relate to their physical origins and will be described in this section. Taking each type of the three potentials in turn, the transmembrane potential difference is considered first as this is the most well known and features prominently in undergraduate text books dealing with biochemistry and physiology. Historically this membrane potential is also known simply as the membrane potential, unfortunately has lead to some confusion as it shares nomenclature with the 'other' membrane potentials. In this article the same nomenclature is adopted as that described in a comprehensive treatise on biological membrane potentials (1) that is consistent with the 'other' discipline of electrochemistry, so unifying the chemistry and biology communities' versions of the nomenclature.

The transmembrane potential difference (\(\Delta \phi_m\))

Membranes represent permeability barriers to the movement of ions (and electrons) and this phenomenon is utilised in a number of ways in biological systems as mentioned in the previous section. In terms of generalising and for the sake of briefness, this gradient of ions has two principal uses, the first as an energy store albeit labile and secondly a means of shuttling ions across the membrane. In terms of an analogy with the charging of a capacitor, this well known latter phenomenon is described by the following expressions:

\[
V = \frac{q}{C} = \Delta \phi = \text{Em} \text{ or } Vm = \Delta \phi_m \tag{1}
\]

where, \(V\) is the electrical potential difference across the membrane ie. \(Vm\) or \(Em\) and \(\Delta \phi\) is the capacitance of the membrane (dielectric) and \(q\) is the number of charges transported across the membrane. The term, \(\Delta \phi_m\) is included as this represents a more consistent nomenclature, and would be better recognised by disciplines outside biology (e.g. Electrochemists, \(\Delta \phi\) is the form utilised by biochemists (2) and \(Vm\) or \(Em\) is the form utilised by Physiologists (6).

The transport of electrical charge across a membrane may take the form of cation, anion (both inorganic and organic) or electron transport all requiring an energetic input. It can be passive due to membrane leaks or mediated by ionophores such as valinomycin. All other things being equal it is possible to utilise equation 1 to calculate a trans-membrane (ie trans-dielectric) potential difference (19). Thus, under these circumstances there is a simple and direct equivalence of the voltage across the membrane (ie \(\Delta \phi_m\)) to the Gibbs Free energy (ie. \(\Delta G\)). In biological membranes, however, the transport is frequently coupled either to the movements of other ions due to co- or anti-porting enzymes, or to chemical reactions as in the case of oxidative phosphorylation or hydrolysis of ATP. The value of \(\Delta \phi_m\) established in such complex systems is governed by the relation \(\Sigma Z_i \phi_i = 0\), where the sum has to include all flows \(j\) of charged species \(Z_i\) across the membrane (1). When this relation is satisfied, charging of the membrane capacity has ceased, and a (pseudo) steady state with a (approximately) constant \(\Delta \phi_m\) is then attained. Its value depends on the difference in chemical potential (\(\Delta \mu\)) of all transported species and on the affinities of coupled chemical reactions. The pertinent relations are usually transcendental and cannot be solved explicitly except for some special cases. In particular, if only one species permeates through the membrane, a true equilibrium state is reached and the resulting transmembrane electrical potential difference is described by what is known as the Nernst equation (below). This equivalence, however, becomes more complicated because biological membranes are not totally impermeable to ions (i.e. there is leakage back across the membrane - e.g. Reference 20), the concentration gradients must also be included in the Gibbs free energy of the expression and finally there is usually a coupling of the transport mechanisms to the movements of other ions as well as to the energy input (as in oxidative phosphorylation or the hydrolysis of ATP – see Reference 2). Nevertheless, ion gradients are established across many membranes whereby the ion leakage is balanced by further energy-linked ion transport. Biological membranes, therefore, exhibit quasi-equilibrium transmembrane concentration gradients of charge based on selective transport of the ions found in physiological electrolytes. This is described by the well known Nernst equation which incorporates the transmembrane concentration gradient of electric...
charge as follows:

$$\Delta \phi_{i} = \left( e Z_i \right) \log \left( C_{\text{inside}} / C_{\text{outside}} \right)$$  \hspace{1cm} (2)$$

In which $\phi$ is an abbreviation of $\text{K}T/e \approx RT/F$; with $F$ = fara-
day constant, $R$ gas constant and $C_i$ indicate the concentration of
the respective ionic species in the aqueous phases separated
by the membrane.

Thus, the Gibbs Free energy expression for such an ionic gra-
dient across the membrane consisting of $n$ positively charged
species $C_i$ would be of the form of the following expression:

$$\Delta G = - nF \Delta \phi + 2.3RT \log \left( C_i^{\text{inside}} / C_i^{\text{outside}} \right)$$

The electrical component ($\Delta \phi_{el}$) and the chemical component
($\Delta \phi_{ch}$) are combined to indicate the thermody-
namic potential difference due to the ionic gradient.

Finally, its also worth emphasizing that related to the trans-
membrane potential difference, a transmembrane gradient of
electric charge known as the Donnan potential is also known to ex-
ist. The Donnan potential arises from the inability of larger
charged macromolecules or other fixed charges to move across
a membrane and for counterions to compensate this gradient.

The membrane surface potential ($\phi_s$)

Many of the electrical phenomena that feature on membrane
surfaces are similar to those found in electrochemistry. Thus
derivation of formalisms describing interactions of ions with
charged membrane surfaces have evolved from and are in many
ways similar to those that have been developed to describe the
ion relations of electrodes in electrochemistry. It's worth empha-
sizing, however that care must be taken when simply reworking
concepts that originate from electrochemistry, as there are im-
portant differences between membranes and electrodes that may
complicate formal descriptions. Electrodes are essentially ho-
mogeneous hard solid metal surfaces, whereas membranes are
soft, fluid, highly heterogeneous interfaces that interact and actu-
ally require their aqueous environment to exist (i.e. through the
hydrophobic effect).

Ions are attracted towards the charged membrane surface but
the hydrated radius of the ions in the layers adjacent to the
surface prevents them from moving to the membrane surface
without becoming actually adsorbed. The initial treatments of
this phenomena by Gouy and Chapman in terms of the interac-
tions of ions with electrodes involved treatment of the ions as
point charges but a modification of this model was presented by
Stern who developed the much earlier Helmoltz-Perrin model
with that of Gouy-Chapman (21). The Helmoltz-Perrin model
described the double-layer concept that counterions formed an
ionic “sheet” upon a charged surface, the ions being at least
partially dehydrated in the direction of the surface. Much of the
history of the development of these ideas as directed towards
metal surfaces in aqueous media (i.e. electrodes) may be found
in comprehensive electrochemistry texts (21).

The forces involved in the specific ionic-interactions with
membranes were initially thought predominantly to be electro-
static and van der Waals in nature, and large enough to counter
the thermal motion/diffusion of the ions away from the interface
to the bulk medium. Later, it was suggested that water molecules
may adopt a specific orientation upon the charged surface due
to their permanent dipoles (this latter feature has some bearing
on an understanding of the dipole potential described below).

A number of diffuse phases or layers as illustrated in Fig. 3a,
have been identified which are thought to exist adjacent to the
surfaces of membranes. These include the so-called Stern layer
(Fig. 3a) which represents contact adsorbed counter-ions (either
totally or partially dehydrated) and oriented water molecules.
The plane running through the centre of the contact adsorbed
ions is referred to as the Inner-Helmholtz plane. The first layer
of hydrated ions is then referred to as the Outer-Helmholtz
plane. The Gouy-Chapman-Stern model was formalised by con-
sidering the adsorption of counter-ions only, the process being
approximated by a Langmuir-type adsorption isotherm (1, 21).

Surface potentials at the electrode-solution interface have
been described by a number of formalisms. The most successful
of these was offered originally by Gouy and Chapman with sub-
sequent elaborations from Chapman, Stern, Bockris etc.,
(outlined in ref 1 & 21). M-L aughlin (22) and others (inlined in
1) suggested that a combination of the Poisson and Boltzmann
equations best describes the state of affairs in the space between
the membrane surface and the bulk phase aqueous solution i.e.
the electrode-water interface. The Poisson-Boltzmann equation,
with defined boundary conditions can be solved analytically
(1, 22) to yield an expression for the surface potential as follows:

$$\phi_s = \left( 2e o \right) / \left( 2A \ln \left( 2e o / (2e o) \right) \right)$$

where $\phi_s$ as in Eqn. 4 and the abbreviation $\phi_s = \phi_{G,C}$
whereby $\phi$ represents the surface charge density, and $\lambda_s$ is known
as the Debye length. At 25 $^\circ$C ($e o = k T/2e N_A)^{1/2} = 0.304$ nm
$M^{-1/2}$ for an aqueous phase with $o = 80$.

Thus the potential profile in the diffuse layer shown in Fig. 3a
is given by:

$$\phi(x) = \phi_s \ln \left( \left( 1 + \tan h (x / \lambda) \right) / \left( 1 - \tan h (x / \lambda) \right) \right)$$

for $x \geq 0$  \hspace{1cm} (5)

The membrane dipole potential ($\phi_d$)

Many membrane components such as phospholipid include moi-
eties such as the $\text{C}^\text{ii} = \text{O}^\text{ii}$ and $\text{C}^\text{ii} = \text{P}^\text{ii}$ exhibit polarisation.

The membrane dipole potential $\phi_d$ has its origins in the dipole
moments of polar groups from the lipidic components of the
bi-layer, it seems likely that the water molecules at the molec-
ular surface of membrane also make a contribution (1). The
organisation of the membrane components that contribute to this
potential have been verified from neutron diffraction studies and
NMR spectroscopy (23) and quite recently using cryo-EM tech-
niques has also added quantitative estimations of the potential
(24). These dipolar groups seem to be oriented in a way such
that the potential located towards the hydrophobic interior of the
membrane is positive with respect to the pole located towards
the external aqueous phases, and $\phi_d$ has a magnitude of several hundred millivolts (typically about 300 mV see Fig. 4).

Symbolically formalising the measured molecular arrangements in order to undertake modelling of the membrane dipole potential is also not as straightforward as it may seem. It is possible to begin by identifying a vector drawn from the point of the negative charge $-Q$ to the positive charge $+Q$ of any dipole which is the familiar electric dipole moment $\mathbf{p}$. The magnitude is $Qa$, with $a$ defining the distance between the centres of the charge density. The potential at a given point with a position vector $\mathbf{r}$ with respect to the centre of the dipole can be expressed as:

$$\phi = \frac{\mathbf{p} \cdot \mathbf{r}}{(4\pi \varepsilon_0 \varepsilon_r r^3)} = \frac{\mathbf{p} \cdot \cos \theta}{(4\pi \varepsilon_0 \varepsilon_r r^2)}$$

for $r \gg a$, where $\theta$ is the angle between $\mathbf{r}$ and $\mathbf{p}$. Higher order terms become influential if $r \gg a$ is not fulfilled.

This expression, however, only deals with a single molecular dipole and as many such dipoles would be required to describe the overall membrane dipole potential but as a mean-field expression, this term is practical and cumulatively offers the approximation of the estimated dipolar organisation shown in Fig. 4. A further complication, however, involves the solvent environment and this too is also often dealt with as a mean field or in a continuum manner. But the relative permittivity (or dielectric constant) $\varepsilon_r$ cannot be considered to possess the same value throughout the multiphase system represented by a membrane in an aqueous medium. The permittivity profile has been measured to vary from about 78.5 in the bulk aqueous.
There are many techniques available for the measurement of membrane surface potential (φs) to around 2 in the membrane interior. Measurements of transmembrane potential differences in living cells are well established with the methods fairly reliable and robust. Many of these membrane potential measurement strategies are generic and hold for plasma membranes of neurons and other excitable tissues.

Chemical Tools & Techniques

Measurements of the transmembrane potential difference (Δφm)

There are many techniques available for the measurement of Δφm (e.g., 1, 2, 5) historically these involved the use of electrodes in which the voltage in one compartment would be compared to that in another separated by the membrane. And whilst this is not a major problem with larger structures such as eukaryotic cells, with bacteria penetration by an electrode would represent a major invasive procedure. Electrodes are still utilized routinely (6, 28) but the use of spectroscopic probes has evolved as perhaps the most commonly utilized current technology (26). Experimental strategies involve measurement of the redistribution of membrane-permeant indicators that migrate according to the Nernst equation, upon rearranging yields:

\[ \Delta \phi_m = \frac{RT \ln 10}{Z_i F} \left( \frac{c_i,s}{c_i,b} \exp(-Z_i F \phi_s/(RT)) \right) \]  

that when introduced into the logarithmic form of the Henderson-Hasselbalch equation, upon rearranging yields:

\[ \log(c_H,b)/c_{HB}) = pH - (pK_a - F \phi_s/(RT \ln 10)) \]  


Measurements of the membrane surface potential (φs)

In our laboratory we have developed novel fluorescence technologies that illuminate the respective membrane potentials. For measurements of φs this involved the synthesis of fluorophores attached to a phospholipid molecule that have the advantage of being virtually non-invasive as they are used at very low concentrations and do not perturb the membrane. One such probe molecule, FluoresceinPhosphatidylEthanolamine (FPE), has proved to be a versatile indicator of the electrostatic nature of the membrane surface in both artificial and cellular membrane systems (12, 17, 25). FPE is sensitive to changes in the surface potential φs at the membrane-solution interface because the fluorescent moiety of the FPE lies precisely at the membrane solution interface (see Fig. 3a). Any changes in the number of surface charges at the membrane, such as the binding of an inorganic ion or a charged oligopeptide, will cause an alteration in φs. These probe molecules operate in the following manner: thus according to the Boltzmann equation the concentration of a charged species at the membrane surface is

\[ c_{i,s} = c_{i,b} \exp(-Z_i F \phi_s/(RT)) \]  


Figure 4  Profile of dipole potential in the membrane.
Measurements of the dipolar membrane potential ($\phi_D$)

The dipole potential $\phi_D$ may be measured using a series of potentiometric fluorescent indicators that operate by electrochromic mechanisms. $1-(3$-sulfonatopropyl)$-4-(p$-$2$(di-n-octylamino)-6-naphthyl) vinyl) pyridinium betaine (known as di-8-ANEPPS) in particular has been successfully applied to the measurement of $\phi_D$ using dual-wavelength ratiometric fluorescence methods (15, 16, 30). This method forestalled problems arising from small differences in dye concentration between different samples, dye bleaching or the influence of light scattering on the fluorescence measurements.

The excitation spectrum of di-8-ANEPPS is altered when it lines up (symmetrically or asymmetrically) with the membrane dipoles causing electronic redistributions within the probe molecule (see e.g. Fig. 5a). This promotes red or blue shifts in the excitation spectrum depending on the magnitude and direction of the dipole moment of the ambient environment that the probe finds itself in as shown in Fig. 5b. Preparation of membranes with sterols etc (ie that possess quite different dipole-moments to PC) promote changes in the membrane dipole potential, and significant variations of the intensity and position of the excitation maximum are observed. The excitation spectrum of di-8-ANEPPS in phosphatidylcholine (PC) membranes for example is significantly altered when 15 mol% of either 6-ketocholesterol (KC) or phloretin are added to such membranes. In the case of phloretin the difference spectrum has a minimum at 450 nm and a maximum at 520 nm (Fig. 5b). In the case of KC however, the difference spectrum has a maximum at 450 nm and a minimum at 520 nm, which is the opposite effect to that of phloretin.

Comprehensive studies of the dependence of the relative magnitude of the dipole potential on the membrane lipidic composition have been reported previously and a summary is reported in ref. 15. Of quite some interest, however, is the possibility that di-8-ANEPPS may be used to indicate the interactions of some macromolecules with membranes as shown in the equation above, the number of molecules that become bound. If $\phi_D$ is measurably influenced by peptides/proteins that insert (at least partially) into the membrane and just as importantly vice versa.

Spatial Imaging of Membrane Potentials

The dominant technologies (outlined above) utilised to determine biological membrane potentials are optically based and particularly involve fluorescence. Thus spatial imaging is a logical extension of these spectroscopic applications. The use of a modified electrode technology has also been utilized to acquire imaging data in which the electrode is rastered over a cell surface to identify localization of ion transport behaviour and by implication spatial variations of the transmembrane potential (28). Measurements that include a spatial element have become extremely important in addressing biological problems as living cells exhibit enormous heterogeneity in the spatial disposition of these potentials. Lateral gradients of each of the membrane potentials, therefore, are implicated in manifold biological processes. In our laboratories for example we have developed the concepts (outlined above) relating to the interesting possibility that elevated dipole fields within membrane microdomains (15) modulate protein structure and that this modulates signaling activity (1, 17) or the activity of ion channels (18). We developed the use of di-8-ANEPPS to determine how the membrane dipole potential may be utilised to reveal macro-molecular interactions with membranes (15, 16). By illuminating spatial variations of this parameter and particularly how it varies in membrane rafts it led us to propose that the value of $\phi_D$ is quite different in membrane rafts as compared to fluid mosaic membranes (1, 17). Figure 6 illustrates this striking heterogeneity of the $\phi_D$ about the cell surface with measures currently being taken to correlate this or co-locate such signals that are emanating from the raft microdomains within membranes. The latter is a significant issue as there remains much confusion as to the in vivo existence of membrane rafts (33).

Similar approaches utilising such indicators as FPE to visualise the membrane surface potential $\phi_D$ are also routinely employed in our laboratories (17). By correlating the change of the fluorescence and hence surface potential with the addition of net electric charges from the macromolecule that becomes bound, it is possible to quantitate on the basis of the poisson-boltzmann equation above, the number of molecules that become bound. This allows us for example to determine localised molecular interactions on the membrane surface (17, 34).

Future Studies of Biological Roles and of Technologies for Measurement

Future aims and understanding of the biological roles of membrane potentials

Each of the membrane potentials described above are fundamental properties of living cells and as such there will be a growing appreciation of their roles in cell biology. In particular the development of an understanding of their roles in complex signaling pathways in which the coupling between the electrical and chemical signaling networks are clarified. Ci-VSP, for example (37) is a recently described protein with sequence similarity to both the voltage-sensing domain of a voltage-gated potassium channel and the phosphate PTEN, and functions as a transmembrane phosphoinositide phosphatase that is regulated by changes in voltage across the plasma membrane (37). More extensive and integrated conceptual developments of the cellular role of all aspects of membrane potentials will undoubtedly emerge as the new biological sub-discipline known as ‘systems biology’ matures (38).
Figure 5. Spectroscopic tools for identifying the membrane dipole potential. Fig. 5a indicates the excitation spectrum of different membranes labeled with di-8-anepps with the emission collected at 580nm. Spectra are shown for membranes made up of 100% phosphatidylcholine (−), membranes made up of phosphatidylcholine and 15mol% 8-Ketocholestanol (⋯⋯) and membranes made up of 15 mol% phloretin (····). These excitation spectra are better visualized as the difference spectra shown in 5b in which membrane made up of different lipid mixtures are compared to a reference spectrum, typically these are normalized and compared to phosphatidylcholine and shown as LHS upper (phloretin) and lower (8-ketocholestanol) spectra. The spectra upper RHS indicates the spectral shift of a membrane following interaction with a peptide and the lower plot illustrates the time course of the interactions in which the upper and lower limbs of the spectra are ratioed and plotted against time.
Future technologies for measuring membrane potentials

With the introduction of the green fluorescent protein (GFP) initially as a fluorescent label to track proteins, subsequent development of the technology to manufacture other colours (X-FP) has been used to visualize protein-protein interactions through confocal microscopy. Di-8-anepps labels all the intracellular membranes as well as the plasma membrane and thus indicates these membranes also possess very localized regions of elevated dipole potentials indicative of intracellular membrane microdomains. The LHS image indicates a bright field image of the same cell.

References

Membrane Potentials in Living Systems: Tools to Measure


Further Reading


See Also

Membrane Assembly in Living Systems
Bioenergetics and Oxidative Metabolism
Lipid Rafts
Ion Transport
Imaging Techniques: Overview of Applications in Chemical Biology
The methods and applications of NMR-based metabolomics are introduced in this article. As important tools for systems biology research, metabolomics, metabonomics, and metabolite profiling have been applied successfully for biomarker exploration at the molecular level. In addition to the detection and identification of putative disease biomarkers, NMR-based metabolomics also can be performed in many areas including nutritional science, disease monitoring, and drug toxicology studies. In the following sections, basic and advanced NMR technologies associated with metabolomics are summarized. Also, to process the complex spectral data and obtain useful information, pattern recognition techniques or multivariate statistical methods are necessary. These methods successfully simplify the multivariable data set, and their improvement may enhance the development of metabolomics. Finally, several important applications of NMR-based metabolomics are summarized, and examples are described.

Introduction

Metabolomics and the closely related metabonomics and metabolite profiling represent a growing field in the study of complex biologic systems (1-5). The aim of metabolomics is the study of concentrations and fluxes of low molecular weight metabolites present in the biofluids, tissues, or tissue extracts of living systems to retrieve information on biologic systems. As an emerging tool, metabolomics may be integrated with genomics, transcriptomics, and proteomics to achieve a comprehensive understanding about the physical condition of the system (6-8). Metabolomics improves that understanding by evaluating the metabolic changes that reflect the integrated responses of the system to internal or external perturbations. Also, compared with the numbers of modified proteins (>100,000) and genes (>30,000), measuring metabolites in biofluids or tissues is a much more direct and in principle simpler analytical problem to collect the useful information. The availability of detailed metabolic maps (9) can make interpretation of the results more straightforward.

Metabolomics is related closely to metabonomics, defined by Nicholson (10, 11), who pioneered several advances that combine NMR with pattern recognition techniques. Metabonomics emphasizes “the determination of systemic biochemical profiles and regulation of function in whole organisms by analyzing biofluids and tissues” (10). For clarification, definitions of common terms used in metabolomics research are summarized in Table 1 (10, 12-15). In this article, metabolomics is used as a general term.

The field of metabolomics has been developing very rapidly in part because of advances in the analytical methods of high-resolution nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) and their combination with multivariate statistical analyses (16-22). Additional methods, such as infrared spectroscopy (23) and electrochemical methods (24), also have been used in metabolomics studies to a lesser extent. Because of the excellent quantitative capabilities and high measurement reproducibility of NMR, 1H-NMR spectroscopy has proven to be highly useful and reliable in metabolomics studies. The positive identification of interesting metabolites has been extremely useful in several studies.
The application of metabolomics has been widespread in areas ranging from systems biology studies to drug discovery, pharmaceutical research, early disease detection, toxicology, food and nutrition science, and others (25-28). Recent clinical examples of this approach include promising studies of ovarian cancer and insulin resistance (29, 30), the latter having been detected by NMR for over 20 years (31, 32). NMR spectroscopy plays an especially important role in metabolomics by providing highly reliable and reproducible measurements of the biologic samples that include a variety of biofluids and tissue samples. In this review, we highlight recent developments in NMR-based metabolomics methods for biomarker discovery and describe the main methods of statistical analysis currently used. Finally, we describe applications of the technology in several areas including nutritional studies, toxicology, and early disease detection.

### Current NMR Technologies in Metabolomics

In a sense, high-resolution NMR has been used for metabolic profiling of biofluids and tissue extracts for many decades (12, 33, 34). It is well known that the complexity of biofluids or tissue extracts often contains hundreds of detectable metabolites and thus thousands of spectral lines gives rise to complicated NMR spectra. However, the high reproducibility and quantitative characteristics of NMR data are well matched to the requirements for multivariate statistical analysis (12), which can provide a method to reduce the spectral complexity. Progress also has been made to reduce the congestion of regular one-dimensional (1-D) 1H-NMR spectroscopy by means of the J-resolved experiment (35) and other two-dimensional (2-D) methods. In addition, the NMR sensitivity and resonant frequency do not depend on the pKa value or hydrophobicity of the sample, which makes it an excellent choice for broad-based analyses especially suitable for biological samples in different conditions, whereas the required sample preparation for standard NMR measurements is limited. With advanced high-throughput NMR methodology, 100-200 samples per day can be measured routinely with the assistance of flow-injection probes and automated liquid handlers. The detection limit is decreased to tens of nanograms by the use of microcoil probes (36). Nevertheless, the complexity inherent in biologic samples generates many peaks within a small chemical shift range (3-10 ppm) in the 1H-NMR spectrum, and therefore spectral overlap is still a problem, especially in the aliphatic region (1-5 ppm). Therefore, potentially important compounds present at smaller concentrations often are overshadowed by larger peaks and thus are less likely to be detected.

Typically in metabolomics, NMR measurements are carried out on biofluid samples such as urine, blood serum, or plasma, although the analysis of intact tissue samples also is gaining in popularity. Unless the amount of sample is limited, approximately 200-400 µL of the biofluid typically is used for a standard measurement. If some samples, buffer solution is added to stabilize the pH, followed by centrifugation to remove insoluble species. In addition, 10% deuterium oxide is mixed with the sample and used for frequency locking. Also, a compound that serves as the standard chemical shift reference (0 ppm) is added to allow the alignment of the spectra. An added benefit is that the peak area of the standard compound may be used.
1-D Water Saturation Pulse Sequences

In NMR-based metabolomics studies, especially when biofluid samples are used, the overwhelming signal of the water resonance must be suppressed by using special NMR pulse sequences to measure the useful metabolites. The simplest water saturation pulse sequence consists of a continuous low-power irradiation over the period of an acquisition delay, which is usually 1–2 s long. The irradiation is applied at the water frequency (~4.8 ppm) that is to be suppressed. This simple pulse sequence usually is referred to as presaturation or PRE-SAT, as shown in Figure 1a where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times. Other water suppression pulse sequences include WATERGATE (WATER suppression by gradient tailored excitation), the PURGE sequence, and excitation sculpting (39–42). For example, the new PRESAT with 1-D NOESY (nuclear Overhauser effect spectroscopy), as shown in Figure 1b where presaturation occurs during the relaxation and mixing times.

CPMG experiment

When large molecular weight species such as lipoproteins are present in the biofluid sample, a profile that contains broad peaks will be detected. The broad lines result from a combination of fast relaxation, multiple and unresolved couplings, and the overlap of similar but distinct molecular subtypes. In such cases, the CPMG (Carr-Purcell-Meiboom-Gill) spin echo pulse sequence is used to attenuate the broad signals that result from macromolecules according to their molecular and relaxation time, as shown in Figure 2. In particular, the number of spin echoes, n, and the delay time, τ, which determine the mixing time of the experiment, are critical for the suppression of broad protein signals. Ideally, these parameters should be properly optimized for each sample type (44). 1-D PRESAT and CPMG pulse sequences of the same human serum sample are shown in Figure 2. The effect of CPMG is evident when comparing the panels a and b, in which the broad signal has been suppressed effectively, especially in the aliphatic region (0.4–5.5 ppm). However, the suppression of broad signals by CPMG also may result in a loss of useful information. It is necessary, sometimes, to perform multiple experiments with and without CPMG when measuring blood serum or plasma samples.

2-D experiments

2-D NMR experiments improve the ability to interpret spectra because the peaks are dispersed extensively, which increases the resolution and information content. Several commonly used homonuclear 2-D experiments include correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY). A typical 1H–1H correlation spectrum provides information on spin–spin couplings between protons that are linked by covalent bonds. When hundreds of peaks appear in the 1-D 1H–1H NMR spectrum, 2-D spectroscopy provides a means to assist the identification of key metabolites.

Homonuclear experiments take advantage of the high sensitivity of the 100% natural isotopic abundance of 1H. Although heteronuclear (1H–13C) 2-D experiments are less common in metabolomics because of their lower sensitivity (caused by the 1% 13C nature abundance), several are useful in helping identify
unknown metabolites. In particular, HSQC (heteronuclear single quantum coherence) (45) is used in metabolomics to assist in metabolite identification (46). HSQC experiments, especially $^{13}C$–$^1H$ and $^{15}N$–$^1H$ HSQC, in which isotopic $^{13}C$ and $^{15}N$ labels are introduced biochemically are used frequently in protein NMR spectroscopy. In metabolomics research, when homonuclear spectroscopy is not adequate for structure elucidation, a $^1H$–$^{13}C$ HSQC spectrum often is acquired. A cross peak indicates the presence of an H–C bond. A typical HSQC spectrum of human urine sample is presented in Fig. 3. Another similar pulse sequence HMBC (heteronuclear multiple bond coherence) is used to detect longer-range couplings through two or three bonds.

An improvement in the resolution and sensitivity of 2-D experiments of complex samples will increase the ability to identify more metabolites of interest. A recent approach developed by Shanaiah et al. enhanced the sensitivity of $^1H$–$^{13}C$ HSQC for metabolomics studies (47). A certain class of metabolites, amino acids in this case, was derivatized with labeled 1,1$'$-$^{13}C_2$ acetic anhydride, which produced N-acylated products and increased the sensitivity by 100-fold. Cross peaks of these metabolites could be resolved in the HSQC spectra.

Pattern Recognition Techniques

Statistical pattern recognition methods (48–50), such as principal component analysis (PCA) and partial least squares (PLS), when combined with NMR measurements, create enormously powerful opportunities for metabolomics research. In particular, they provide important data reduction and simplification methods, as well as extensive capabilities for distinguishing very similar spectra. Most multivariate analyses used in the field of metabolomics are based on the projections of the original data set onto a new coordinate system. If a certain degree of similarity exists between rows (spectroscopic measurements) in the data matrix $X$, such projections in principle can predict the outcome of a new data set. The advantages of projection methods are:

1. The multivariate model is based on the complete data set.
2. It can tolerate a certain amount of missing data.
3. It provides visual representation of the classification within the data matrix $X$ (the score plot).
4. It assumes the existence of noise in the data.
5. In the case of PLS, it can interlink different data sets X and Y and generate results that easily can be interpreted (51).

Preprocessing
Before multivariate analysis, certain preprocessing steps have to be carried out for several reasons. The overall goal of preprocessing is to achieve a better discrimination and obtain the results with better interpretability. First, spectra are aligned using the chemical shift reference. Next, unwanted peaks are removed. In the case of 1H-NMR spectra, unwanted peaks include the residual water signal after suppression and the chemical shift reference peak. In the case of urine samples, the urea signal is broad and partially attenuated because of its fast isotopic exchange with the (saturated) protons in water and, thus, also is removed.

Normalization
For biofluid samples such as human urine, the concentrations of different metabolites are highly dependent on the amount of consumed water. Therefore, it is necessary to normalize the whole spectral region to the same level to eliminate the effect of the unwanted overall concentration variations. The most common way to normalize the data sets is to use total integrated intensity normalization, which simply sets the total intensity of the whole spectral region to a constant value. An alternative method is vector length normalization, which sums the square of each variable and sets it to a constant. In the view of multidimensional space, each spectrum can be represented by a vector. After normalization, the length of each vector has the same value. Another common method is to divide the whole spectrum by the peak intensity of a concentrated metabolite. For example, in the case of an 1H-NMR spectrum of urine, creatinine often is used by the medical community as an intensity reference. This use is because of the underlying assumption of constant excretion of creatinine into urine. Thus, creatinine can be used as an indicator of urine concentration.

Mean-Centering
Mean-centering is carried out by subtracting the mean value from each variable. In other words, an average spectrum is calculated from the whole data set and then subtracted from each spectrum. By the use of mean-centering, the center of the data distribution in the multidimensional space is set to the origin.

Scaling
Scaling is used to change the emphasis from metabolites with high concentrations to those with moderate or small concentrations. Mean-centering must be applied before performing any scaling process. Variance scaling, in which each variable is divided by its standard deviation (square root of variance) is the most commonly used. In this way, all variables are scaled to unit variance. The purpose of this operation is to decrease the dominant effect of large and variable NMR signals on statistical models and to allow the contribution of small peaks to be included. A combination of mean-centering and variance scaling is termed autoscaling. Variance scaling is recommended when the variables have different units. However, this situation is not the case for NMR measurements. Variance scaling may cause an undesired emphasis on noise. Another popular scaling method is Pareto scaling, which uses the square root of the standard deviation as the scaling parameter. If metabolites with moderate or small concentrations are important, then Pareto scaling tends to be useful as it is less dominated by noise.

Principal component analysis (PCA)
The goal of principal component analysis (PCA) is to eliminate the noise-associated variance and reduce the dimensionality.
of complex spectral data sets. In metabolomics research, the unsupervised method of PCA usually is the first multivariate analysis applied to the data to detect possible clusters and outliers in the sample set. PCA refers to a method that builds linear multivariate models of complex data sets. The PCA model is built on the basis of orthogonal vectors, in other words, eigenvectors. The central step of PCA is to solve for the eigenvectors and eigenvalues of the covariance or related correlation matrix. The covariance matrix X' calculated from the original data set X usually is computed using the following expression:

\[ \mathbf{X'} = \mathbf{X} \mathbf{X'} \]

The eigenvectors of X', or principal components, are calculated and then arranged in decreasing order according to their corresponding eigenvalues. Usually the first several principal components will explain most of the variance generated in the whole data set X. The rest of the PCL, which mostly contain noise, will form the residue matrix E. Thus, PCA acts as a powerful data reduction method.

The results of PCA can be written in the following matrix equation:

\[ \mathbf{X} = \mathbf{TP} + \mathbf{E} = \mathbf{t}_1 \mathbf{p}_1 + \mathbf{t}_2 \mathbf{p}_2 + \ldots + \mathbf{t}_k \mathbf{p}_k + \mathbf{E} \]

where the score matrix T contains all the projection scores of spectra in the data matrix X on the chosen PCs, and the loading matrix P is the ranked eigenvectors according to their corresponding eigenvalues. The scores \( t_1 \), \( t_2 \), which correspond to the first two PCs (\( p_1 \) and \( p_2 \)), usually are used to provide an overview of all data in the form of a scatter plot. The original data set is projected onto the PCA model for consistency and is performed by dividing the samples into subsets, generating a statistical model based on one subset, and testing or validating the model on the other subset. Next, new observations are projected into each PCA model and the residual distances are calculated to assign the class membership of each observation. The SIMCA model is constructed on a training set. Once the model is defined, a separate data set is used for testing its accuracy.

**Partial least squares (PLS)**

Another popular method of projection is partial least squares (PLS), which uses two data matrices X and Y. The two data matrices can be, for example, different spectroscopic measurements or different omics data from the same sample set. As a bilinear model, PLS fits data matrices X and Y and recasts these data as score plots, loading plots, and weight plots. Similar to PCA, the PLS score plot is generated by projection onto the new coordinate system. The only difference is that the selection of the axis is based on the regression of X against Y. Compared to PCA, the PLS model seeks to maximize the variances from both X and Y matrices.

In particular, when Y has only one variable that contains the group assignment of each sample or measurement, PLS always is referred to as PLS-DA (partial least squares-discriminant analysis). In this case, Y is called the “dummy matrix.” As a supervised method, PLS-DA modeling establishes the difference between preassigned sample groups. PLS-DA is performed to maximize the separation among classes and to identify distinguishing features or potential biomarkers.

For example, when building a PLS-DA model of male and female biofluid samples, the class membership of each sample is input into the Y matrix as a dummy variable (i.e., 0 = male; 1 = female). Corresponding loading plots (of LVs) largely reflect the effects that result from gender (52). Typically, models based on supervised methods such as PLS-DA require validation that uses an additional sample set because of the ease and danger of overfitting the initial trial sample data. Careful and rigorous validation is necessary to avoid the common problems related to overfitting that seem to be too prevalent in supervised statistical studies in metabolomics.

**Soft independent modeling of class analogy (SIMCA)**

Introduced by Wold (53), soft independent modeling of class analogy (SIMCA) first uses PCA, which is performed individually on each class. The PCA scores and loadings thus are used to define each class, with the number of principal components usually being preselected or decided by a cross-validation process (54). Cross validation involves the testing of a data set for consistency and is performed by dividing the samples into subsets, generating a statistical model based on one subset, and testing or validating the model on the other subset. Next, new observations are projected into each PCA model and the residual distances are calculated to assign the class membership of each observation. The SIMCA model is constructed on a training set. Once the model is defined, a separate data set is used for testing its accuracy.

**kth nearest-neighbor analysis (KNN)**

kth nearest-neighbor analysis (KNN) is a method of classification based on the similarity within classes. Each spectrum can be treated as a point in a multidimensional space. In KNN, the Euclidean distance between every pair of spectra is calculated first. The class assignment of one sample is based on the majority vote of its nearest neighbors. K is a set of integers starting from one. The best choice of K also can be chosen by cross validation.

Beckonert et al. (55) performed a toxicologic metabolomics study in which they applied KNN for classification and prediction. All the urine samples collected from the subjects were measured using 1-D NMR spectroscopy and subjected to KNN analysis. Leave-one-out cross validation was used, and the same number of samples were chosen in the training and test sets. As the results show, KNN analysis resulted in a good predictability of liver and kidney toxicity (55).

**Orthogonal signal correction (OSC)**

The orthogonal signal correction (OSC) filter was developed to remove variations in the data matrix X that are uncorrelated to their corresponding eigenvalues. Usually the first several principal components will explain most of the variance generated in the whole data set X. The rest of the PCL, which mostly contain noise, will form the residue matrix E. Thus, PCA acts as a powerful data reduction method.

The eigenvectors of X', or principal components, are calculated and then arranged in decreasing order according to their corresponding eigenvalues. The scores \( t_1 \), \( t_2 \), which correspond to the first two PCs (\( p_1 \) and \( p_2 \)), usually are used to provide an overview of all data in the form of a scatter plot. The original data set is projected onto the PCA model for consistency and is performed by dividing the samples into subsets, generating a statistical model based on one subset, and testing or validating the model on the other subset. Next, new observations are projected into each PCA model and the residual distances are calculated to assign the class membership of each observation. The SIMCA model is constructed on a training set. Once the model is defined, a separate data set is used for testing its accuracy.
NMR-Based Metabolomics Technology in Biomarker Research

The Y data matrix (56). Especially when Y is a dummy matrix, the OSC filter is extremely useful for removing X variations that are orthogonal to class membership. Combined with PCA or PLS-DA, unwanted variation thus can be separated from signals of interest. For example, by applying OSC-coupled PCA and PLS-DA, circadian-related metabolic variation can be removed from the spectral data in the study of strain-related metabolic differences between two phenotypically different mouse strains (57). As with PLS, the use of OSC needs to be rigorously cross-validated, preferably with a new data set to avoid problems associated with overfitting.

**Statistics-assisted NMR spectroscopy**

In addition to the differentiation of sample subgroups, statistics also may assist in the improvement of the ability to detect a biomarker in NMR spectroscopy. Statistical total correlation spectroscopy (STOCSY) (58) takes advantage of the inherent colinearity of multiple peaks from a compound in a set of NMR spectra. Based on correlation alone, the identity of a metabolite with multiple correlated peaks can be determined. Statistical heterospectroscopy (SHY) (59) uses the Pearson correlation to combine spectroscopic measurements from NMR and MS on the same sample set and, thus, additionally improves the efficiency of the molecular biomarker identification. It also is possible to identify compounds metabolically related to known biomarkers using SHY, and this works across spectroscopic platforms (60).

**Advanced NMR Technologies**

New NMR methods have focused on simplifying and improving the interpretability of complex spectra for metabolomics studies. For example, a relatively simple approach such as liquid-liquid extraction to separate biologic samples into hydrophilic and hydrophobic fractions can improve the resolution (61). As an alternative approach, the selective-TOCSY experiment can focus the analysis on certain target molecules with much improved resolution and sensitivity compared with the normal 1-D NMR. A rather important technique, projected j-resolved spectroscopy, can be used to simplify the 1-D NMR spectrum by removing j-coupling effects (35). These two techniques are summarized in the following sections.

---

**Figure 4** PCA results of human urine samples from 20 healthy controls and 6 patients with IEMs: (a) the score plot and (b) the loading plot. When analyzed individually, well-known biomarkers for the IEMs are more apparent in the loading plots.
2-D J-resolved spectroscopy

J-resolved spectroscopy (J-Res) separates the chemical shifts and spin–spin couplings (J-couplings) along two (nontopolog- nal) frequency axes. Compared with a normal 1H-NMR mea- surement, the acquisition of J-Res spectra is time-consuming because of the 2-D data set, and typically requires 20-minute acquisition times or more (62). The projection of the J-Res spec- trum onto the chemical shift axis yields a proton-decoupled 1H-NMR spectrum. This projection can simplify greatly the spectrum of biofluid samples, especially in the region where peaks are highly overlapped, because J-coupling causes a high degree of additional peaks in the 1-D 1H spectrum. J-Res spec- troscopy therefore improves the spectral dispersion and thus is potentially useful when measuring complex biofluid samples in metabolomics (35). Figure 5 clearly shows the difference be- tween regular 1-D NMR and the projection of the J-Res (P-Jres) spectrum. For both types of measurements, PCA score plots show a distinct classification. However, the corresponding load- ing of the P-Jres data is much easier to interpret because of its simplicity. The only drawbacks of this approach are the long acquisition times and a loss in direct quantification compared with the 1-D spectrum.

Selective-TOCSY

Another promising NMR technique for metabolomics is selective-TOCSY (total correlation spectroscopy) (63). Selective-TOCSY is a 1-D pulse sequence that uses homonu- clear Hartmann-Hahn matching to transfer magnetization between protons through J-coupling (covalent bonds). In a typical 2-D TOCSY experiment, cross peaks can be observed among all J-coupled members within a molecule. Magnetization is transferred successively over up to seven bonds, as long as successive protons are coupled. The number of transfer steps can be adjusted by changing the mixing time and can be tuned from 1-2 bonds up to 5-7 bonds. A one-step transfer would give a spectrum very similar to a COSY spectrum. In the 1-D selective-TOCSY, an isolated proton usually is irradiated by using a selective pulse (shaped soft pulse) and magnetization is transferred through J-coupling from that 1H spin to the others in the molecule during the mixing time. All resonances from the same coupling system then may be observed after the detection period. The selective-TOCSY technique is highly useful in isolating certain components in the NMR spectrum of a complex mixture (63). Sandusky (64, 65) applied selective-TOCSY in a metabo- lomics study and showed that the classification of samples could be improved significantly. When differen- tiating sample groups using unsupervised PCA, species that have larger concentrations usually dominate the statistics. However, low concentration species also may be important. Selective-TOCSY may be used to emphasize such com- pounds by selectively irradiating their peaks without being overwhelmed by more concentrated, overlapping species. The effectiveness of selective-TOCSY is demonstrated in Fig. 6. It should be emphasized that the intensities of the selective-TOCSY peaks are affected greatly by the mixing time. However, if calibrated, the results can be quantitative, and the selective-TOCSY signals are proportional to the concentration. In summary, selective-TOCSY can be used to emphasize important metabolites, which may appear as low-concentration species, to achieve a better discrimination in metabolomics (64).

Applications in Metabolomics

A growing number of exciting applications for NMR-based metabolomics exists. A few areas of interest chosen from many are presented below to indicate the breadth of opportunities developing in the field.

Nutrition

The integration of metabolomics and nutritional science is one of the most natural applications of metabolomics. It is clear that metabolic variations in biofluids can result largely from dietary effects. By measuring the varying chemical compositions of biofluid samples, dietary factors that may contribute to the metabolic changes thus can be understood and controlled. For example, the metabolomics study of the effects of dietary supplements can provide useful information on the metabolic changes involved. A recent study compared the effects of black and green tea consumption on human metabolism as detected by NMR-based metabolomics. Urinary excretion of hippuric acid and 1,3-dihydroxyphenyl-2-O-sulfate was studied (66), and the presence of various amines in individuals who ate fish was observed also.

Dietary changes can be observed rapidly by using NMR-based metabolomics, as shown in a recent paper by Gu et al. (67) that focused on the study of rats by both NMR and MS approaches. Moreover, when developing models for disease detection, biomarkers must be distinguished from metabolic variations because of diet. For this objective, one approach is to identify certain metabolic pathways that are connected to dietary changes and use this information to distinguish these effects (67). Given the large changes observed, NMR-based metabolomics can play an important role in evaluating dietary effects in biofluid samples. To explore this issue, a human study was carried out by obtaining urine, plasma, and saliva samples from 30 healthy volunteers whose food intake was controlled and recorded during the study (68). Samples were analyzed us- ing 1H-NMR spectroscopy followed by PCA and PLSDA. It was found by PCA that interindividual variation in urine may be reduced by restricting dietary intake. However, this finding was not observed when measuring plasma or saliva. Moreover, the PLSDA score plot highlighted the metabolic changes over time, which indicates that the changes were largely because of creatinine in urine and acetate in saliva. Several possible nutri- tional studies can be contemplated, especially given the current and growing importance of studies that link nutrition with health and disease. However, to date, NMR-based metabolomics stud- ies have yet to exploit this rich research area fully.
Figure 5 ¹H–NMR data of medaka embryo extract: (a) 1-D spectrum and (b) the 1-D projection (P-Jres) of (c) a 2-D J-Res spectrum, (d) the pre-processed p-J-Res spectrum. (Reproduced with permission from Reference 35. Copyright 2003 Elsevier.)

Toxicology

NMR-based metabolomics has been applied very successfully to the study of toxicology (4, 11–13, 16, 50, 55, 69, 70), and many early metabolomics methods were developed for applications in this area (4, 71–73). The goal of toxicological studies is to find the organs that are affected by the toxin, the magnitude of the effect, and the mechanism by which the toxin alters the living system or its biochemical processes. When a living system is perturbed by certain toxic compounds, the altered concentrations of endogenous metabolites in biofluids can be measured by NMR spectroscopy and emphasized by multivariate analysis.

Typically, NMR-based metabolomics studies of toxicology are carried out on biofluids, especially urine, acquired typically from rodents. In a prototypical preclinical toxicity study, Han–Wistar (HW) and Sprague–Dawley (SD) rats were dosed with either hydrazine or HgCl₂ (74). Urine samples from control and dosed rats were collected and prepared for NMR measurements under the same protocol. PCA and SIMCA were used to evaluate the ¹H–NMR spectral data. Clear classifications in the PCA score plots were achieved between the dosed and control groups, as shown in Fig. 7. Biomarkers corresponding to each toxin were elucidated from the PCA loadings. It was found that several organic acids, amino acids, and sugars were altered in the case of mercury dosing, whereas taurine, β-alanine, creatine,
L-proline and 10 mM L-arginine with selective excitation on the proline (1.98 ppm). (c) Selective-TOCSY spectrum for a mixture of 10 mM defined: control, clear toxicologic lesion, and marginal lesion; as research subjects. For each study, three subclasses were pre-
responses to 147 toxins and treatments. Rats and mice were used
a database that would contain the NMR spectra of rodent re-
nitional collaboration between research institutes and pharmaceu-
tical companies (75). The objective of this study was to generate
the purpose of prediction. Using the model from the first three
doses, respectively. Several multivariate analyses were applied
to construct chemometric models. Although it was not possible
to define the toxic response for each of the toxins uniquely,
this database provides a comprehensive approach for the ex-
ploration of biomarkers in toxicologic screening and provides a
wealth of data on the metabolic toxic responses. More studies in
the area of toxicogenomics also were suggested by the project
coordinators.
As a result of the early metabolomics toxicology studies,
several methods were developed that now are used widely by
researchers in the field, including normalization (76), trajec-
tories, correlation methods, and so forth. (59, 77) NMR-based
metabolomics study of toxicology remains a vibrant area of
research.

Disease detection
NMR-based metabolomics also shows promise for detecting
molecular biomarkers of disease and for understanding the
biochemical disease processes. Although several studies have
identified interesting molecules, these putative biomarkers need
to be validated on a large sample size (many hundreds to
thousands of samples) before they can be used for clinical
diagnosis.
One promising application is the detection of metabolic dis-
eases such as a variety of inborn errors of metabolism (IEM).
Current standards of care include the routine screening of
newborn infants for 20–40 of the most common IEMs, typ-
ically using gas-chromatography with mass spectrometry de-
tection. However, borderline and undiagnosed cases require
the development of alternative approaches. It is hoped that
metabolomics-based approaches may improve the detection of
such difficult cases. NMR has been used to diagnose IEMs for
many years (31, 32), but the application of metabolomics-based
approaches may provide additional diagnostic power. Recently,
such a study has been carried out by using blood spot sam-
pies for the detection of phenylketonuria (PKU) and maple
syrup urine disease (MSUD) (44). Samples from patients and
healthy controls were measured by 1-D CPMG NMR. PCA
was performed to discriminate healthy and disease samples.
A similar study involved the diagnosis of phenylketonuria (PKU)
and maple syrup urine disease (MSUD) (44). Samples from patients and
healthy controls were measured by 1-D CPMG NMR. PCA
was performed to discriminate healthy and disease samples.

Figure 6 (a) 1-D proton spectrum of a 7-fold dilute honey acquired using
an 1-D NOESY sequence with presaturation for water suppression during
the relaxation delay and mixing time. (b) Selective-TOCSY spectrum of a
7-fold dilute honey with selective excitation on the proline γ peak
(1.98 ppm). (c) Selective-TOCSY spectrum for a mixture of 10 mM
L-proline and 10 mM arginine with selective excitation on the proline γ
peak (1.98 ppm). (Reproduced with permission from Reference 64.
Copyright 2005 American Chemical Society.)

4.5 4.0 3.5 3.0 2.5 2.0 1.5 ppm

PROLINE
Serum samples were used for the 1-D 1H-NMR measurements. The PCA score plot could classify all serum specimens from the 38 patients with EOC from all of the normal samples and from all the sera from patients with benign ovarian disease. Also, it was possible to separate 37 of 38 (97.4%) cancer specimens from 31 of 32 (97%) postmenopausal control sera. SIMCA was used to provide a validation for the classification. Particularly, by applying ROC (receiver operating characteristic) analysis, it was found that the lipid signals at 2.77 and 2.04 ppm and 3-hydroxybutyrate were altered in the EOC samples, although the two lipid signals were not associated with particular metabolites.

Disease detection using NMR-based metabolomics currently is extremely promising and potentially very useful. Many diseases such as cancers, type II diabetes, and heart diseases cannot be detected easily in their earliest stages. Many validated biomarkers such as CA125, a protein that is found to be elevated in some cancers, are not very sensitive or specific. Along with the development of better protein-based tests, metabolomics provides promising, noninvasive, and cost-effective approach with competitive accuracies. However, metabolomics seems to provide a validation for the classification. Particularly, by applying ROC (receiver operating characteristic) analysis, it was found that the lipid signals at 2.77 and 2.04 ppm and 3-hydroxybutyrate were altered in the EOC samples, although the two lipid signals were not associated with particular metabolites.

For the future, more advanced NMR and statistical methods are likely to appear. In addition, the combination of highly sensitive mass spectrometry with NMR spectroscopy in various ways provides several new avenues for development (59, 60, 78). Both NMR and MS provide unique metabolic profiles and can be combined to explore novel biomarkers that may not be identified easily with a single method. In the longer run, this combination of methods may show more potential to correlate metabolomics extensively with proteomics and genomics, although this problem has been an overwhelmingly challenging problem to date. In summary, NMR-based metabolomics likely will provide an extremely useful approach for the investigation of key problems in systems biology.

Summary

The field of metabolomics is growing rapidly and attracting increasing interest. NMR-based metabolomics, sometimes coupled with other analytical methods, has broad appeal and a promising future for many applications. The common objective in metabolomics is the selection of a set of reliable, key metabolites that can be used to identify and follow changes in biologic systems. In particular, NMR spectroscopy provides quantitative and reproducible measurements of the metabolic profile, whereas multivariate analyses are used to select important metabolites. For these purposes, more sensitive, reliable, and high-throughput analytical tools to perform metabolomics studies are highly desirable. New statistical tools, such as advanced multivariate analyses, will likely be developed and applied to different types of sample sets.

For the future, more advanced NMR and statistical methods are likely to appear. In addition, the combination of highly sensitive mass spectrometry with NMR spectroscopy in various ways provides several new avenues for development (59, 60, 78). Both NMR and MS provide unique metabolic profiles and can be combined to explore novel biomarkers that may not be identified easily with a single method. In the longer run, this combination of methods may show more potential to correlate metabolomics extensively with proteomics and genomics, although this problem has been an overwhelmingly challenging problem to date. In summary, NMR-based metabolomics likely will provide an extremely useful approach for the investigation of key problems in systems biology.

References


44. M A H, Keyes D. Pre-SAT180, a simple and effective faraway water suppression method. In press.


Combinatorial Libraries: Overview of Applications in Chemical Biology

H. Peter Nestler, Combinatorial Technologies Center, Tucson, Arizona
doi: 10.1002/9780470048672.wecb405

The sequencing of genomes gave access to the complete set of building blocks for organisms of various species. A plethora of "-omics" technologies has been developed to investigate the dynamic interactions of the building blocks to understand the functioning of living organisms. Synthetic molecules have proven to be powerful tools to modulate that states of biological systems, but the challenges to find suitable probes are tremendous. Combinatorial libraries allow preparing and testing diverse sets of molecules with high efficiency. In this article, we will discuss how combinatorial chemistry enables to investigate and modulate biochemical function in the quest to chart chemical and biological spaces.

The scholar disciplines of chemistry and biology have for a long time lived in separate academic universes, although they heavily depend on each other. Combinatorial chemistry started to bridge this intellectual gap by introducing the concept of diversity and selection into synthesis, which mimics biological evolution. It was triggered by the frustration of chemists trying to study biological systems with insufficient chemical matter. The recognition that chemical structures cannot be tailor-made to interact with biological space stimulated the development of methods for fast synthesis and screening of small molecules. Albeit combinatorial chemistry was recognized in its infancy as one basic pillar of the proteome into target families, assuming that members of the same protein family share structural features. This assumption holds true especially for enzymes and GPCRs but less though for proteases and ion channels, which are the four most common target families in drug discovery. Other protein families such as nuclear hormone receptors and transcription factors display even higher structural variations. However, chemical biology is also founded on the understanding of the biological function of the protein families, information about assaying as well as privileged ligand motifs that can be transferred from one target family member to another. Although combinatorial chemistry and chemical biology may seem to be opposite concepts, in reality they are highly complementary and interacting. Using the insights gained from chemical biology to bias the design of combinatorial libraries will help to assess a chemical space suitable to probe biological systems. In this article, we will discuss the roots of combinatorial chemistry, the technologies for generating and assaying compound collections, and studies performed on biochemical systems.

Chemical Biology—One Short Definition

Chemical biology covers many aspects of the influence of chemical molecules on biological function. In a broad definition, the understanding and investigation of toxicological effects, modulation of gene and protein expression, transformation and transport of the molecules in cell and whole organisms, and the change of the metabolic pathways and patterns are topics of chemical biology. With the definition below, we put a focus on one subtopic of chemical biology pertinent to the discussions in this article.

The sequencing of the human genome (1) marked the apex in the transition of biology from an observational and descriptive activity to a hypothesis-driven science. With the information about complete sets of building blocks for cells available, methods could be devised to group proteins on a structural level and to investigate the phenomenology of organisms on a molecular and structural level. Similarities in protein structures have been investigated for a long time, which cover all levels from primary (sequence) via secondary (domain folds) to tertiary (overall three-dimensional) structures. Investigations of tertiary structures help to predict functional sites and roles for uninvestigated proteins from their primary sequence. Especially through bioinformatic analysis, it has been possible to...
identify homologous reaction mechanisms even within proteins with lower sequence similarity and different biochemical functionality (2, 3), as highlighted by the cases of Leukotriene A4 hydrolase and angiotensin-converting enzyme (a zinc metalloprotease), which are both inhibited by bestatin but have distinct biological roles (4). Primary structure investigations have been preferentially of interest for evolutionary analyses. These phylogenetic analyses have been crucial in defining target families, which are groups of proteins that have a similar gene and therefore protein sequence, which lays the foundation for "chemical biology." Kinases are the prototypic target family as their active sites are structurally highly homogenous and bind the same co-substrate, adenosine triphosphate (ATP). Other gene families include G-protein coupled receptors (GPCR), ion channels and transporters, and proteases, although the structural diversity in these families is higher and they therefore group in structurally and mechanistically diverse sub-families, such as cysteine or metalloproteases.

Soon after the sequences became available, researchers discussed how many of the approximately 27,000 genes that had been assigned (1) would be "druggable" (i.e., their associated protein products could be modulated with small molecules in a directed fashion to achieve a desired physiological effect (5). For a small molecule to exert its biological action, it is crucial that the molecular shape of the molecule complements the cast offered by the target protein. This fact has been recognized first by Emil Fischer who phrased it as a "Key-Lock" principle (6), being unaware of the dynamic and flexible nature of protein structures. Today, we examine the interactions of two molecules more in a "Hand-Glove" fashion with strong elelements of induced fit (7). We term the ensemble of available interaction shapes in the proteome the "biological space," whereas the "chemical space" is considered the ensemble of shapes offered by small molecules. With our structural understanding constantly evolving through molecular biology and crystallography, the efforts to design matching chemical structures rationally have increased and led to successes in drug development; the HIV-protease inhibitors are one prime example. Rational design depends on structure-activity relationship to bias the contribution of various molecular motifs to the intermolecular interactions properly; therefore, it is powerful for the optimization and understanding of activity and selectivity on the protein target level. Rational design suffers shortcomings when we attempt to address the challenges of finding novel starting structures from scratch. High-throughput screening efforts try to tackle this challenge through engaging in a high-numbers trial-and-error game. As the screening collections reflect the target history of the respective company, they often cover narrow aspects of chemical space. This challenge is addressed by combinatorial chemistry.

Considering that most novel therapies would rely on oral administration of drugs, such molecules have to fulfill requirements to achieve suitable pharmacokinetic behavior. The most quoted and commonly used guidelines are Lipinski’s “Rule-of-5” (8) and Veber’s “rotational bonds” (9) that have been based on a statistical analysis of marketed oral drugs. Considering such boundaries, it has been estimated that about 10–15% of the human genome would be "druggable" (10). Although this number may seem low, it should be considered that only one third of these mechanisms are currently targeted, and a significant fraction of drugs, even those in development, still act through undefined molecular pathways. Furthermore, the hype around the sequencing of the genome and the assumed impact on drug discovery meanwhile has vanished, as it was recognized that biological networks are too complex and redundant to allow control through one molecular dial. Systems biology tries to address this challenge by exploring the interactions of proteins and the resulting pathways of transferring biological signals and actions. Chemical biology is the matching component in drug discovery that tries to synergize on structural relationships of proteins to address the druggable genome efficiently (11). Furthermore, the concepts of chemical biology are transferred towards the understanding of pharmacokinetics by identifying and investigating proteins that are involved in transport through the organism and the metabolic transformation of the molecules. Models to predict unwanted effects (commonly referred to as "toxicity" or "side effects") are developed based on our increased understanding of biological pathways and the proteins involved therein that may offer interaction sites for the molecules. As those investigations do not involve combinatorial chemistry approaches, they will not be topic of this article.

Chemical biology has reshaped all stages of the drug discovery and today is a widely used discovery paradigm in pharmaceutical industry. The focus as well as the impact of using targeted family knowledge has been on the early stages from target identification via structural understanding through lead finding efforts. The later stage of the drug-discovery process, which included the optimization of lead compounds into drug candidates, is not as amenable to technological solutions that can be provided through target family concepts as the challenges become very specific for each lead series. Still, transferring insights and understanding compound interactions with targets and other proteins help to avoid entering dead-end alloys of modification. However, the focus of this article will be on the biochemical level of chemical biology. Most chemical examples will be of peptidic nature, as most reports around the topic of this article are based on libraries derived from peptides, although this may change in the near future. The main theme of this article will be more on the general concepts and approaches than on specific molecules, but the interested reader can find a bounty of literature on specific molecules that modulate biological systems especially from a viewpoint of drug discovery.

**Combinatorial Libraries—A Brief Overview**

Historically, the art of chemical synthesis demanded that individual molecules are prepared efficiently with high yield and are extensively characterized. Over decades, more and more complex molecules were prepared by dissecting the target molecule at key reactive bonds, devising strategic options for the assembly, and developing novel reactions to address the synthetic needs. Although this "logic of synthesis" laid the foundation for
Combinatorial Libraries: Overview of Applications in Chemical Biology

Today’s capabilities to prepare a wide variety of conceivable target structures, it did not provide the numbers of molecules necessary for and efficient investigation of the “biological space.” The resolution of this conundrum required a conceptual rethinking by chemists. Assuming that our hypotheses about evolution are correct, the natural evolution of complex organisms occurred through generation of molecular diversity and selection and amplification of suitable structures. Today, we observe this process at a level of high structural complexity; genetic selection of alleles and somatic selection of antibodies are among the most prominent aspects. Thus, it was not too surprising that the concept of combinatorial chemistry was devised based on immunological challenges (12). To be successful in the game of evolution and life, four aspects have to be addressed: 1) generation of diversity, 2) compartmentalization of the individual members, 3) selection for desired properties, and 4) reproduction of selected members. The third aspect will be the topic of other articles, whereas the other aspects will be discussed here. We will not review the variety of chemistry possible for the preparation of libraries, which has tremendously developed over the last decade and has been comprehensively reviewed on a regular basis by Dröffe et al. (13).

**Synthesis of combinatorial libraries**

Biology handles diverse gene and protein populations by compartmentalizing the members in cells. The selected library members are amplified by growing clones of the cells that survive the selection step, which allows ultimately for the reporting of the structures for selected members through gene sequencing. The “yeast-two-hybrid” (14) or the “phage display” technique (15) are used to study protein/protein and protein/peptide interactions in vivo, and the “SELEX” technology allows for the in vitro evolution of short oligonucleotides, which are called aptamers, that interact with small molecules and proteins (16). Such biological libraries will not be topic of this contribution.

Synthetic molecules by amplified by growing clones, and chemists who handle libraries face two major problems: the compartmentalization of the individual library members and the determination of the molecular structure of the selected species (17). Three approaches have been devised to tackle these challenges:

1. **Spatial arraying of compounds during synthesis and assay**
2. **Synthesis of mixture libraries and deconvolution of active sublibraries**
3. **Encoding of particles carrying individual compounds to be tested**

Spatial arraying has been used traditionally for the assaying of historic compound collections. As the identity of the molecule is correlated to position in the array, this approach carries the least challenge for structural assignment for active array positions. As an additional aspect to ease the implementation of arrayed libraries, arrayed libraries are commonly handled as solutions. A most biological assays are based on homogeneous test systems, providing the libraries in solution is the obvious choice. Today the preparation of arrayed libraries in solution is mostly addressed through robotic means (which basically provides high-throughput parallel syntheses), but significant approaches are available to provide arrayed libraries by synthesis at predetermined locations of solid supports. The earliest experiments used functionalized polystyrene mounted on plastic rods, which were arranged in a 96-well microtiter plate format and were used for the synthesis of the libraries and the compounds were released into the wells for their biological testing (18). Various implementations of this strategy have been developed, such as spot synthesis on cellulose membranes (19, 20) or photolithographic wafer techniques (21). However, these techniques impose limitations on the numbers of compounds that can be prepared and usually the 96-well format is used to test sublibraries and mixtures of compounds. Instead, DNA and RNA libraries prepared through photolithographic technology are used more widely used in genomic diagnostics.) To ameliorate the restrictions caused by the spatial addressing, Berk and Chapman (22) devised a strategy for the generation of two-dimensional spatial arrays of compounds that allow for determination of synergistic effects of residues in the same molecules, which combines the advantages of solution-based biological assays with an efficient testing of large numbers of compounds.

Several groups have reported the generation of combinatorial libraries in solution (23–26) as well as the fractionation of complex libraries by HPLC (27). Although the handling of compounds and libraries in solution makes them more amenable to the formats used for biological tests, the task of elucidating the structures of the active compounds requires complex deconvolution through sublibraries. This task can be performed by simultaneous preparation of the required sublibraries that control residues in one or two variable positions, which is called “positional scanning,” or by subsequential narrowing of the options by repeated testing and resynthesis of the most active sublibraries with additional defined positions. Both approaches are based on the hypothesis that the most active compound in a library is defined by the combination of the most active moieties in the various positions. In that respect, the deconvolution approach is conceptually related to fragment-based approaches, that will be discussed later. Despite all doubts about the reliability of results that stem from the testing of complex mixtures, there have been successes to the deconvolution strategy, which have been reviewed by Houghton et al. (28). Most recently, the mixture screening and deconvolution was applied in vivo (29).

To avoid this caveat of screening mixtures and the identification of active compounds through deconvolution, approaches to compartmentalize the library on moieties of solid supports have been devised. Houghton (30) suggested the use of labeled “teabags” that contain peptide synthesis resin for the compartmentalization and tracking of the library members (30). This approach was modified by using radio chips for the labeling of the “tea bags” (31, 32), which allow the robotic sorting of the containers for synthesis and spatial arraying. Yet, the size of the bags still puts limits on the diversity of such combinatorial libraries. "Split synthesis" (33, 34) lifts the restrictions for the complexity of libraries by creating “one-bead-one-structure libraries” that display the individual library members separated on beads of a solid support. Successive cycles of distributing and
mixing portions of the support lead to a combinatorial increase of the diversity of structures contained in the library. As each bead of the resin reacts with only one set of reagents per synthesis cycle, each bead carries one individual structure, which is determined by the sequence of reactions that occurred during the synthesis. Library members that interact with the targets can be isolated by selecting beads detected in assays performed on the beads. However, the application was limited to peptides and oligonucleotides because of the analytical means of determining the structure of the molecules on the beads. Molecular encoding was the breakthrough that opened the way for the application of the whole repertoire of synthetic chemistry for generating combinatorial libraries by solving the problem of structure determination from a single bead: Easily detectable molecular tags are attached to the resin beads as they proceed through the split synthesis during the library construction, which thereby records and reports the reaction history of each individual bead (35) (corresponding to the genetic coding of proteins in cells). After screening the library, these molecular tags are cleaved from each selected bead and then analyzed to report the structures of the library members on these beads. Several encoding methods are currently in use, which include microsequencable oligonucleotide (36) or oligopeptide strands (37, 38), as well as nonoligomeric schemes that employ small molecules that can be analyzed by gas chromatography (35, 39), HPLC (40), or mass spectrometry (41).

Fragment screenings and dynamic combinatorial libraries

In recent years, two variations of combinatorial chemistry approaches have been explored that do not present preassembled molecules to the biological target (42). Both variations make use of the detection of interactions instead of using biochemical inhibition as a readout. One approach, "fragment screening," uses fragments/building blocks of the molecules and relies on the detection of low-affinity interactions. Several detection approaches have been described, based on NMR (43), cocrystallization (44), or surface plasmon resonance (45, 46). The second approach, which is called "dynamic combinatorial chemistry," relies on forming the complete molecules from fragment in presence of the target molecule and enriching molecules with high affinity in the equilibrium reaction (47). This approach has been applied to generate artificial receptors for peptides (48, 49) as well as inhibitors for enzymes with detection by cocrystallization (50).

Mapping Enzymatic Activities

One key aspect of applications of combinatorial libraries in chemical biology is the mapping of substrate specificities for enzymes. Although the early studies that use peptide libraries were focused on studying antibody epitopes (12), the substrate mapping for enzymes that transform peptides and the development of molecular probes found widespread interest.

Proteases

The earliest applications focused on proteases. The major challenge that had to be tackled was the generation of a read-out signal. Although most protease assays rely on generating a chromophoric or fluorophoric molecule on cleavage from the carboxy-terminal amino acid (e.g., nitrophenyl, coumarins), this method was not applicable to the screening of peptide libraries as the cleavage does not necessarily occur at the carboxy-terminus of the sequence. The problem was resolved through the introduction of fluorophor-quencher pairs at the ends of the peptide sequence. As long as the peptide is intact, the quencher at the distal end of the peptide suppresses the fluorescence. After cleavage, the quencher diffuses away from the support, and the support becomes fluorescent because of the fluorophor at the proximal end of the peptide (51, 52).

Another challenge that had to be addressed for the screening of proteases was the penetration of the supports and its handling in aqueous environment during the assays. The classic supports used in peptide synthesis show a limited amount of swelling and penetration by proteins in aqueous environment, which makes only peptides on the surface accessible for study (this behavior was used to develop a molecular encoding approach by discriminating surface and interior positions of the bead (53)). Grafting of hydrophilic chains of polyethylene-glycol and preparation of supports based on cross-linked acrylamide yielded supports that could be handled in organic solvents for synthesis and water for assaying (54). With all these different approaches in hand, substrates for all protease subfamilies have been identified from combinatorial libraries (55–59), but the application of the data has been extended the most for cathepsins (60). Based on the sequence information obtained, Greenbaum et al. (61) designed inhibitors by incorporating the epoxide motif from E-64, which is a well-known unspecific and irreversible cysteine protease inhibitor. These inhibitors allowed pulling down various cathepsins from cell extracts or labeling them with some selectively using fluorescently labeled inhibitors, which thus fingerprinted cathepsin activity in cells. The surprising observation in these studies was that specificity could be instilled into the inhibitors starting from the unspecific E-64 by attaching it to the various substrate sequences. The latest step forward was the demonstration that these substrate based inhibitors can be used to monitor cathepsin activity in living mice. The inhibitors were generated using an acyloxymethyl ketone-inhibitor motif, which acts as a suicide substrate. The two sides of the substrate were labeled with infrared fluorescent dyes and a quencher and were injected into the tail veins of mice xenografted with tumors that express high levels of cathepsins B and L. The inhibitor became fluorescent on cleavage, and the labeled cathepsins allowed a distinct imaging of the tumors (62). The use of covalent inhibitors offers the advantage of localizing the fluorescent signal to the locus of generation as the diffusion of the signal is limited by the slow diffusion of the cathepsins. Recently, a new approach based on substrates was described, which will lead to the ability for fast real-time monitoring of proteases in vivo (63).
The screening of combinatorial libraries for kinases faced another challenge. Whereas kinases allow the straightforward generation of a readout signal on peptides through the transfer of radioactive (32P)-ATP, the association of the radioactive signal to individual members of highly diverse libraries posed a problem. Autoradiography using photographic film was used to visualize groups of beads immobilized in agarose gels, and individual beads were isolated through repetitive steps of isolation of beads from the gels and dilution, which made the process work intensive (64, 65). The process was dramatically simplified by embedding the beads directly into the photographic emulsion, which allowed for the microscopic identification of individual radioactive beads (66, 67). The other approaches for library preparation and screening like positional scanning (68) or arrayed synthesis (69) were successfully applied as well.

For the main purpose of the kinase chemical biology of routine profiling and substrate identification in a high-throughput mode, the array synthesis seems to be established because of the ease of readout. Recently, a novel approach lifted the need of using radioactive ATP for the screening for tyrosine-kinase substrates. The library of fluorescein-labeled peptides is encoded with PNA tags and incubated with a kinase. The peptides are then incubated with a DNA-microarray that complements the PNA strands of the individual peptides in solution. After such attachment of the peptides to the array, the array is treated with a phosphotyrosine-recognizing antibody sandwich, which carries the fluorescent dye Cy3. The microarray is then read out at two wavelengths to allow the detection of fluorescein and Cy3. The quantitative determination allows for a sensitive calibration against PNA/DNA complexation differences and the identification of kinase substrates (70).

Unlike in the protease field, the straightforward translation of kinase substrate information into inhibitors by using non-cleavable phosphate mimics (71) and their use for biological investigation has been limited due to the substrate size and the challenges for biological application. Most kinase inhibitors target the ATP-binding pocket and the associated hinge region, and their use has been found mostly in drug discovery with detailed pharmacophore insights (72). After the sequencing of the human genome (1), the use of substrate sequence information has been successful for the identification of kinase substrates (70). Most kinase inhibitors target the ATP-binding pocket and the associated hinge region, and their use has been found mostly in drug discovery with detailed pharmacophore insights (72).

Phosphatases and other enzymes

Conceptually, the screening for the substrates of proteases and kinases is rather straightforward, as both enzyme classes allow for detection by either fluorescence or radioactivity of their altered substrates. Other enzyme classes are not that amenable and require more elaborate assay designs. Cheung et al. (73) described an elegant scheme to identify substrates for leukocyte antigen receptor (LAR) phosphatase. Although it may be conceptually envisioned to detect the substrates for phosphatases by reduction of radioactivity after hydrolysis of a radiolabeled phosphate, the screening of combinatorial libraries for signal differences proves to be unsuitable, because of the varied concentration of the individual members. Thus, all readouts for the screening of combinatorial libraries should be laid out to yield the appearance of an effect. The screen for phosphatase substrates capitalized on the stability of phospho-tyrosine peptides against proteolysis by chymotrypsin, whereas their dephosphorylated counterparts are very susceptible to cleavage.

A 186-bead combinatorial N-acetyl peptide library that features phosphotyrosine in the center of each peptide sequence was incubated with LAR. Subsequently, the library was treated with chymotrypsin, which led to the cleavage of peptides that had been dephosphorylated. The resulting free amino-terminal were derivatized with amine-reactive chromophores or fluorophores to visualize the individual beads. After sequencing a coding peptide strand on the beads, six substrates were identified. Each of these substrates shows faster dephosphorylation than one native substrate, which is the D_WGbpG_WGpeptide of the EGF receptor (73). Several enzyme classes have been targeted using the positional scanning approach, such as farneyl transferase (74) and β-glucosidase (75), the latter establishes that the screening of peptide libraries is not limited to enzymes that transform peptides.

Studying Protein–Protein Interactions/Whole-Cell Systems

The use of combinatorial libraries for the study of protein–protein interactions is currently focused on the diagnostics and the fingerprinting of interaction surfaces. The starting point for combinatorial libraries was the investigation by Lam et al. (12) of antibody epitopes with a one-bead-one-compound peptidomimetic library. An α β-endorphin antibody recognized only 6 peptide sequences out of 2.5 million possible sequences. The intellectually intriguing aspect of the study was the use of an antibody as a biological probe; antibodies are the prime example for somatic combinatorial chemistry in biological systems.

The same study also demonstrated that combinatorial chemistry can identify ligands to proteins that are unrelated to the natural ligands. The tripeptide IPPN was the privileged binder to streptavidin, and competition experiments clearly showed that it binds at the same site as biotin.

From the screening for antibody epitopes it was a short step to address another larger class of cell-surface molecules, which is called GPCR. Using a 5000-membered dinner and trimer-peptidomimetic
IN-alkylated glycine polymers library, new ligands to adrenergic and opioid receptors were discovered. The peptoids discovered as ligands to the adrenergic receptor carry three aromatic residues and thus mimic an ephedrine polymer; the peptoids discovered for the opioid receptor resemble Met-enkephalin. Sequences for both receptors are very lipophilic, which points to a challenge that affects many library screenings. However, both types of sequences are nanomolar antagonists for their respective receptors (76). In an interesting twist, Spatola et al. (77) designed a cyclic peptoid library in a way that only one round of screening was necessary to deconvolute the most active peptoid sequence. The library contained 82,944 peptides in 48 sublibraries with various defined positions. One round of screening revealed the cyclo-Pro-D-Val-Leu-D-Trp-D-Asp(T) sequence as the most active antagonist of endothelin 1. This sequence already had been established earlier from a microbial peptide (77). Although the concept of self-deconvolution is very intriguing, the effort in design and limitations on diversity prevented further widespread application of this approach. Combinatorial libraries have also allowed to look into the intracellular side of GPCRs. Martin et al. (78) studied the interaction of the G-protein α subunit with Rhodopsin using a 11-mer peptide library displayed on phage. After several rounds of binding to rhodopsin-load membranes and amplification of the phage, several sequences similar to a sequence in the carboxy-terminus of the G-protein α subunit could be identified. After resynthesis, these sequences bound to rhodopsin and helped to stabilize the meta-II rhodopsin state. Yet, phage-display libraries are not crucial to the study of intracellular protein–protein interactions. One of the most common motifs in protein–interaction domains is the Src homology 2 (SH2). This domain is an integral part in kinase-signaling binding to the phosphorylated peptides. Thus, peptides that bind to the SH2-domain could be valuable tools for dissecting the cellular-signaling cascades. Sweeney et al. (79) used a hexapeptide library while carrying a phospho-rylserine in the center. The library beads were incubated with various biotinylated SH2-domains and the SH2-binding beads were visualized using streptavidin conjugated alkaline phosphatase. Sequences binding selectively to the SHP-1, SHP-2, and SHIP-2 SH2-domains could be identified. Furthermore, the resynthesized peptides could be used to pull down the full-length proteins from cellular extracts. The peptide sequence information was used to search for binding partners to the respective SH2-domains in genomic databases, which identified known proteins and a variety of hitherto unknown binding candidates.

For routine diagnostics in which the number of probes to be tested is smaller than in a de novo interaction screen, arrayed libraries show the higher potential because of the simplified handling. Today, these libraries are produced by spotting the libraries on microscope slides that can be read out in fluorescent microscope readers after assaying. Reddy et al. (80) reported a study that fingerprinted three different proteins against an array of 7200 octamer peptides. Each protein showed distinct interaction patterns, although the peptoid library was naïve against the proteins. Using scatter plots of the interaction intensities, clear dependencies of detection approaches (direct labeling of protein vs. antibody detection) in the read outs could be found. However, the quality of the assays and the ease of handling makes the approach using peptoids promising as a complementation of DNA and protein arrays used in diagnostics today.

One of the most recent examples of combinatorial library applications comes back to the theme of in vivo imaging discussed earlier in the protease section. A one-bead-one-compound library of urea-capped octapeptides built around a central LDV-motif was screened against JURKAT cells. Beads were selected based on observed cell adhesion under a microscope, and the compounds on these identified beads was characterized by EDMAN sequencing of encoding peptide tags in the interior of the beads, which yields only to different sequences. Because of the biasing with the central LDV motif, these compounds showed high affinity and specificity to αvβ3-integrin. After complexation with the NIR dye Alexa680-streptavidin conjugate, the peptides were used for the imaging of various tumors in nude mice. Tumors that expressed the αvβ3-integrin became fluorescent, whereas tumors not expressing the integrin did not provide imaging signals (81).

Most recently, the utility of combinatorial libraries in proteomics was demonstrated by enriching platelet proteins. Using columns loaded with a combinatorial peptide library, low abundant proteins from platelet extracts could be enriched through capturing and eluting them. The diversity of the library allowed capturing proteins without knowledge of their identity and binding partners. Using two different libraries, 175 new proteins could be identified and sequenced (82).

**Modulating Biological Systems—Artificial Receptors**

Traditionally, the approach to interrupting biological processes aims to find small molecules that fit into a binding site offered by one binding partner. This approach has been and still is very successful for drug discovery. Combinatorial chemistry opened the way to another approach for the modulation of biological processes. For the first time, it became possible to develop receptor molecules for peptides without detailed knowledge of the peptide conformation and without intricate design for the rigidification of the receptor and fitting of the intermolecular interactions. Using cheno(12-deoxy)cholic acid as a scaffold to orient combinatorial decapeptide chains, Boyce et al. (83) identified receptors that would bind and distinguish Leu- and Met-enkephalin, which are two pentapeptides. Thus, it became conceivable that one could prevent an enzymatic transformation by trapping the substrate instead of blocking the active site of the enzyme. The viability of the approach was shown on the example of the farnesyl transferase inhibitors. Farnesyltransferase inhibitors have been under preclinical and clinical investigation for a long time, but they suffer severe unwanted side effects, as a plethora of other
proteins require the farnesylation of their respective CaXX-box sequences. Although these sequences are similar enough to be transformed by the same enzyme, each protein carries its characteristic CaXX-sequence. Using a library of receptors based on four identical combinatorial tetrapeptides that grew from a trilysine scaffold, receptors that recognize the CVLS-sequence of H-Ras could be identified in an on-bead screening assay. 

After resynthesis, these receptors suppressed the farnesylation of H-Ras, while still permitting the farnesylation of K-Ras (CVM) and Laminin (CAVM) (84). One receptor prevented the localization fluorescent GFP-H-Ras to the cell membrane after microinjection into Xenopus eggs, which demonstrated the use of such receptors and molecular biology tools. In another demonstration of this approach of “epitope protection,” Zhang and Kodake (85) showed that a linear 15-mer peptide identified from phage-display library can bind to the cleavage site of interleukin-1β. This cleavage site is prototyped by Caspase-1, which is a key initiator of intracellular processes. A subset of the key switches for inflammation and apoptosis, Caspase-1 has a broad substrate specificity, which has made the therapeutic development of Caspase-1 inhibitors very challenging. The identified peptide could suppress the cleavage of interleukin-1β, while at the same time having no effect on the catalytic activity of Caspase-1, one of the key factors in Caspase-1 mediated signaling (85). Yet, not only peptide epitopes can be recognized with peptide derived receptors. Using the same combinatorial receptor library as in the H-Ras study (see above, (84)), Fukase et al. discovered receptors for liposaccharides (86). Echerichia coli Lipid A, which is a strong endotoxin, is responsible for inflammatory and lethal septic events during bacterial infections. Thus, molecules that absorb the endotoxin could be of high therapeutic value. In another affinity screening, tritium-labeled Lipid A analogs were equilibrated with the library, and beads that bind the radioactive liposaccharides were visualized by microscography. Decoding of the binding beads and resynthesis of the respective receptors yielded “molecular forceps” that can bind and distinguish the various Lipid A analogs, whereas another E. coli liposaccharide, O111:B4 LPS, was not recognized by the forceps (86). The field of using artificial receptors to camouflage enzyme substrates or to absorb toxic biologicals is still in its infancy, and no biological activity of artificial receptors beyond the cellular level has been reported to date. Yet the ability to generate and identify receptor molecules that show binding properties similar to antibodies opens intriguing avenues to chemical tools for the study of molecular biology. Once we learn how to handle the tools provided by combinatorial libraries, we will have access to a tool chest for the investigation of cells and organisms at the molecular level, the key aspect of chemical biology.

Epilogue

The use of combinatorial libraries in chemical biology is plentiful. This article focused on the plethora of possibilities and challenges in a fast developing field, which highlighted some achievements. With the evolution of synthetic repertoire and analytical technologies, high-throughput screening starts to move from structurally related libraries as discussed here to libraries assembled from historic collections. With the increased diversity of structures, we will observe many more molecular tools for chemical biology. As many efforts are focused on drug discovery, one should also expect novel therapeutic approaches based on our increased understanding of chemical and biological spaces and networks.

References
Combinatorial Libraries: Overview of Applications in Chemical Biology


Combinatorial Libraries: Overview of Applications in Chemical Biology


Further Reading

See Also

Chemical Libraries: Screening for Biologically Active Small Molecules
High Throughput Screening (HTS) Techniques: Overview of Proteins, Chemistry and Chemical Reactivity of Synthetic Peptides and Proteins to Elucidate Biological Function
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Kim E. Sapsford, George Mason University, Manassas, Virginia and the Center for Bio/Molecular Science and Engineering Code 6900, U.S. Naval Research Laboratory, Washington, D.C.

Lorenzo Berti, National Research Center on NanoStructures and BioSystems at Surfaces, Modena, Italy

Igor L. Medintz, Center for Bio/Molecular Science and Engineering Code 6900, U.S. Naval Research Laboratory, Washington, D.C.

doi: 10.1002/9780470048672.wecb174

Fluorescence spectroscopy in all of its variations can be considered among the most powerful types of analysis available to chemical biology. However, to be useful almost all applications are dependent on optimal labeling of biomolecules with a fluorophore and on the appropriate choice of analytical technique. In this article, we examine the applications and contributions of fluorescent spectroscopy to chemical biology in three inter-related sections. We first examine the properties of the common fluorophores available from many disparate structural and functional classes, which includes a discussion of their individual benefits and liabilities in the context of their application. The available conjugation chemistries used to attach fluorophores to myriad biomolecules are next reviewed. As each class of biomolecule differs in both structure and function, the focus here is on strategies for the specific labeling of different functional groups. Last, some major types of fluorescent spectroscopy and the associated biologic questions and analysis that can be addressed with them are covered briefly.

Fluorescence can be defined as the emission of a photon from an excited singlet state electron returning to the ground state orbital of a molecule where it is paired with, but of opposite spin to, the second ground-state electron (1). Fluorescence, although it typically originates from molecules containing multiple double-bonded aromatic groups, can also be emitted from many structurally diverse substances. The power of fluorescent analysis results from the ability to label myriad biomolecules with fluorophores (or to use their intrinsic fluorescence), discretely excite and visualize them in a heterogeneous sample, and then monitor their function either in vitro or in vivo with a variety of different techniques. For the purpose of this review we define chemical biology as “applying the tools of chemistry to the understanding of biological problems” and focus in this article on the diverse classes of available fluorophores, their labeling chemistries, the predominant fluorescent techniques in use, and some biologic questions that can be answered by their application. As fluorescence is a diverse and complex discipline, an in-depth description of each aspect is beyond the current scope; the focus here is more a generalized overview of fluorophore structure/function, fluorescent applications and important considerations within a chemo-biologic context. The continuing challenge in this field remains in labeling the biologic molecule(s), both in vivo and in vitro, in a specific manner such that useful data can be derived from the configuration. Fortunately, many fluorophores are available commercially along with affordable instrumentation (2). Interestingly, the most popular application of fluorescent labeling is in microscopy, for example, antibody labeling or in situ hybridization, and the many facets of this particular technique are discussed in other related articles. Additionally, myriad other prominent analytical techniques rely on fluorescent detection,
including fluorescence activated cell sorting (FACS), real-time polymerase chain reaction (PCR), and various microarray analyses, all of which are covered to some extent in other articles. For the interested reader, Lakowicz’s Principles of Fluorescence Spectroscopy can be considered the primary go-to reference for almost all questions on fluorescent analysis, including history, theory, basic and advanced concepts, techniques, fluorophores, instrumentation, and applications (1). Although primarily a catalog of fluorophores, Haugland’s The Handbook, A Guide to Fluorescent Probes and Labeling Technologies is another excellent resource (available from Invitrogen, Carlsbad, CA) (2).

Properties of Common Fluorophores

Fluorophores come in a huge diversity of structures based on the materials they are derived from and can be divided into three primary classes: organic, inorganic, and biologically derived materials. Fluorophores manifest many intrinsic physical properties that are exploited in the various experimental formats, which includes their absorption/emission profiles (spanning the UV-to-near-infrared (IR) regions of the electromagnetic spectrum), varying fluorescent lifetimes, stokes shifts and quantum yields, and in some instances their sensitivity to their local environment. The exact choice of fluorescent probe for a particular application obviously depends on several factors, including the nature of the system under investigation, the question to be addressed, the pertinent biolabeling chemistries, and the available analytical instrumentation.

Biologic materials

Amino acid fluorophores

Biologic materials are a class of fluorophores that can range from amino acids and enzymatic cofactors to fluorescent proteins. As their fluorescence originates from their structures, they are sometimes referred to as intrinsic fluorophores (1). The aromatic amino acids, tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe) are the simplest structural class of these fluorophores and are ubiquitous in most naturally occurring proteins (see Fig. 1a). The UV absorbance at 280 nm, which is commonly used for protein quantitation, and the resulting emission at 340–360 nm originates mostly from the indole ring of the tryptophan residue (1, 3). Tryptophan emission also reflects the polarity of its local solvent environment and can be sensitive to the binding of substrates/ligands, protein-protein association interactions, protein denaturation, and global conformational changes in structure (1, 4–6). These interactions can be used for protein characterization and sometimes can form the basis of certain types of biosensing (see below). Because of its UV emission, tryptophan has also been used as a donor in various fluorescence resonance energy transfer (FRET) studies examining intraprotein distances and conformational changes (3). Phenylalanine has a low quantum yield, and tyrosine residues are less prevalent in proteins and have been far less used for fluorescence.

The advent of recombinant DNA technology has enabled researchers to introduce these three residues into any desired site(s) within a cloned protein for potential fluorescent usage. It is particularly advantageous for FRET configurations in which optimal placement of a donor fluorophore is desired (2). Although it is an advantage in some cases, the widespread occurrence of these residues in almost all natural protein structures can also be considered a disadvantage, especially in applications in which observing a specific protein in a heterogeneous environment is desired. A further potential disadvantage includes excitation confined to the UV region, which can result in high background signals from scatter and autofluorescence within cells or other biomolecules in a sample matrix.

Fluorescent cofactors

Enzymatic cofactors, such as nicotinamide adenine dinucleotide (NADH), nicotinamide adenine dinucleotide phosphate (NAD(P)H), flavin adenine dinucleotide (FAD), flavin mononucleotide (FMN), and pyridoxal phosphate, are fluorescent and commonly found associated with various proteins where they are responsible for electron transport (see Fig. 1b and Table 1). NADH and NADPH in the oxidized form are nonfluorescent, whereas conversely the flavins, FAD and FMN, are fluorescent only in the oxidized form. Both NADH and FAD fluorescence is quenched by the substrate found within their cofactor structures, whereas NADH-based cofactors generally remain fluorescent when interacting with protein structures. The fluorescence of these cofactors is often used to study the cofactors’ interaction with proteins as well as with related enzymatic kinetics (1, 9-12). However, their complex fluorescent characteristics have not led to widespread applications beyond their own intrinsic function.

Fluorescent proteins

Fluorescent proteins (FPs) are a class of fluorescent probes of which green fluorescent protein (GFP) from the jellyfish Aequorea victoria is probably the best characterized and can be considered the prototype (13-16). Their utility is derived primarily from the ability to append genetically and to coexpress these proteins as a chimeric fusion with a desired target in a wide range of host cells, including bacteria, lower eukaryotes such as yeast, almost all higher eukaryotic cells, and even within transgenic animals (17). This process makes FPs, in general, a very powerful chemical biology tool for many applications, including in vivo labeling of specific proteins in cells as well as monitoring of intercellular signaling and intracellular processes (18-25). Beyond basic research, many FPs are used in a variety of applications related to drug discovery. The field has progressed rapidly to the point where now several GFP variants exist, such as blue, cyan, yellow, and various dimeric (Ds) and monomeric (m) red fluorescent proteins (BFP, CFP, YFP, DsRed, mRFP) (26-28). Continued cloning from many different species, including coral (29), in conjunction with functional optimization has led to a variety of commercially available FPs with emission maxima that range across the spectrum from 470 nm to 630 nm (Fig. 2). Tsien’s monograph on the green fluorescent protein is perhaps the best source in many important photophysical considerations when using all fluorescent
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Proteins (14). Issues of relevance when using FPs include size, maturation time; absorption and emission shifts; sensitivity to pH, O₂, temperature, and certain ions; obligatory dimerization or oligomerization; optimal placement; and photobleaching (21, 26, 30). Another important consideration, when appending such large fluorescent probes onto a protein of interest, is verifying that the primary function/activity of the endogenous parent protein is not perturbed in any way (20, 21). Rationally designed FRET sensing in live cells to monitor signaling events would not have been possible without the ability to append genetically these FPs to species of interest, with some commonly used FRET FP pairs, including GFP/BFP and CFP/YFP; see below (3, 13–18).

Phycobiliproteins, such as phycocyanin and phycoerythrin, are members of a family of fluorescent accessory, non-chlorophyll-based pigments found in cyanobacteria and eukaryotic algae. The phycobiliproteins have characteristic broad absorption profiles spanning 450–600 nm, emissions ranging 570–660 nm, and small stokes shifts; see Fig. 3. The major structural subunits, phycoerythrobilin (PEB) or phycocyanobilin (PCB), have multiple covalently attached bilin chromophores (open chain tetrapyrol, between 6 and 34) per moiety, which result both in the large molecular weight (∼10 times that of GFP) and a remarkable extinction coefficient (∼100 times that of GFP). The strong brightness and photostability of these fluorophores have led to their wide use as a very sensitive probe (31, 32). Although appearing to succumb to classic photobleaching under constant illumination, the loss of specific fluorescence of some phycobiliproteins, such as B-phycoerythrin (B-PE), may actually be from the exciting of more than one chromophore per protein that can annihilate the excited state and decrease the quantum yield (1, 33). These FPs are currently available from several commercial sources (Martek Biosciences Corporation, Columbia, MD; Invitrogen; Cyanotech Corporation, Kailua-Kona, HI; and Europa Bioproducts Ltd., Brussels, Belgium) and are commonly conjugated to antibodies or streptavidin for use in a variety of immunoassay and other detection formats.

Phytochromes are a family of bilin binding proteins that function as photoreceptors, regulating many physiologic processes.
Table 1: Physical properties of selected fluorescent probes

<table>
<thead>
<tr>
<th>Species</th>
<th>Approximate molecular weight (Da)</th>
<th>λ\text{abs max} (nm)</th>
<th>λ\text{em max} (nm)</th>
<th>Quantum yield†</th>
<th>Lifetime (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biologic Materials</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>204</td>
<td>295</td>
<td>353</td>
<td>0.13</td>
<td>3.1</td>
</tr>
<tr>
<td>NADH</td>
<td>709</td>
<td>340</td>
<td>460</td>
<td>0.4</td>
<td>2.3</td>
</tr>
<tr>
<td>FAD</td>
<td>829</td>
<td>450</td>
<td>525</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BFP</td>
<td>25–30,000</td>
<td>380</td>
<td>440</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>GFP</td>
<td>25–30,000</td>
<td>396,475</td>
<td>508</td>
<td>0.79</td>
<td>3.2</td>
</tr>
<tr>
<td>R-phycoerythrin</td>
<td>240,000</td>
<td>495,536</td>
<td>576</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td><strong>Organic Materials</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluorescein</td>
<td>332</td>
<td>490</td>
<td>514</td>
<td>0.95</td>
<td>4.0</td>
</tr>
<tr>
<td>FITC</td>
<td>389</td>
<td>495</td>
<td>525</td>
<td></td>
<td>4.1</td>
</tr>
<tr>
<td>TAMRA</td>
<td>534</td>
<td>555</td>
<td>580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROX</td>
<td>817</td>
<td>595</td>
<td>615</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>Cy3</td>
<td>766</td>
<td>550</td>
<td>570</td>
<td>0.15</td>
<td>&lt;0.3</td>
</tr>
<tr>
<td>Cy5</td>
<td>792†</td>
<td>649</td>
<td>670</td>
<td>0.28</td>
<td>1.0</td>
</tr>
<tr>
<td>Alexa Fluor 488</td>
<td>643</td>
<td>495</td>
<td>519</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>Alexa Fluor 555</td>
<td>1250</td>
<td>555</td>
<td>565</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alexa Fluor 647</td>
<td>1250</td>
<td>650</td>
<td>668</td>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Inorganic Materials</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lanthascreen Tb</td>
<td>915</td>
<td>343</td>
<td>495,545,570</td>
<td>µs–ms</td>
<td></td>
</tr>
<tr>
<td>Ru chelate(s)</td>
<td>940</td>
<td>445</td>
<td>615</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>550 nm CdSe/ZnS QDs</td>
<td>&gt;100 kd</td>
<td>UV = 550 nm</td>
<td>550</td>
<td>≥0.20</td>
<td>5–10</td>
</tr>
</tbody>
</table>

Drawn from References 1, 3, 7, and 8.

†Quantum yield added where available; the quantum yield and fluorescent lifetimes can be highly dependent on the local environment. A fluorophore free in solution may have a different quantum yield than the same fluorophore attached to a protein, which in turn also depends on the extent of protein-to-fluorophore labeling (1–3).

‡Sigma-Aldrich Product # 71603 Ru(bpy)2(phen-5-isothiocyanate) (PF6). In plants, cyanobacteria, and other microorganisms (34, 35). Unlike their light harvesting phycobiliprotein counterparts, native phytochromes binding their natural bilins (phytochromobilin or phycocyanobilin) result in a nonfluorescent conjugate caused by deactivation by an efficient double-bond isomerization. If, however, the apophytochrome is allowed to bind phycoerythrobilin (PEB; found naturally in phycobiliproteins), a highly fluorescent protein results, which is referred to as a phyfluor (36). These proteins have large extension coefficients, high quantum yields (0.7), and good photostability and pH stability. Highly fluorescent proteins can also result from site-directed mutagenesis of the apoprotein while maintaining the natural bilin (37). In contrast to phycobiliproteins, phytochrome-based FPs have not found widespread application as fluorescent probes primarily because in their natural form, they are not fluorescent and only become so through either genetic mutation of the bilin binding site or insertion of a non-native bilin.

Enzyme-generated luminescence

Chemiluminescence and bioluminescence are unique processes in the sense that the excited state species is generated enzymatically through a chemical reaction rather than photophysically. Although closely related, chemiluminescence (CL) originates from reactions with synthetic substrates. Both processes result from an enzymatically catalyzed reaction that generates an excited state chemical product that decays to produce light emission, generally between 400 nm and 550 nm (3). Although a variety of BL and CL enzymes/substrates combinations are available from which to choose, the firefly Luciferase/Luciferin pair is the most commonly exploited BL reporter system and Horseradish peroxidase (HRP) is the most popular CL system (3). BL and CL are dark field techniques, which means they do not require an excitation source, therefore, reducing the background fluorescence and greatly improving the sensitivity and potential limits of detection (LODs) when used for detection. BL/CL have been applied widely, including in microarrays and nanoarrays, in vivo imaging ranging from whole animals down to single cells, in numerous biosensors, and as tracers in immunoassays such as enzyme-linked immunosorbent assays (ELISAs) (38–40). The only major disadvantages of using BL and CL are that multiple wash and reagent addition steps are often required in the “development” of the signal, which can make the analysis time longer, and the user typically has a limited time in which to collect the generated signal as the local substrate is consumed rapidly. Therefore, when multiple samples are run
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

1.0
0.5
0.0
450 500 550 600 650 700
Wavelength (nm)
Normalized fluorescence (arbitrary units)

(a)
(b)
Emission profiles
Cyan
GFP
Zs Green
YFP
DS Red2
AS Red2
HC Red

Figure 2 (a) Ribbon structure of the GFP and (b) fluorescence emission profiles from various representative fluorescent proteins. (Modified from Reference 3.)

B-Phycoerythrin
MW ~240,000
Quantum yield ~0.98
Molar extinction coefficient ~ 2,410,000 M⁻¹ cm⁻¹

300 400 500
Wavelength (nm)
Normalized absorbance/emission

Figure 3 Absorption and emission profile of the phycobiliprotein B-phycoerythrin (B-PE), which is a multi-subunit multi-chromophore fluorescent protein with exceptional absorption and emission properties.

simultaneously in microtiter plates or microarray formats for high-throughput analysis, luminescent imaging techniques are preferred over the more traditional plate readers (41).

Organic materials

Standard organic fluorophores

Organic dye materials represent the largest and best characterized class of probes used in all manner of fluorescent analysis. As an overall class, these dyes are used in almost all areas of biotechnology, including biosensing, cellular imaging, clinical immunofluorescence, and in DNA/protein microarrays (42–45). Several major structural classes of organic fluorophore span the UV-to-near-IR spectrum; see Fig. 4. UV dyes are typically pyrene-based, naphthalene-based, and coumarin-based structures, whereas the Vis/near-IR dyes include a variety of fluorescein-based, rhodamine-based, and cyanine-based derivatives; see Fig. 5. Fluorescein dyes are extremely popular because they have good quantum yields, are relatively cheap, water soluble and readily bioconjugated, and easily excited using a standard argon-ion laser (488 nm). However, fluorescein has a high rate of photobleaching, is sensitive to pH (sometimes considered an advantage; see below), and can self-quench at high degrees of substitution onto biomolecules. Various alternatives are available such as Oregon Green dyes (Invitrogen, Carlsbad, CA), which are fluorinated fluorescein analogs; AlexaFluor dyes (Invitrogen; (46)); Cy dyes (GE Healthcare, Buckinghamshire, United Kingdom); and BODIPY dyes (Invitrogen), all of which claim to alleviate some of these issues. For some redder dyes, overlabeling can induce protein precipitation because of their low solubility in aqueous environments (47). A fos, several organic-based dyes are used as fluorescent stains for visualizing cell membranes, proteins, and DNA in cells or separation gels (1, 2, 48). These dyes also include fluorescent dyes attached to lipids that allow membrane labeling and intercalating probes like ethidium bromide, which are typically weakly fluorescent until bound to DNA (1, 2, 48). In general, organic dyes have several issues that have to be considered before use; these include broad absorption/emission profiles with small stokes shifts (which can be problematic for FRET-based applications or multiplexing), solubility issues, and susceptibility to environmental influences. Cumulatively their many advantages, which include extensive characterization in the literature, their wide availability from several commercial sources, relative cost, ease of use, and many different available bioconjugation chemistries, often outweigh the above liabilities and mean that they are usually the first choice in most experiments. Haughland’s handbook is an excellent reference on many aspects of most commercial organic fluorophores (2).
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Figure 4  Examples of commercially available fluorophore families. (Modified from Reference 3.) Absorbance and emission maxima along with spectral regions covered by a particular dye family are highlighted. The major suppliers are as follows: Molecular Probes, Inc. (Eugene, OR; Fluorescein, rhodamine, TAMRA, TAMRA and ROX, AlexaFluor, BODIPY, Oregon Green, and Texas red), GE Healthcare (Cy dyes); AnaSpec, Inc. (San Jose, CA; HiLyte Fluors), ATTO-TEC GmbH (Siegen, Germany; ATTO dyes); Molecular Biotechnology (DY dyes); and Pierce Biotechnology, Inc. (Rockford, IL; DyLight 547 and 647 dyes).

Figure 5  Chemical structures of common UV/Vis fluorescent dyes. Typical groups at the R position include CO$_2$-, SO$_3$-, OH, OCH$_3$, CH$_3$, and NO$_2$; R$^*$ marks the typical position of the bioconjugation linker. (Modified from Reference 3.)

Environmentally sensitive fluorophores

Environmentally sensitive fluorophores that exhibit some change in their absorption or emission properties as a function of their environment are often referred to as indicator probes (1, 3). Such dyes may be sensitive to changes in pH, ionic strength, ionic type, oxygen, solvation, or polarity. Although many dyes will exhibit some sensitivity to perturbation of their local environment, fluorescein is probably the best known; with sensitivity to pH, see Fig. 6. A gain Haughland’s handbook is an excellent reference for organic dye probes optimized for environmental sensing (2). Many new indicator probes, which are sensitive to neutral and ionic molecules as well as to oxygen reactive species, have been reviewed recently in the literature (49, 50). These indicator dyes have applications for chemical sensing both in vitro and in vivo. Important considerations when loading cells with these indicator probes include whether the probe is as follows: toxic to the cell, and at what concentration, and where the dye will likely be sequestered during loading into the cell, such as within the organelles.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Fluorescent polymeric microspheres and nanospheres

One photophysical limitation of organic fluorescent dyes is the tendency toward self-quenching that occurs when attempting high conjugation ratios to improve both sensitivity and LODs. Researchers have overcome this limitation by functionalizing polymeric microspheres and nanoparticles with multiple fluorophores (100s–1000s) that result in highly fluorescent particles. This strategy also allows one to label biomolecules with dyes that would otherwise lack reactive groups for bioconjugation or are inherently insoluble in an aqueous environment (3). Fluorescently functionalized microspheres and nanospheres are available commercially from several sources, spanning the UV-to-IR, in a variety of sizes ranging from 20 nm to 5 µm (3). Fluorescent microspheres are also commonly labeled with primary or secondary antibodies and are used as solid supports in many sandwich immunoassays, where they take full advantage of solution-based kinetics (51–53). Alternatively, microspheres can be obtained with carboxyl, amine or other surface-displayed functional groups for chemical conjugation to biologic molecules (3). The recently developed FloDots, which are functionalized with either antibodies or DNA, have demonstrated applications in several areas, including bioimaging, cell detection, gene detection, and protein arrays (54). Similar sol-gel-derived silica nanoparticles have also been investigated (55). Coded microspheres, which are internally labeled with specific concentrations of two fluorescent dyes, are an integral part of the Luminex flow cytometry technology and have been used in many biomedical research and diagnostic applications (56, 57). Environmentally sensitive dyes have been immobilized in various cross-linked polymers to produce PEBBLE (Probes Encapsulated by Biologically Localized Embedding) nanosensors for several ionic and neutral species. When delivered into cells, these PEBBLE nanosensors are used to image and monitor the presence of intracellular chemical species (58, 59). In general, the liability of working with a microsphere that is far larger than an organic or intrinsic fluorophore can be offset greatly by the increased sensitivity and stability.

Inorganic materials

Metal chelates and long-lifetime fluorophores

Long-lifetime dyes typically consist of either the luminescent lanthanides or the ruthenium-metal chelates. The principle advantage of working with long-lifetime dyes over conventional fluorophores originates from the ability to gate out, through time-resolved measurements, any background fluorescence originating from matrix components in the sample (such as autofluorescence and scattering), thus greatly improving the sensitivity of detection. Of the four lanthanides, terbium (Tb), europium (Eu), samarium (Sm), and dysprosium (Dy), that emit in the visible region, Tb and Eu are the most commonly used probes (60–63). In particular, chelate ligands of the lanthanide ions are typically used for biophysical applications. The chelates are designed to 1) tightly bind the lanthanide ion, imparting high thermodynamic and photochemical stability; 2) position a sensitizing chromophore in close proximity to the ion; and 3) contain a reactive group allowing bioconjugation (see Fig. 7). Because of their relatively low extinction

Figure 6. (a) Normalized absorption and emission (b) changes of fluorescein as a function of its ionization state (c). The monocationic and dicationic species are the more fluorescent. (Modified from Reference 3.)
coefficients (~1 M\(^{-1}\) cm\(^{-1}\)), lanthanide ions are almost impossible to excite directly so the sensitizing chromophore is an essential part of the lanthanide chelate probe as it functions as the initial light-harvesting antenna. Lanthanide probes are characterized by unique, sharp emission profiles (see Fig. 7), which originate from both magnetic and electric dipole transitions, along with large Stokes shifts (excitation 343 nm, emission > 500 nm). The fluorescent lifetimes of these materials range from ms to ms, and this allows lanthanide probes to be used in luminescence resonance energy transfer (LRET) applications along with various clinical bioassays (60, 61, 63-65). Commercial sources for lanthanide probes include CIS-Bio International (Yvette, France) and Packard Instrument Company (Downers Grove, IL; cryptate-based probes), PerkinElmer, Inc. (Waltham, MA; LANCE, a pyridine-based system), Invitrogen (Lanthascreen; polyaminocarboxylate chelates, coupled with a CS124 sensitizer), and GE Healthcare who market a europium (Eu) complex in the iso(thio)cyanate form for biocorjugation to amines. Also, an increasing number of reports appear in the literature of lanthanide-doped nanoparticles being used as bifunctional labels where their benefits can be compounded for exploitation (66-68).

Long-lifetime ruthenium metal–chelate complexes were originally used by Lakowicz as anisotropy labels for measuring the rotational dynamics of proteins (69, 70). These transition metal complexes typically contain one or more bi-dentate or tri-dentate ligands and use metal-to-ligand charge transfer processes; see Fig. 8 (1, 71). Sigma-Aldrich Corporation (St. Louis, MO) offers a series of reactive ruthenium complexes for biolabeling that are derived from Lakowicz’s work. They are characterized by relatively small extinction coefficients (34,500 M\(^{-1}\) cm\(^{-1}\) and low quantum yields (0.05), but they possess long lifetimes, typically >500 ns, coupled with a fairly large Stokes shift. As with the lanthanide probes, several applications have used the ruthenium complex as a donor in LRET-based studies. The major advantage of these transition metal complexes is that they can be activated to give an acceptor dye, which leads to significantly improved sensitivity (72).

Fluorescent metal nanoparticles

Atomic clusters consisting of a few noble-metal atoms have demonstrated intense fluorescence emission as well as some interesting nonlinear optical properties. Gold (73), silver (74, 75), and copper (76) nanoparticles have both size-dependent and shape-dependent absorption and luminescent profiles, with the resulting emission exhibiting strong resistance to photobleaching. The potential benefit here would be the extremely small size; however, these nanoparticles are currently not available commercially and to date have found limited use as fluorescent labels, mostly because of complex synthesis and stabilization along with lack of biocorjugation routes.

Semiconducting nanoparticles: quantum dots

The II-VI group of semiconductor nanocrystal materials or quantum dots (QDs), namely CdSe and CdS, are extremely well characterized in the literature and have proven their applicability as fluorescent probes in a wide range of biologic applications, including FRET-based investigations (3, 77), biosensing (78), drug discovery (23), cancer diagnostics and investigative treatments (79, 81), and many other areas of biomedical research (82, 83). QDs or Qdots have several unique properties that make them ideal fluorescent probes compared with organic dyes: 1) high quantum yields and molar extinction coefficients (>10 to 100 times those of organic dyes that steadily increase toward the UV); 2) high resistance to photobleaching and chemical degradation; 3) broad absorption coupled with narrow emission profiles (see Fig. 9b); 4) large Stokes shifts, >200 nm for the red-emitting QDs; and most importantly, 5) size-tunable photoluminescence; see Fig. 9a (83, 88). Although QDs have been synthesized from many binary and ternary semiconductors, the most commonly used are a CdSe QD core overcoated with a ZnS shell, which acts to passivate and protect the core (83, 89). Critical to the development of QD biotechnology has been the ability to make these inherently insoluble materials biocompatible through the use of bifunctional ligands (83, 87). QDs are available commercially from Invitrogen and Evident Technologies, Inc. (Troy, NY); they are typically overcoated with a protecting layer and then precoated with avidin or other proteins that facilitate biocorjugation. Several detailed discussions have occurred on the synthesis of QD core-shell materials (89, 92). Many applications specifically aim to minimize the large protecting outer layers, which is often desired in FRET-based applications where acceptor distance from the QD donor core can be an issue (77, 93).

Silicon and more recently germanium semiconductor-based nanocrystals have also demonstrated intense fluorescence emission (94-98). Si-based nanoparticles have found some applications as biolabels (98, 99), whereas Si-based and Ge-based nanoparticles are still in the early stages of development. However, all are characterized by immature synthesis and stabilization that has limited their biologic use to date. That said, they still have a promising future as optical probes that potentially may be a nontoxic alternative to semiconductor materials.

Fluorophore Bioconjugation

Chemistries

To exploit analytical fluorescence in biologic assays fully, it is often necessary to label target biomolecules by introducing a fluorophore, usually through a chemical reaction. Ideally, the reaction should take place in aqueous conditions, be fast, high yielding, and nondestructive; target reactive groups already present on the biomolecule; and allow easy purification of product from unreacted dye. Many reactions have been developed that fit some or most of the above criteria, and this section is only intended to give the reader a general outline of the common/practical chemical methods of achieving fluorescent labeling. Several excellent reviews (100) and books (3, 101, 102) cover this extensive topic exhaustively, and the reader is referred to these and to the current literature in the field of biocorjugation for additional insight.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

**Figure 7** Luminescent lanthanides. (Modified from Reference 3.) (a) Structure of the LanthaScreen Tb probe from Invitrogen. The bioconjugation group is typically a NHS ester, isothiocyanate, or maleimide group. (b) The unique sharp emission profile of the LanthaScreen Tb probe $\lambda_{\text{ex}} = 343$ nm.

**Figure 8** Transition metal chelates. (Modified from Reference 3.) (a) Structure of the commercially available ruthenium complex, Ru(bpy)$_2$(phen-5-isothiocyanate)(PF$_6$)$_2$. (b) The resulting absorption and emission profiles of the ruthenium complex.

### Synthetic considerations

Labeling biomolecules with fluorophores can be challenging because of the differences in the nature of the two reactants and the small quantities of reagents employed (pmol to µmol). Because of their widespread incorporation in countless assays, many fluorophores are now available commercially in reactive form, targeting chemical groups already present in biomolecules, with amines and thiols being by far the most common targets (2). Although sometimes expensive, reactive fluorophores are the most convenient to use as a stock solution can be prepared and the conjugation reaction can be performed by adding a known concentration of fluorophore to the biomolecule dissolved in the appropriate buffer. Fluorophores are often hydrophobic molecules with limited water solubility, whereas biomolecules are soluble in water at a wide pH range. To facilitate the reaction between these disparate entities, a large excess of dye is usually employed. As a result, a common problem often encountered in fluorescent labeling of biomolecules is the presence of a heterogeneous reaction mixture: soluble biomolecule and solid dye in two immiscible phases. One way to address partially this issue consists of adding small quantities (usually between 20% and 50% of the total reaction volume) of a carrier solvent, such as dimethylsulfoxide (DMSO) or N,N-dimethylformamide (DMF). The carrier solvent facilitates the solubilization of some hydrophobic dye in the water medium and diminishes the problems associated with the formation of aggregates. Labeling reactions usually take place under some moderate-to-vigorous stirring, and if the desired substitution level is not achieved in a reasonable reaction time (usually 1-2 hours), repeating the conjugation is preferred over extended reaction times as many reagents tend to hydrolyze over time. Because of the necessity of...
lowering background fluorescence in the final assay, purification of the fluorescent adduct from the unreacted fluorophore is necessary and sometimes complex. The large excess of fluorophore employed in the labeling reaction can compound purification of the labeled adduct, and several rounds of purification using different techniques may be necessary. Centrifugation and precipitation are the first steps that can help in removing suspended unreacted fluorophore and in facilitating additional purification steps. Gel-permeation, reverse-phase high-performance liquid chromatography, electrophoresis, dialysis, and combinations thereof are the techniques more commonly employed to achieve a high-purity labeled product. Depending on application, site-specific labeling may be a preferred strategy as opposed to nonspecific or generalized labeling of a common functional group, for example, a single-labeled protein-thiol versus multiple lysine-amines. Lysines are abundant on most proteins, there is also the possibility of a distribution of labeled-adducts expressing various degrees of functionalization. Furthermore, lysine residues may be the key to subsequent function, and labeling can interfere with this process.

Polypeptides and nucleic acids as biomolecular targets

Proteins, antibodies, and enzymes consist of long polypeptide sequences folded in ways that favor the positioning of hydrophobic residues at their interior, whereas more hydrophilic or easily ionizable residues are at the periphery. Water-exposed residues are therefore important and available targets for bioconjugation, most notably the carboxyl group of aspartic (103) and glutamic acid (104), the primary amine of lysine and arginine (105, 106), the secondary amine of histidine (107, 108), the thiol of cysteine (109, 110), and the phenolic hydroxyl of tyrosine (111). All behave as nucleophiles under appropriate conditions, such as specific pH ranges, and their reactivity can be exploited for the attachment of several reactive probes (112, 113). Such target-expressing residues can also be introduced recombinantly at well-defined specific sites in proteins (114, 115). Besides the presence of these residues, it is also possible to introduce alternative functional groups, or to interconvert existing ones, through the use of versatile and readily available bifunctional reagents (116). For example, 2-bromoethylamine can be used to transform thiol-containing amino acids (cysteine) into primary amines, whereas N-succinimidyl-s-acetylthioacetate (SATA) reacts with amine containing amino acids to form a protected sulfhydryl group, which can be deprotected readily using hydroxylamine to a free thiol (101). The use of bifunctional reagents provides versatility to the bioconjugation chemist, extending the repertoire of reactions and functional groups available and permitting the fine-tuning of other factors, such as the spacing or the solubility of the fluorophore.

Nucleic acids do not display the same promiscuous chemical reactivity of proteins. Instead, individual synthetic nucleotides can display a unique functional group that can be exploited even more for direct attachment of probes (117). The simplest method of labeling DNA uses high-affinity bis-intercalating dyes such as ethidium bromide, acridine, and thiazole orange monomers...
and dimers (118). These dyes, which have almost no emission in water, display fluorescent enhancements of 30–1000-fold after binding to DNA and remain bound even during electrophoresis, hence, their use as gel stains. Different DNA samples can be prestained with differentially emissive dyes before electrophoresis, then mixed, and finally covalized (119). Perhaps the biggest liability of these dyes is their toxicity, which necessitates careful handling and disposal. The 5′-backbone phosphates can be targets for modification through carboxylimide-mediated formation of a phosphoramidate (120, 121). Such reactions, however, are not easy to control, and the expertise required makes this type of approach not very common. Two other techniques are usually preferred for tagging DNA with a fluorescent probe: 1) direct enzymatic labeling and 2) postsynthetic modification. Direct enzymatic labeling usually uses some type of enzymatic introduction of modified ddNTPs (122, 128), which results in the generation of an oligonucleotide sequence that contains the probe or an appropriate functional moiety such as an amine to which a probe can be attached. Several enzymatic techniques are available, including PCR, nick-translation, random-primed labeling, and 3′ tailing with terminal transferase, which are all methods reviewed in References 129 and 130. Issues of this enzymatic insertion include the limited types of modified nucleotides that can be introduced, as the enzymes employed can be very selective when it comes to incorporating non-natural substrates, cost, and the necessity of long incubations in some cases. Postsynthetic modification of nucleotides usually consists of the automated synthesis of a short DNA fragment incorporating a specific chemical moiety—usually an amine or a thiol—within or at the termini of the sequence (117, 130–132). The functionalized oligo is then tagged with the fluorescent probe, which exploits reactions very similar to those employed for the proteins above or listed in Table 2. These reactions, however, are not easy to control, and the expertise required makes this type of approach not very common. Two other techniques are usually preferred for tagging DNA with a fluorescent probe: 1) direct enzymatic labeling and 2) postsynthetic modification.

**Type of linkages**

Table 2 presents a list of the most common target groups used for fluorescent biodiagnosis along with their cognate reactive partner and the type of product formed. Amines and thiols are by far the most commonly exploited target functional groups as they are both good nucleophiles, readily available and/or easily introduced in biomolecules. Figure 12 shows the most common reactions for targeting amines. Isothiocyanate (ITC) (153–155) and succinimidyl ester (OSu) (151, 152) derivatives are the most common amine-reactive groups readily forming, respectively, stable thioesters and amide linkages (Fig. 1a, 1b). ITCs, however, present long-term storage stability problems, which makes the OSu derivatives more popular. Nonetheless, certain types of reactive ITC probes, such as the fluorescein (FITC) and tebramethylrhodamine (TRITC) derivatives, are still used widely. OSu derivatives are extremely common, and many of them are available commercially (3). OSu derivatives can also be prepared as starting material from a carboxylic derivative and N-Br-Succinimide as reagent. Less common, but still viable, reactive partners for amines are sulfuric chlorides, unstable in aqueous (156), which leads to more stable sulfonamides (Fig. 1c) and aldehydes (157) yielding Schiff bases (imines) (Fig. 1d). Carboxylic acids can also be reactive toward amines, ultimately forming amide bonds, but they require prior activation usually achieved by employing a water-soluble carboximide reagent (Fig. 1e) (158, 164).

Maleimides (133–138), iodoacetamides, and alkyl halide derivatives (133, 139–141) are the most common thiol-reactive groups (Fig. 1a, 1b). All of these reagents readily react when a free thiol is present to give stable thioethers. Disulfide exchange of a free thiol with an activated piriylisulfide is another common and effective reaction (146, 147, 171, 174) (Fig. 1c), but the resulting disulfide containing product may not be stable under reducing conditions or in the presence of other free thios.

Other chemical groups such as aldehydes, ketones, and alcohols are also exploitable targets for bioconjugation. The reactivity of alcohols in water is very low, and usually intermediate steps are involved to transform the hydroxyl into a more reactive or more easily exchangeable group. The resulting multistep scheme, although not an alternative to more readily labeled groups, is an option when other reactive groups, such as carbohydrates, are not present (175). Aldehydes and ketones react with amines to give imines (149, 150). However, both groups are not commonly found and often have to be generated by oxidation of the corresponding alcohol or vicinal diol. Hydrazines are also very good reactive partners for ketones and, to some extent, aldehydes, which yield hydrazones (148). The hydrazones can be stabilized even more by reduction yielding an irreversible product. Interestingly, these same groups are also common targets for modifying sugars and carbohydrates using

*Glen Research (Sterling, VA; www.glenres.com) and Trilink Biotechnologies (San Diego, CA; www.trilink.com) offer a variety of fluorescent phosphoramidates for automated DNA synthesis.*
some of the same chemistry. However, sugar modification is a
less well-developed chemistry in general and thus less prevalent,
although strong research in this area continues (2, 101).

Another commonly exploited labeling method takes advan-
tage of the very strong interaction between the protein Avidin
(or Streptavidin) (176) and its natural ligand biotin, as reviewed
in Reference 2. This labeling scheme usually starts with the bi-
otinylation of the target biomolecule by employing one of many
commercially available biotinylating reagents. The biotinylation
chemistries available are essentially similar to the ones already
mentioned for the fluorescent labels above. After biotinylation, fluo-
rescently labeled Avidin or Streptavidin is added, which results
in strong binding to the biotin and forms a basically irreversible
linkage that has found use in countless biologic applications
(177, 178).

Beyond the above chemical methods for the introduction
of fluorophores, several emerging technologies target in vivo
fluorescent labeling for applications where the probe has to be
located inside a target cell. Fluorescent proteins such as the
GFP can be appended onto the target protein by recombinant
techniques resulting in the coexpression of fluorescent protein
cimers (14, 179). FLASH technology can be used in the specific
in vivo tagging of proteins expressing a Cys–Cys–Pro–Gly–Cys–Cys
sequence by employing a cell-permeable dye that becomes fluores-
cent only during labeling (180–182). The HaloTag method consists of a fusion

---

Table 2: Selected biologic functional groups and their corresponding target chemistries

<table>
<thead>
<tr>
<th>Target</th>
<th>Reactive group</th>
<th>Product</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Thiol</td>
<td>Maleimide</td>
<td>Thioether</td>
<td>(133–138)</td>
</tr>
<tr>
<td></td>
<td>Haloacyl/Alkyl Halide</td>
<td>Thioether</td>
<td>(139–141)</td>
</tr>
<tr>
<td></td>
<td>Aroylating agents</td>
<td>Thioether</td>
<td>(142)</td>
</tr>
<tr>
<td></td>
<td>Aziridine</td>
<td>Thioether</td>
<td>(143)</td>
</tr>
<tr>
<td></td>
<td>Aroyl derivatives</td>
<td>Mixed disulfide</td>
<td>(144, 145)</td>
</tr>
<tr>
<td></td>
<td>Pyridyl disulfides, 5-thio-2-nitrobenzoic acid (TNB)</td>
<td></td>
<td>(14, 147)</td>
</tr>
<tr>
<td></td>
<td>Hydrazine</td>
<td>Hydrozone</td>
<td>(148)</td>
</tr>
<tr>
<td>Aldehyde/Ketone</td>
<td>Amines</td>
<td>Schiff’s base (imine)*</td>
<td>(149, 150)</td>
</tr>
<tr>
<td></td>
<td>N-hydroxysuccinimide</td>
<td>Amide</td>
<td>(151, 152)</td>
</tr>
<tr>
<td>Free Amines</td>
<td>Isothiocyanates, isothiocyanates</td>
<td>Urea, Thiourea</td>
<td>(153–155)</td>
</tr>
<tr>
<td></td>
<td>Acyl azides</td>
<td>Amide</td>
<td>(101)</td>
</tr>
<tr>
<td></td>
<td>Sulfonyl chlorides</td>
<td>Sulfonamide</td>
<td>(156)</td>
</tr>
<tr>
<td></td>
<td>Aldehydes, Glioxals</td>
<td>Imine, secondary amine*</td>
<td>(157, 158)</td>
</tr>
<tr>
<td></td>
<td>Epoxides, Oxiranes</td>
<td>Secondary amines</td>
<td>(101)</td>
</tr>
<tr>
<td></td>
<td>Carbonates</td>
<td>Carbamate</td>
<td>(101)</td>
</tr>
<tr>
<td></td>
<td>Aroylating agents</td>
<td>Arylamine</td>
<td>(101)</td>
</tr>
<tr>
<td></td>
<td>Imidoesters</td>
<td>Amidine</td>
<td>(159, 160)</td>
</tr>
<tr>
<td></td>
<td>Carbodimides, Anhydrides</td>
<td>Amine*</td>
<td>(120, 121, 161)</td>
</tr>
<tr>
<td></td>
<td>Diazaoalkanes, Diazaoacetyl</td>
<td>Ester</td>
<td>(162, 163)</td>
</tr>
<tr>
<td>Carboxylate</td>
<td>Carboxyldiimidazole, Carbodiimides</td>
<td>Amides*</td>
<td>(159, 164)</td>
</tr>
<tr>
<td></td>
<td>Epoxides</td>
<td>Ether</td>
<td>(101)</td>
</tr>
<tr>
<td>Hydroxyl</td>
<td>Carboxyldiimidazole, N,N’-diisuccinimidylic carbonate, N-hydroxyisocyanide</td>
<td>Carbamate or Urethane*</td>
<td>(165, 167)</td>
</tr>
<tr>
<td></td>
<td>Chloromethacrylate</td>
<td>Chloroformate</td>
<td>(168, 169)</td>
</tr>
<tr>
<td></td>
<td>Alkyl halogen</td>
<td>Carbamate</td>
<td>(170)</td>
</tr>
</tbody>
</table>

* Might be followed by reducing amination to form a stable product.
† After reduction.
‡ Via reactive intermediate.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Figure 10 Schematic representation of the most common reactions for labeling an amine: (a) reaction with isothiocyanate to give a thiourea; (b) reaction with a succinimidyl ester to give an amide; (c) reaction with a sulfonyl chloride to give a sulfonamide; (d) reaction with an aldehyde to give an imine (Schiff's base); and (e) reaction with a carbodiimide-activated carboxylic acid to give an amide.

A recent and growing conjugation methodology worth mentioning is the implementation of click chemistry to bioconjugation. Click chemistry is a fairly generic term referring to a certain class of quick, high-yielding reactions. For bioconjugation, this term usually indicates the Cu(I) catalyzed 1,3-dipolar (Huisgens) cycloaddition between an alkyne and an azide to give a 1,2,3-triazole as the product. This reaction proceeds very rapidly and is compatible with both aqueous chemistry and a variety of functional groups, which makes it an excellent candidate for future development and applications (184, 185). However, the current drawback to this chemistry is the unavailability of widely applicable commercial kits for introducing the necessary alkyne and azide cognate precursors onto both target and probe. However, the Click-iT kit (Invitrogen) targeting glycoproteins probably represents the first of many applications to come that will use this exciting chemistry.

Selected Fluorescent Techniques

Fluorescence anisotropy

The extent of the polarized emission from excited state fluorophores in a solution can be described in terms of their anisotropy (r), and measuring this can provide insight into the angular displacement of the biomolecule(s) to which the fluorophores are attached (1). Within a homogeneous solution, the ground-state fluorophores are all oriented randomly. However, when exposed to a polarized excitation source, the molecules with their absorption transition moments oriented
Fluorescent labeling and fluorescent spectroscopy: Overview of applications in chemical biology

**Figure 11** Schematic representation of the most common reactions for labeling a thiol: (a) reaction with an alkyl halide to give a thioether; (b) reaction with a maleimide to give a thioether; and (c) reaction with an activated pyridyl-disulfide to give a mixed-disulfide through disulfide interchange.

Along the electronic vector component of the light will be excited preferentially (1), which means the excited population is not random in orientation, but their transition moments are all oriented. For chemical biology, the factors that affect the rotational correlation time, or anisotropy, can then be investigated, including biomolecule-biomolecule interactions, modifications, or denaturation. Anisotropy is also a powerful tool for measuring viscosity in select environments such as membranes and lipid composition. Again, the interested reader is referred to Lakowicz’s text for more detailed reading on all facets of this subject (1).

Anisotropy, which is sometimes used interchangeably with polarization, although the former is preferred, is a dimensionless quantity that is independent of sample intensity. A fluorometer and appropriate polarizing filters (parallel and perpendicular) are the simplest instrumental setup that can be applied (1). Measuring the steady-state fluorescence anisotropy provides data on only the average anisotropy decay and the interpretation can be complex. Far more information is gained from measuring the time-resolved anisotropy; however, the equipment and analysis required is more complex. Direct monitoring of fluorescent anisotropy provided insights into the binding modes of the G protein-coupled type A and B cholecystokinin receptors (186). These receptors, which share homology to rhodopsin and β-adrenergic receptors, have important regulatory functions in certain hormonal responses. Monitoring changes in anisotropy of differentially labeled probes while interacting with the receptors confirmed that each type uses a distinct mode of high affinity binding. The structure of UrEc, which is an essential Bacillus pasteurii protein required for the in vivo activation of the enzyme urease, has also been probed with both steady-state and time-resolved anisotropy analysis (187). Although this protein behaves as an intrinsic unstructured dimer, the conformation it assumes is unknown, for example, fully folded, molten globule, or random coil. Direct analysis of steady-state anisotropy and intrinsic tryptophan fluorescence wavelength shift allowed monitoring of transitions between native and unfolded states upon increasing concentrations of a denaturant (see Fig. 12). Furthermore, the hydrodynamic parameters obtained by time-resolved fluorescence anisotropy in the presence of a denaturant confirmed the existence of a stable but disordered dimer formed at a unique cysteine residue. This key bond acts to stabilize the dimer under native conditions.

**DNA and fluorescence**

Fluorescent labeling and spectroscopic analysis have had a profound impact on two areas: research into DNA structure/function and myriad diagnostic and sequencing applications. For the former area, the focus is to understand complex
nucleic acid chemistry and more recently to exploit the inherent self-assembled structures for creating precisely formed nanoscale architectures (188, 189). For the latter arena, the focus has been on all aspects of genomic analysis from clinical/genetic diagnostics to the creation of the genomic databases (190). Although this field is relatively young, with fluorescent DNA sequencing described just ~20 years ago (191), commercial applications have driven progress and the completion of the Human Genome map (192); many focused reviews are available in this area (3, 193-195).

Fluorescent sensing

Fluorescent sensing can provide a powerful tool to the chemical biologist especially for in vivo applications. One version of this analysis originates almost exclusively from the environmental sensitivity of selected fluorophores discussed earlier (1-3). Fluorescein, for example, has been exploited for intracellular pH monitoring (see Fig. 6). Intracellular calcium-sensing techniques are a powerful and widely used technology with many applications in neuroscience and in cardiovascular and signal transduction research. It has also proven useful recently in monitoring receptors associated with odor detection (194). This particular technique is discussed extensively in other articles; see, for example, the article “Calcium Signaling.” The ability of membrane-localized receptors to “sense” and transduce an odorant was monitored by their ability to elicit a coupled-intracellular calcium transient. Invitrogen also offers a variety of commercial probes targeting diverse ions as modified endogenous intracellular analytes. "Fluorescent Dyes and Other Photochemically Active Molecules," in Haughland's handbook is an excellent reference on this subject (2). This type of fluorescent sensing is well developed but may be limited by the lack of multiplexing capability. That is, only one or two of these dyes can be used simultaneously because of their broad absorption/emission profiles.

Fluorescent sensing has been exploited even more by implanting environmentally sensitive fluorophores into select proteins to create a variety of in vitro and in vivo sensors. The proteins provide biologic specificity for target recognition, whereas the fluorophores provide signal transduction. This type of sensing relies on some change in the proteins structure during binding, which in turn causes a change in the local environment of the fluorescent dyes and, thus, its photophysical state. The superfamily of bacterial periplasmic binding proteins has provided an excellent source of sensing proteins for diverse analytes from which to begin designing such sensors (197), but long-term fluorophore instability has remained an issue and may necessitate preparation of fresh sensors for each experiment. The prototype for this sensor design has been the maltose binding protein (198), and many different maltose sensors have been assembled to test a variety of signal transduction modalities, including a variety of surface-tethered versions with complex kinetic functions; see Fig. 13 (198, 199). Hellinga and Frommer (197, 200) have been at the forefront of this field where they have applied computational design to identify critical sensing sites within a variety of natural and de novo rationally designed proteins.

Energy transfer

As FRET is a significant analytical technique for a variety of in vivo and in vitro biosensing configurations, the concepts and applications will be discussed in other articles. However, FRET is heavily dependent on optimal placement of two or more fluorophores either within a single biomolecule or in two separate biologic entities (1, 3, 201). As such, discussion is warranted on some relevant issues. The first is choice of fluorophores with appropriate spectral overlap and their placement such that the proximity exists between the fluorophores for efficient FRET (1, 3, 201). Depending on the structure and the available functional groups, site-specific sequential or orthogonal labeling of a single biomolecule with multiple fluorophores is still challenging. DNA and other oligonucleotides are synthesized readily with both site-specific amine- or thiol-functions for facile labeling with appropriately reactive dyes (2, 101). Single cysteines can be introduced recombintantly into specific protein sites; however, the presence of other cysteines can cause “thiol-scrambling” of the structure and subsequent loss of function (3). Hellinga and colleagues (202) have described recently a method for the sequential labeling of multiple cysteines within a single protein to address this issue specifically. Primary amines are ubiquitous to proteins, and thus, fluorophores targeting these functions tend to be nonspecific. The same considerations for thiois and amines usually apply to labeling synthetic peptides.

Proteins can also be engineered to express a variety of fluorescent proteins appended at different points within the structure. Frommer and colleagues (203, 204) have designed advanced sensors that couple two fluorescent proteins at select sites within sensing proteins and ligand binding is signaled by changes in their FRET efficiency. Figure 14 is an example of a cell-surface expressed version of such a sensor that detects glutamate release from neurons (205). The recently developed FLASH, HaloTag, and SNAP tag techniques offer alternative methods for labeling specific recombinant sequences in vivo with proprietary reactive dyes (3). The use of two disparate fluorophore classes as donors/acceptors in FRET configurations is steadily growing as exemplified by quantum dot—dye or fluorescent protein—dye pairings (3, 7). Interestingly, although not fluorophores per se, fluorescent quenchers and gold nanoparticles have found utility as acceptors in myriad FRET applications (3). In vivo sensing of toxins and second messengers with FRET-based fluorescent protein sensors is an important and related area covered in several other chapters.

Single-molecule analysis

In theory, single-molecule analysis (SMD) represents the ultimate analysis as it can provide the highest sensitivity while providing stochastic sensing. In an ideal experiment, fluorescent signal transduction is suited particularly to this analysis as a lone fluorophore can emit photons against a dark background. Practice has proven far more complex, however, and this remains a technically challenging and still nascent technique (3). The two confounding issues remain sample immobilization coupled with optical detection configuration. Almost all SMD is dependent on microscopy and the 1D immobilization of biomolecules by surface tethering or fixation within a network.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

Figure 13  (a) Schematic of a surface-tethered maltose biosensor with complex binding kinetics. The modular arm consists of DNA, a dye, and terminates in the maltose analog beta-cyclodextrin (β-CD). The maltose binding protein (MBP) is site-specifically dye labeled, and both it and the β-CD are assembled onto a Neutravidin (NA) functionalized surface using biotin (b). For MBP, this uses a biotin–nickel nitroloacetic acid (Ni-NTA) intermediary to bind the MBP’s 5-histidine sequence, 5-HIS. MBP binding of the β-CD-dye-DNA arm assembles the final sensor by bringing dyes 1 and 2 into proximity, which allows FRET or FRET quenching. Maltose displaces the β-CD-dye-DNA arm in a concentration-dependent manner. The addition of second modulator DNA that hybridizes to the arm can alter and extend the binding kinetics. (b) Representative binding curve and approximate binding constant (Kapp) for titrating the sensor against maltose. (Reprinted from Reference 199 with permission of the ACS.)

or 2) monitoring and averaging of a continuous sample of individual molecules to overcome any photobleaching effects. The required instrumentation is still some combination of confocal imaging, cooled avalanche photodiode (APD) detector, and total internal reflection or similar microscope, although continuous technological innovations have made these simpler and more affordable. SMD has been applied recently to detecting the cystic fibrosis transmembrane conductance regulator (CFTR) localized to the erythrocyte plasma membrane (206). In this case, a novel experimental approach that combined atomic force microscopy with quantum-dot–labeled anti-CFTR antibodies was employed to detect individual CFTR molecules. The results suggested that quantification of CFTR in a blood sample could be useful in the diagnosis of CFTR-related diseases. An important issue for every SMD experiment is whether immobilizing the analyte affects its function, and thus, monitoring in this state may reflect “unrealistic” data. The benefits of SMD include bypassing ensemble averaging completely, taking measurements from fixed “quantities” of analyte, and the ability to monitor intermediary reaction states that can be “masked” in an ensemble. For all pertinent issues, the reader is referred to References 1, 207, and 208 for a comparison of ensemble versus single-molecule FRET studies. The interested reader is referred to other pertinent articles on SMD of various analytes.

Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) is a technique that allows monitoring of single molecules but does not necessitate surface immobilization as above. At its most basic level, this process is a monitoring of fluorescent translational diffusion into/out of a minute volume defined by a focused laser and imaged with a confocal aperture (1). Diffusion continuously replenishes the analyzed molecules and their short transit times through the laser focal point to mitigate any photobleaching issues. The fluorescent bursts are collected and analyzed with the Stokes–Einstein equations to correlate their diffusion coefficients and thus provide insight into their size, rotational properties, concentration, and so on. FCS has found extensive applications in biochemical reaction monitoring and kinetic...
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

The authors realize that we have left out far more than we have included, and beyond our apologies for this, we hope that this article will serve rather to stimulate more reading and experimentation.

References

42. Waggoner A. Fluorescent labels for proteomics and genomics.

38. Dodeigne C, Thunus L, Lejeune R. Chemiluminescence as a

36. Murphy JT, Lagarias JC. The phytofluors: a new class of fluo-

34. Schaferling M, Nagl S. Optical technologies for the read out

30. Shu XK, Shaner NC, Yarbrough CA, Tsien RY, Remington SJ.


33. Glazer AN. Phycobiliproteins—a family of valuable, widely used

48. Miller I, Crawford J, Gianazza E. Protein stains for proteomic

45. Rojas B, Taylor P, Hater-Khalil M,类型的 muscle responses to

46. Panchuck-Voloshina N, Haugland RP, Bishop-Stewart J, Bhal-

54. Burns A, Ow H, Wiesner U. Fluorescent core-shell silica

57. Chen J, Yang X, He Z. A new class of chromosensor for DNA

47. Buschmann V, Weston KD, Sauer M. Spectroscopic study and

50. Soh N. Recent advances in fluorescent probes for the de-

58. Buck SM, Koo YEL, Park E, Xu H, Philbert MA, Brasuel MA,

55. Burns A, Ow H, Wiesner U. Fluorescent core-shell silica


51. Pataki J, Szabo M, Lantos E, Szekvolgyi L, Molnar M, Hegedus

52. Wang C, Zhang Y. A microsphere coupled micropatterning

60. Selvin PR. Principles and biophysical applications of lanthanide-

61. Bunzli JCG, Piguet C. Taking advantage of luminescent lan-

62. Hemmila I, Laitala V. Progress in lanthanides as luminescent

49. Soh N. Recent advances in fluorescent probes for the de-

43. Taitt CR, Ander son GP, Ligler FS. Evans wave fluorescence

44. Schaferling M, Naji S. Optical technologies for the read out

40. Michelini E, Guardigli M, Magliulo M, Mirasoli M, Roda A,

37. Fischer AJ, Rockwell NC, Jang AY, Ernst LA, Waggoner AS,

41. Roda A, Pasini P, Mirasoli M, Michelini E, Mirasoli M. Bio-

47. Buschmann V, Weston KD, Sauer M. Spectroscopic study and

39. Roda A, Pasini P, Mirasoli M, Michelini E, Mirasoli M. Bio-

53. Wang C, Zhang Y. A microsphere coupled micropatterning

56. Yamasaki Y, Takahashi S, Hara S, Shikauchi K, Murakami

57. Chen J, Yang X, He Z. A new class of chromosensor for DNA

58. Buck SM, Koo YEL, Park E, Xu H, Philbert MA, Brasuel MA,

62. Hemmila I, Laitala V. Progress in lanthanides as luminescent

40. Michelini E, Guardigli M, Magliulo M, Mirasoli M, Roda A,

55. Burns A, Ow H, Wiesner U. Fluorescent core-shell silica


31. Glazer AN. Phycobiliproteins—a family of valuable, widely used

26. Shaner NC, Steinbach PA, Tsien RY. A guide to choosing

24. Dooley CT, Dore TM, Hanson GT, Jackson WC, Remington


27. Baird GS, Zacharias DA, Tsien RY. Biochemistry, mutagenesis,


20. Day RN, Schaufele F. Imaging molecular interactions in living


18. Waggoner A. Fluorescent labeling and fluorescent spectroscopy:
overview of applications in chemical biology.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications, in Chemical Biology


Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology


140. Biophys. 1966;115:593.


144. Biophys. 1966;115:593.


146. Biophys. 1966;115:593.


Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications, in Chemical Biology
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology


187. Hariharan KG, Piron DJ, Miller LJ. Fluorescent indicators distributed throughout the pharmacophore of cholecystokinin provide insights into distinct modes of binding and activation of type A and B cholecystokinin receptors. J. Biol. Chem. 2006;281:27072–27080.


Further Reading

Fluorophore spectra: http://probes.invitrogen.com/resources/spectraviewer/.
Fluorescent Labeling and Fluorescent Spectroscopy: Overview of Applications in Chemical Biology

See Also

Chemical Labeling: Overview of Applications in Chemical Biology
Chemiluminescence and Bioluminescence Techniques
Fluorescence in Living Systems: Overview of Applications in Chemical Biology
Fluorescence Lifetime and Anisotropy Measurements: Proteins
Fluorescence Resonance Energy Transfer: Proteins
Fluorescence Techniques: Lipids
The rapid testing of chemical libraries for biological activity is the primary aim of high throughput screening (HTS). Advances in HTS have paralleled those in molecular biology, instrumentation and automation, and informatics, and the increased availability of arrayed compound libraries. Sophisticated high sensitivity assays and the associated technologies required to implement these assays in HTS have been largely developed within the pharmaceutical industry for the identification of new chemical matter for drug development. However, HTS approaches are now widely applied to the broader questions within biological research. By way of introduction, we will describe the components of HTS and provide examples of strategies used to identify novel chemtypes for specific biological targets using large chemical libraries. We then will illustrate how more narrowly defined compound collections (e.g., targeted libraries or bioactive compounds) have been profiled against related targets or cell types for the purpose of discovering or defining, compound class/gene family selectivity, off-target activity or "hidden phenotypes", toxic fingerprints, or any other relationship between chemical structure and bioactivity. In this way, HTS systems can expand the scope of an experimental hypothesis to address questions of chemical biology, be they at the level of an isolated enzymatic activity or that of a complex cellular phenotype.

High throughput screening (HTS) is a technologically enabled field of applied science that creates an interface between chemical libraries and biological assays for the purpose of rapidly exploring and identifying chemical bioactivity. The technologies that make HTS of large compound collections possible have been supported principally by the pharmaceutical industry because of the needs inherent to drug discovery but are increasingly being used to expand the boundaries of traditional academic disciplines such as enzymology and cell biology. "Chemical biology" that emphasizes the application of synthetic chemistry in the study of biological processes or "chemical genetics" in which compounds are used to recapitulate the effect of genetic mutations are particularly enabled by HTS as it allows the identification and profiling of wide-ranging chemotypes that modulate individual gene products or cellular phenotypes (1, 2). HTS laboratories are now an integral part of many universities, and in 2004 the U.S. National Institutes of Health (NIH) implemented a specific initiative to expand access to HTS for translational research (3–5). Whether the aim is chemical biology or drug discovery, HTS is most effective when biological assay systems are designed and screening results interpreted with the technological capabilities and limitations of the entire process in mind.

As a primer to this area, we outline the basic process of HTS and synergisms between pharmaceutical and chemical biology research endeavors and provide examples of platforms that use HTS technologies to enable chemical biology.
Bridging Chemistry and Biology with HTS

Compound discovery or profiling that uses HTS involves merging diverse types of chemical libraries and biological assays (6) (Fig. 1), and generally involves four major components that include the compound library itself, high-quality assays, process engineering platforms, and informatics procedures to track, analyze, and annotate the results. As a consequence, implementation of HTS requires expertise from multiple disciplines. In this section, we highlight some of the important features for each of these processes and how these technologies currently are being applied to the goals of chemical biology.

Chemical libraries

Multiple categories, including natural products (7) and synthetic bioactives (e.g., metabolites, carcinogens, and approved drugs (8)); privileged scaffold-based libraries (e.g., untested analogs of synthetic drugs or natural products such as benzodiazepines (9), indoloquinolizidine (10), or diketopiperazines (11)); biologically uncharacterized compounds of low diversity but high density (e.g., combinatorial chemistry-derived libraries (12, 13)); and consolidated samples/collections that represent extensive structural diversity (e.g., Molecular Libraries Small Molecular Repository, see PubChem (14)) have been used to describe the general character of library collections in use today (Fig. 1a). These libraries range in size from small focused or diverse sample collections of ∼1000 compounds or less to very large collections that may contain a million or more compounds. The cost can range from between a few dollars per mg for specialty sets of a few hundred compounds to ∼$1500 compounds or less to very large collections that may contain a million or more compounds. The cost can range from between a few dollars per mg for specialty sets of ∼1000 compounds or less to very large collections that may contain a million or more compounds. The cost can range from between a few dollars per mg for specialty sets of ∼1000 compounds or less to very large collections that may contain a million or more compounds.

The nature of the library is dependent on the aim of the experiment. For example, if an inhibitor of a novel protein kinase is desired, then a targeted library derived from the aminoquinazoline scaffold (31) (Fig. 1a, i) might be considered, whereas an enzyme that belongs to a gene family with no precedent of small-molecule modulation may benefit from a more random or diverse collection (32). Aiming for many dissimilar scaffolds (the central rigid part of the molecule capable of positioning key interactions) and analogs could lead rapidly to very large libraries, and, therefore, the complexity and the redundancy of the library needs to be carefully managed. The quality of the compounds should also be determined with respect to the stability, synthetic tractability, resupply, and purity of the samples. Various chemical descriptors can be used to measure library diversity; however, a universal method has not been found (21, 33). General screening compound collections aim to contain many different scaffolds with few analogs and are expanded in size by similarity searches using parameters such as the Tanimoto coefficient (34) to consider how new compounds may complement or add to the novelty of the collection. At the other extreme are combinatorial libraries that contain few scaffolds but potentially thousands of related analogs that systematically explore the activity relationships between two or more structural variables. Several thousand different combinatorial libraries have been described since 1992 (35, 36). Combinatorial libraries are synthesized using a variety of methods and can be very large in size (10^9), and although these libraries are not classically diverse, they provide the advantage of containing many subtle changes in structure that can be enormously important for biological activity (37). Knowledge around compound classes that target specific members of a gene family, as in the case of protein kinases or G-protein-coupled receptors (GPCRs), can form the basis of targeted libraries that can increase the probability of finding useful leads (12, 38). More specialized libraries are often required to address certain target classes, such as protein-protein interactions that have been historically refractory to conventional small-molecule approaches. For example, fragment library approaches in which small scaffolds (MW 150–300) are screened at very high concentration to identify weak interactions using nuclear magnetic resonance (NMR)- or X-ray crystallography-based methods have been successful for these difficult target classes (16). Once active fragments are found, the affinity is improved by merging or growing fragments into nearby sites on the protein through the use of structure-guided techniques.

Chemical library handling

Collection sample preparation, formatting for screening (39), short- and long-term storage, and analytical quality control (QC) (40) comprise important variables in the management of HTS libraries. Generally, libraries are arrayed with a single compound per well, but strategies that employ pooling of compounds have been described to compensate for limited screening capability (41). For the purpose of HTS, compounds are routinely dissolved in DMSO, a so-called "universal solvent" at high (typically 10 mM) concentrations, and stored in polypropylene microtiter plates. Critical to the "stability" of these collections is the control of humidity and temperature. DMSO is highly hygroscopic, and when it absorbs water, because of nonideal behavior, it becomes more structured and viscous and limits compound solubility, which can lead to compound precipitation (42). That precipitation, rather than compound degradation, is the central factor determining library integrity has been supported by studies on modern pharmaceutical screening collections (40, 43, 44). Therefore, refined compound storage units use desiccated inert chambers (e.g., argon saturated environments) and dry conditions. If dry conditions cannot be maintained, then the compound collection can be stored at room temperature (never...
High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology

Figure 1  Chemistry and biology. a) Categories of compounds. i, ii) Representative chemical scaffolds that could be present either as part of a diverse set of compounds (i) or as a combinatorial library template, illustrated with the 4-amino-quinazoline scaffold (ii). Here a “split and mix” strategy is used in which the library is synthesized on solid support beads that contain the scaffold. Chemical side chains are added to the scaffold by reacting building blocks (or synthons) in separate batches and then pooling all the batches and dividing the beads equally for the next round of synthesis. In the example shown, 10 unique synthons are added at each of three positions (X, Y, and Z) to create a library of $10 \times 10 = 1000$ different 4-amino-quinazolines in three steps. iii) Compounds that represent known drugs. Shown are anti-inflammatory compounds (aspirin and celecoxib) and the HMG-CoA reductase inhibitor, Atorvastatin. (iv) Natural product libraries represented by trichostatin A, forskolin, or a compound from a natural-product-derived synthetic library (Boston University). b) Categories of different assay methodologies based on purified (i, ii) or partially purified (iii) targets and intact cellular systems (iv). Examples of purified molecular targets (i) or partially purified (ii) targets and intact cellular systems (iv). i) Examples of purified molecular targets (i) or partially purified (ii) targets and intact cellular systems (iv). Examples include, enzyme catalyzes turnover of a profluorescent red emitting substrate (i) or a substrate/product (e.g., ATP) of the target enzymatic reaction catalyzed, for example, by an ATPase (ii) is detected by the “reporter” enzyme firefly luciferase via luminescent (iv).
Table 1: Example Chemical Libraries Used in Drug Discovery and Chemical Biology

<table>
<thead>
<tr>
<th>Library</th>
<th>Category</th>
<th>Size</th>
<th>Comments</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigma LOPAC</td>
<td>Pharmacologically active.</td>
<td>1208</td>
<td>Major target classes: GPCRs, kinases, ion channels, nuclear receptors, metabolic enzymes, cell signaling, apoptosis, cell cycle, etc. Used extensively for assay/HTS validation.</td>
<td>*</td>
</tr>
<tr>
<td>ChemBridge Fragment Set</td>
<td>Low MW (≤ 300), and cLogP (≤ 3) designed for high aqueous solubility (≤ 3mM).</td>
<td>5000</td>
<td>Useful for high concentration screening (e.g., NMR/SPR-based approaches).</td>
<td>16</td>
</tr>
<tr>
<td>TimTech Natural compound library (NPL400)</td>
<td>Purified natural products.</td>
<td>480</td>
<td>Structures retrievable from website. Mainly plant derived, ~16% flavonoids.</td>
<td>17</td>
</tr>
<tr>
<td>Angular epoxyquinol library a</td>
<td>Diversity oriented synthesis (DOS).</td>
<td>244</td>
<td>Chemical Methodology and Library Development (CMLD) initiative to develop novel libraries.</td>
<td>18, 19</td>
</tr>
<tr>
<td>National Toxicology Program (NTP1408)</td>
<td>Toxic agents. Includes compounds tested in traditional in vivo and in vitro toxicologic assays.</td>
<td>1408</td>
<td>Chemical descriptions are publicly available in PubChem.</td>
<td>20</td>
</tr>
<tr>
<td>NIH Molecular Libraries Small Molecule Repository (ML SMR)</td>
<td>Diverse collection: procured, QC-ed, stored and distributed to 10 network labs.</td>
<td>&gt; 250 K</td>
<td>Used widely in academia.</td>
<td>21</td>
</tr>
<tr>
<td>Pfizer compound file</td>
<td>Large pharma collection.</td>
<td>&gt; 2 × 10⁶</td>
<td>~U.S. $1 billion file enrichment program to expand diversity, drug likeness, synthetic tractability, and rapid analog access.</td>
<td>15</td>
</tr>
</tbody>
</table>

*must consider the nature of potential assay liabilities (e.g., sensitivity to compound library fluorescence) and match counter or orthogonal assays (e.g., an assay that is insensitive to compound fluorescence) to stringently and accurately verify or eliminate the primary assay findings. Ideally, a counterscreen is integrated into the primary HTS either as multiple measurements from the same assay well or as independent assays run in close proximity. As an example of the latter, a 13536-well robotic screen used fluorescent polarization to measure the interaction between the C-terminus of the tumor suppressor protein BRCA1 and a phosphopeptide derived from an interacting protein pBACH. Here, two assays that differed only in the labeled phosphopeptide ligand, carrying either a green or red fluorescent probe, were interleaved into a single HTS (53). The experiment was designed to discriminate signal changes from fluorescent samples, likely to affect the assay pair differentially, from genuine inhibitors of the protein-phosphopeptide complex that should have an unbiased effect on the assay. For cell-based assays, dual-luciferase reporter systems, for example, where one luciferase is constitutively expressed to track cell number, have been used to segregate compounds that modulate the biology in question from nonspecific effects on the assay format or cytotoxicity (54, 55).
High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology

Figure 2  Chemical probes versus drugs. Structures across the top are a) monastrol (25), an inhibitor of the kinesin Eg5 identified through forward chemical genomics, b) an activity-based chemical probe (26) used to label cellular metalloproteases, c) the natural product fumagillin (27) from Aspergillus fumigatus that acts as an angiogenesis inhibitor through inhibition of a methionine aminopeptidase, d) CKD-731, a semisynthetic analog of fumagillin currently in clinical trials as an anticancer agent, e) LG335 (i, 28), an inactive analog of the drug Targretin® (ii, Bexarotene), a pharmaceutical developed to bind RXR for the treatment of cutaneous T-cell lymphoma, paired with a designer RXR nuclear receptor to enable conditional gene expression (29), f) the approved HMG-CoA reductase inhibitor Atorvastatin (8), g) the natural product toxin Taxol used as an anticancer agent, h) structures of the opioid analgesics fentanyl—(i) (8) and 1-methyl-4-phenyl-4-propionoxypiperidine, (MPPP, ii) and its metabolite 1-methyl 4-phenyl 1,2,3,6-tetrahydropyridine (MPTP, iii) that is neurotoxic resulting in Parkinson disease-like symptoms (30)—and i) a polychlorinated dibenzodioxin (TCDD) that represents an environmental pollutant that can reach toxic levels by accumulating in fatty tissues.

Assay technologies for modern drug discovery have been fueled by achievements in molecular biology and the numerous genome sequencing projects that have lead to the cloning and expression of novel proteins, highly engineered cell lines (54, 56, 57), and model organisms (58, 59) with convenient reporters such as luciferases (55), fluorescent proteins (57), or enzymes that act on fluorescent substrates such as β-lactamase (60). In current drug discovery, HTS involves a variety of assay types (Fig. 1b) based on purified molecular targets, reconstituted enzyme cascades, cell extracts (61), and cellular/organism phenotypes (62). The use of reconstituted systems or extracts is particularly attractive as multiple molecular targets/interactions can be tested without the complication of the test compounds binding to serum, having limited cell penetration, or undergoing rapid cellular efflux. Also, such systems can be used to identify the target through techniques analogous to genetic suppression (63).

Decisions surrounding specific HTS assay technologies will take into consideration factors including pharmacological relevance, HTS compatibility, follow-up strategy, and costs, among others. For example, whether a membrane receptor antagonist is sought using a SPA-based radioligand binding assay with an enriched membrane preparation or through a cell-based reporter gene assay may depend on the availability of a suitable radioligand and ability to prepare sufficient plasma membrane versus developing a stable cell-based assay responsive to receptor activation. Also, in this example, a consideration of the cost of assay implementation including handling/disposal of radioactive versus the counter or orthogonal assays required to sort out the often multiple possible mechanisms of action from reporter gene responses are critical evaluation criteria. In a recent review (64), we provided detailed information and references on technologies used in HTS assays and the considerations in assay protocol design and reporting (65) to aid in evaluating the options for HTS assay development and communication of protocols and results.

Optimizing assays for HTS performance involves several steps including reducing the number of required reagent additions; assessing QC and availability of the reagents; determining
the stability and batch variability testing of the reagents; determining the DMSO tolerance of the assay (particularly important for cell-based assays); minimizing incubation and measurement times; and calibrating and optimizing detector settings. Therefore, an HTS assay can differ greatly from typical laboratory assays, and not all assays that function adequately in bench-top experiments will be adaptable to HTS. Typically, libraries are screened at one concentration in a manner in which every compound is assayed once. This "n = 1" experiment is rapidly performed on as many as 1 million samples, and, therefore, the assay must show excellent precision; but this does not guarantee nor should it be confused with biological fidelity (66). Therefore, during HTS implementation, the entire procedure is validated with test plates to assess the assay precision. The goal is to minimize variation and provide adequate signal:background (S:B) with-out adversely effecting the sensitivity or biological relevance of the assay. A measure of the assay performance or quality is typ-ically made using a parameter, the Z-factor (67, 68) (Table 2) (67-71), that is a measure of both assay variation and S:B in which assays that exhibit Z-factors > 0.5 are generally consid-ered acceptable. The background can be taken from specified control wells (Z') or the median of the sample field (Z).

To obtain statistical significance, these parameters should only be reported from entire microtiter plates (e.g., 96-wells or higher densities) and are best measured by performing the experiment in triplicate on different days. Also, the accuracy of the assay response in reflecting the biology under investi-gation can be measured when control compounds are present. IC50/EC50 can be compared with accepted values typically ob-tained from the literature and should include a consideration of the assay precision. For this purpose, the minimum-significant ratio can be calculated from control titrations across individual assay plates, which provides the smallest potency ratio between two compounds that can be considered real based on the preci-sion of the assay (72).

**Engineering processes in HTS**

To achieve a "high-throughput" process requires that systems are available to test compounds rapidly by processing of 96-well or higher density plates (73). The throughput of HTS can vary from low (10,000-50,000 data points/day) to ultrahigh (uHTS) where > 100,000 data points are collected/day. Efficiency in HTS requires highly engineered systems and components, such as high-density microtiter plates, precision, low-volume liquid handlers for dispensing assay reagents (nl to l) or concen-trated compound solutions (pl to nl) (74-76), microtiter plate readers varying in sophistication from reporting changes in a well's total absorbance to the subcellular distribution of target proteins within a cell (77, 78), and robotic systems to inte-grate these components (Fig. 3). Fully integrated systems that leverage miniaturized parallel assay processing (e.g., 1536-well plates) served by robotic arms having direct access to compound libraries, reagent/compound dispensers, and plate incubators and readers can achieve throughputs as fast as 10 samples/sec (79, 80). Operating within a thin margin for error, automated assays by virtue of their speed and parallelism should be validated carefully on the robotic system, and the entire process of as-say implementation should be managed to minimize costly loss of reagents that can occur quickly if the screen is improperly initiated and monitored (73).

**HTS data analysis and informatics**

Several independent streams of data are tracked and integrated throughout the HTS lifecycle. The library samples' structures, plate locations, and array positions, concentrations, and other parameters and annotations comprise one group; a second group is the detector output readings for the assay responses, and a third group include process information, such as time stamps and trace files that track compound and assay plate histories. For smaller-scale screens (a few 1000 compounds) Excel spreadsheets and an Access database can suffice, but as the scale of the HTS increases, organizing this large amount of chemical and biological assay data requires efficient laboratory informa-tion management systems (LIMS) that cover and provide tools to facilitate the entire experimental process. Flexibility has been cited as one of the primary requirements in LIMS design as both the processes and nature of the information content will change over time (81).

HTS assay data is statistically analyzed to identify genuine outliers from systematic or random assay noise, which allows the level of activity for each compound tested in the assay to be determined. The values used in the final analysis are derived from a normalization procedure that converts the raw data to the percent activity relative to the control values (Table 2). Wells that contain controls can be placed in any number of wells on an assay plate, but a common practice is to use ≤ 10% of the plate for controls, often in columns along the left or right edges (Fig. 4a).

A threshold cut off is the simplest method to select putatively active compounds (Table 2). For example, in Fig. 4, samples to the left of (Fig. 4b) the 3 standard deviation (σ) threshold or below it (Fig. 4c) are considered "active" (red values; often referred to as "hits") and are selected for "confirmation" or prioritized with additional criteria (e.g., clogP < 4 or availability of analogs). However, several problems have been noted with this method (70). For example, the controls may be unstable or, in the case of uncharacterized biology, may be absent entirely, which makes it difficult to derive accurate normalized percent activity values. Therefore, methods such as the z-score (69) have been used where the test samples act as the controls (Table 2).

Several methods have been employed to correct for positional variations within the screening data that can be very common in HTS and include edge effects, dispensing artifacts, and changes in reagent stability during the course of the screen. Correction of systematic errors requires special software and is facilitated by incorporating blank plates (e.g., containing DMEM alone without test compounds) uniformly throughout the screen to capture a sample-independent signature of the plate signal variation (Fig. 4). One method that considers positional effects uses the b-score (for "better" score) that is analogous to...
### Goals of chemical biology

The goals of chemical biology have been stated in various ways. In the “gene-centric” view, the aims are borrowed from classical genetic studies in which genes were mutated to determine their role in phenotypes. In this analogy, small molecules rather than gene mutations are used to affect protein function and perturb the biological system. Terms have been coined such as “forward chemical genomics” to represent phenotypic or response assays that are used to identify compounds that generally affect a cellular or model organism phenotype (Fig. 8). However, no single technology is sufficient to identify a target for a small molecule from such phenotypic assays, and target identification remains a challenging area in chemical genomics. In “reverse chemical genomics,” the starting point is often based on an assay that uses an isolated protein, and with subsequently identified compounds, one backs away from the “gene product” to determine the effect on phenotype. In this respect, classical drug discovery resembled forward chemical genomics in which isolated compound testing was often conducted in physiological model systems (e.g., organ bath), whereas modern drug discovery follows a reverse chemical genomics paradigm in which isolated molecular target-based assays are used for compound discovery, and these then are progressed into disease models and eventually to testing in humans. It is now widely appreciated that the proteome is considerably larger than the sum of its genes, and defining what constitutes a “gene” has become increasingly complex (84). Mechanisms such as alternative splicing result in an average of three protein isoforms/gene leading to a larger and more complex proteome. In this “proteome view,” chemical biology has the goal of annotating every protein with compounds capable of inhibiting, activating, or allosterically regulating their function. Ultimately, both the gene-centric and the proteomic viewpoints converge to create a pharmacological database that relates compound activity to biological effects (85). This “chemical probe directory” will list compounds useful for the interrogation of basic biological questions and initiation points to develop novel therapeutics.

---

**Table 2** Assay Diagnostic and Hit Scoring Methods Used in HTS

<table>
<thead>
<tr>
<th>Equation</th>
<th>Explanation</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-factor</td>
<td>Factor used to evaluate assay performance where a value &gt; 0.5 represents an acceptable assay.</td>
<td>67, 68</td>
</tr>
<tr>
<td>Scoring Methods</td>
<td>Control-based. Normal distribution assumed.</td>
<td></td>
</tr>
<tr>
<td>%Inhibition</td>
<td>( \mu_{\text{sample}} - \mu_{\text{control}} ) ( \mu_{\text{max}} - \mu_{\text{control}} )</td>
<td></td>
</tr>
<tr>
<td>Threshold value (T)</td>
<td>( T = k \cdot \mu_{\text{max}} )</td>
<td></td>
</tr>
<tr>
<td>z-score</td>
<td>Method to score the relative potency of actives on a plate-by-plate basis without correction for row and column effects. Controls not used. Normal distribution assumed.</td>
<td>69</td>
</tr>
<tr>
<td>b-score</td>
<td>( y_{ijp} = \mu_{ijp} + R_i + C_j + \epsilon_{ijp} )</td>
<td>70</td>
</tr>
<tr>
<td>r-score</td>
<td>( y_{ijp} = \mu_{ijp} + R_i + C_j + \epsilon_{ijp} )</td>
<td>71</td>
</tr>
</tbody>
</table>

The z-score except positional and plate effects are additionally taken into account and result in a marked reduction in false positives (Table 2) (70). However, such methods can fail in screening-focused collections, such as combinatorial libraries in which many analogs around a common scaffold are present, because the assumption that the majority of the tested wells are inactive may not hold when many genuine actives occur on the same plate. Therefore, for combinatorial libraries, the activity is better addressed using multiple assays to help rank or profile the actives or screening at multiple concentrations where potencies can be resolved (see below, Fig. 4d). More attempts to improve active selection from large HTS data sets include the use of structure-activity relationships (SAR) (82) to help reduce false negatives and advanced population analysis to construct predictive models of activity (83).
High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology

Figure 3  Dispensing and detection instrumentation. a) Example of liquid handling instrumentation. Top: Piezo-electric tip plumbing and pulse control for pL dispensing of liquids. Bottom: Plumbing and valve control of a solenoid-based for nL dispenser. b) Methods for nL dispensing of compound solutions in DMSO (beige wells) to assay plate wells (blue wells) using contact-based dispensing with pin-tools (i) that may contain slots as shown or noncontact-based dispensing as in using a focused acoustic beam (ii). For acoustic dispensing, it is necessary for the destination plate to be inverted over the source plate. c) Reader modalities (i–iii) are depicted for population averaged detection, (i) transmittance using microtiter plates made of clear plastic (e.g., polystyrene or cyclic olefin polymers), (ii) luminescence, typically using opaque white microtiter plates, and (iii) fluorescence where in the example shown epifluorescence is collected by using focusing optics for excitation (green arrow) and emission (red arrow) above the well of an opaque black microtiter plate. d) Detection output (i, ii) for single object enumeration or imaging. (i) Cell cytometry in HTS is possible using either flow or microtiter plate systems (78). Here, the population distribution of green and blue fluorescent cells in a β-lactamase reporter gene assay is analyzed. (ii) High-content screening using automated wide-field or confocal microscopes. High-resolution confocal image of G2/M cell cycle sensor (GE Healthcare) expressed in U2OS cells. Actin and microtubular staining using Texas red- and Alex488-labeled antibodies, respectively, and nuclei stained with DAPI.

Furthermore, as the pharmacological database becomes populated with diverse biological assays and grows in structural scope, it will act to guide the design of higher-quality libraries that will access the darkest recess of the genome.

Chemical probes

To understand the nature of chemical probes, we can evaluate these probes relative to the current drug discovery process. Although the goals of drug discovery have a clear end point, the development of a therapeutic, the goals of chemical biology are broader so that although all drugs are probes because of the special requirements of drugs, not all probes are intended or can be developed into drugs (Fig. 2). Chemical probes can be drawn from the embryonic stages of the drug development pipeline, for example, in the lead optimization stage, once activity has been confirmed and compounds of sufficient potency have been obtained. Early stage absorption, distribution, metabolism, elimination (ADME)-toxicity data such as plasma serum binding, cytochrome P450 inhibition/induction, and cell permeability data can be obtained around chemical probes to determine their usefulness for in vivo testing. Although cell permeability and low serum binding are important, fully optimized ADME-toxicity parameters are not a requirement for many chemical probes. For example, the natural product forskolin (Fig. 1a, iv), an activator of adenylyl cyclase, has been used extensively and primarily in the study of Gi-coupled or "inhibitory" GPCRs in many cellular model systems, with no need to consider ADME-toxicity data around such a tool compound. Furthermore, although selectivity is an important consideration for drug development and many tool compounds, a pan-inhibitor such as staurosporine that is broadly active against the kinase (86) has proven useful in the validation of assays for protein kinases. Also, some pan-inhibitors have been used to develop generic protein kinase assays (87). However, the hydroxamate-containing antifungal antibiotic trichostatin A (TSA; Fig. 1a, iv) has been used as an investigational tool to study histone-deacetylases (88) (more than 2000 publications have resulted from the use of TSA since the initial discovery in 1976) and as a proof-of-principle compound in the validation of novel cancer drug targets. Also, drugs have been used as probes to uncover new targets as in the case of the identification of cyclooxygenase-3 (COX3), a splice isoform of COX-1, through the investigation of the pharmacology associated with analgesic/antipyretic drugs (e.g., acetaminophen and coxibs, Figure 1a, iii) (89).

Academic chemical biology and pharmaceutical drug discovery synergies

Finally, it is worth noting where chemical biology can support the discovery of novel therapeutics. For example, many
High Throughput Screening (HTS) Techniques: Overview of Applications in Chemical Biology

Figure 4  Data processing and Informatics. a) Heat maps depicting 1536-well plate activity for a library of 1408 compounds that was screened from low to high concentrations using qHTS (black bar indicates increasing concentration) in triplicate with each plate containing a different compound concentration. Controls are contained in the left four columns (128 wells) and used to obtain normalized activity from the raw data. Background signatures are derived from plate sample wells containing DMSO/assay buffer alone (not shown) and used to correct for systematic artifacts before scoring and selection of “actives.” The final corrected activity shows actives increasing with concentration (blue: inhibitory activity in well, red: activating activity in well, and white: inactive). The background signature shown (striped pattern) indicates a systematic artifact because of liquid dispenser tip variability. This background signature was subtracted from the normalized data to yield the final corrected activity. b) Traditional HTS is typically conducted at a single concentration (e.g., 10 µM), and treatment of this data is shown. Distribution of activity at one concentration using the highest tested concentration of the qHTS data. A 3σ threshold differentiates the active (red) and inactive (blue) data sets. c) Scatterplot representation of the single concentration-based data set. Example CRC data sets are shown in (d) for a single compound that showed a concentration–response relationship or where (e) multiple CRCs are obtained through the use of qHTS.

rare genetic and infectious diseases are not currently the focus of the pharmaceutical industry but provide fertile ground for public–private partnerships that use chemical biology as the vector to initiate and coordinate programs around these disease areas. The unprecedented open access to screening data enabled by the National Center for Biotechnology Information that manages PubChem (14), where results are found from a diversity of bioassays tested against a growing compound library, will provide new avenues and opportunities for drug discovery and development. Such initiatives have now made a large amount

Figure 5  Reverse and Forward Drug Discovery/Chemical Biology. Reverse chemical genomics starts with a target-small-molecule interaction and progresses to determine the pharmacologic effect on an organism’s phenotype, whereas forward chemical genomics begins with a chemically induced phenotype and aims to identify the target(s) responsible for the phenotypic effect. HTS is enabling to both processes.
of chemical information available on the worldwide web (90). The open access model also brings together laboratories that possess the proper expertise both in the know-how of HTS and the required knowledge of the biology under investigation that is critical to advance compounds in an efficient manner. Recent broad access to HTS and chemical screening data should act to alleviate bottlenecks in our knowledge around new target classes and their potential modulation by small molecules and possibly lower the investment for development by industry.

Example HTS Platforms that Support Research in Chemical Biology

HTS can serve the needs of chemical biology by functioning in the traditional mode of “mining” large chemical libraries with individual assays to identify specific chemical matter for development (91, 92) or “profiling” a specific small ensemble of compounds with bioassay panels (88, 93) or combining both to populate a pharmacological database (94). In this section, we will touch on specific examples from the literature on the profiling of annotated chemical libraries and show the power automation can bring to increase the scope of these experimental designs. Lastly, we discuss several areas of chemical biology that have been beneficiaries of HTS-derived leads, which, with tighter integration to HTS, will enhance the value of new chemical probes.

Profiling assays for understanding mechanism of action

Defining the spectrum of activity for a collection of compounds can be achieved by profiling bioassay panels focused, for example, on gene families (86) or distributed across diverse signaling pathways (95). Bioactivity profiling can be most efficiently performed when the assays share a common format such as a luciferase-reporter system to streamline assay implementation, results quantification, and interpretation. Also, with format homogeneity, nonspecific effects across the profile can be readily identified. As an example, we describe experiments using the cell-based protein complementation assay (PCA) technology (96, 97). In this system, cells are engineered to express two interacting proteins where each protein is fused to complementary fragments of a split yellow fluorescent protein (EYFP). Interaction of the two protein fragments yields a reconstituted EYFP whose signal is at least 10 times brighter than either fragment alone. This system has been used to place cellular “sentinels” along points in signaling pathways where modulation of the sentinel complex reports on a compound’s effect in that pathway. In the work of Macdonald and colleagues (95), a collection of 107 drugs from six therapeutic classes was screened with 49 PCA assays and provided 127 different concentration–response curves (CRCs) against 35 activated protein kinases. Drugs with similar mechanisms of action revealed similar cellular response profiles, for example, PPARγ agonists induced PPARγ/SRC-1 complexes, and differences in profile signatures often attributable to chemical structure. Additionally, the hierarchical clustering of compound activity exposed a supercluster of drugs that did not share any common therapeutic target or mechanism of action but showed a “hidden phenotype” of antiproliferative activity. Therefore, pharmacological profiling enabled with assays such as PCA can be used at an early stage of chemical characterization or optimization to identify such hidden phenotypes, understand off-target activity, or decipher the mechanism of action.

Pharmacologically defined libraries and their uses

Drugs and bioactives from compound collections, such as the Library of Pharmacology Active Compounds (LOPAC, Sigma-Aldrich; Table I), or chemical probe collections available from vendors, such as Tocris, represent libraries that are annotated with biological activity. These types of libraries often show more activity across a broader range of HTS assays compared with nonbiologically biased libraries (e.g., based on synths and scaffolds selected for synthetic tractibility alone) and therefore are useful in identifying controls for assay development and validation. A noted advantage can be used to generate a hypothesis around the mechanism of action for a given response phenotype (98). However, the reliability of the annotation, often derived from multiple databases or fragmented literature sources, can be misleading and should be “triangulated” or corroborated with structurally distinct chemical classes. A group at Amphora, Inc. has approached this problem by employing microfluidic technologies to construct an annotated database in-house by measuring IC₅₀ for 88 protein kinases against a set of 1330 compounds (99). Collecting and improving the annotation of large compound collections against a wide range of biological targets is an active area of research in cheminformatics (100), but ultimately, the activity profile of library molecules will be confirmed experimentally owing to the advances in HTS such as those described below.

Automation of compound profiling

Once an assay platform configurable with highly engineered HTS robotic systems is developed, large-scale compound profiling becomes possible. An example of such an experiment was described by scientists at the Genomics Institute of the Novartis Research Foundation where automated cell culturing through dispensing in either 384- or 1536-well plates was used to test 1400 small-molecule kinase inhibitors to generate concentration–response curves (CRCs) against 35 activated tyrosine-kinase-dependent cellular assays (80). Similar to the PCA system described above, the results of these screens were used to cluster the kinase inhibitors based on the phenotypes observed in the cell-based assays. At the NIH Chemical Genomics Center (NCCG), we have leveraged the advances in HTS technology to determine concentration–response relationships of large chemical libraries routinely across diverse biological systems, a process called quantitative HTS or qHTS (94). To date, >5 million CRCs have been generated in >100 unique assays within 3 years of operation (see PubChem [141] using qHTS at the NCGC. The relatively
short timeframe under which such large-scale experiments can be conducted has enabled the needs of scientists with diverse interests and backgrounds, for example, scientists searching for antiapoptotic agents (79, 101), comparing experimental with in silico screening results (102), and testing hierarchical clustering algorithms for cheminformatics (103). The large data sets achievable from methodologies like HTS will be indispensable to progress in chemical biology and often will serve as the only means by which a foothold at the interface of chemistry and biology can be made.

Chemoproteomic methods: profiling gene families and downstream uses for chemical probes

For investigating the function of enzyme gene families within cells, activity-based probes (ABPs) are particularly useful (104). Requirements for an ABP include sufficient affinity (< 100 nM), pan-selectivity for the enzyme family, an expressed active enzyme to modify covalently, and a reporter tag such as a fluorophore for detection. In a recent exploration of metalloproteases, ABPs were synthesized around two alkyne-tagged hydroxamate-benzophenone libraries, and metalloproteases that bound the probes could be covalently labeled through the benzophenone group (Fig. 2b) (26). The alkyne was placed distal to the metal-binding hydroxamate moiety and readily coupled via click-chemistry with an azido-containing tag (e.g., rhodamine or biotin). The two libraries identified metalloproteases from nearly all branches of this enzyme superfamily, and the authors used the library to detect differences in metalloprotease activity from invasive and noninvasive melanoma cells.

Another method that enables large-scale profiling is the development of generic platforms for entire gene families. Researchers at Aribit have developed a competitive-binding assay that involves the expression of human kinases as fusions to T7 bacteriophage and a set of adenosine triphosphate (ATP) competitive pan-inhibitors tagged with biotin that enable selectivity profiles of kinase inhibitors to be determined on hundreds of protein kinases (more than 250 now available) (86). This format shows exquisite sensitivity with the ability to measure binding affinities as low as 1 pM.

Researchers at Sereis (now part of Pfizer) have turned the compound discovery process on its head using affinity chromatography to "proteome mine" and asking what proteins bind to the compounds rather than searching the library against a single protein (105). In this system, a chemical scaffold such as purine is tethered to a solid support, and a cell, tissue, or organ extract is passed over the surface so that the purine-binding proteins are specifically captured by the affinity resin. Then, library members are added to displace (specifically elute) the targets competitively at single or multiple concentrations, and the proteins that bind to these targets are identified by chromatographic separation and MS analysis. Some considerations in these affinity-based panels are that the affinity of the interaction must be high (Kd < 100 nM), the targets must have relatively good abundance (> 100 copies/cell with 10^6 cells typically harvested), and the identification of specific binding requires optimization of the washing protocols.

Ongoing HTS advancement includes systems that more closely replicate physiological context in the assay format. For example, techniques in which cells are grown under 3-D culture conditions better mimic the biological and pharmacological consequence of the response after compound treatment. An excellent review using 3-D culture of a multicellular tumor spheroid model was recently published (106). The use of 3-D cultures in these tumor models recapitulates the morphological, functional, and mass transport characteristics of the tumor tissue in vivo (107). Such 3-D cultures are also being used in high-throughput toxicity screening and in the analysis of their cytochrome P450-generated metabolites (108). Simple metazoan model organisms (e.g., Caenorhabditis elegans and zebrafish) offer the opportunity to identify substances that modulate complex physiological systems through HTS (58, 59). As these organisms are suitable to both reverse and forward genetics, screens based on target-directed or phenotypic assays are possible. Furthermore, rapid organogenesis (hours to a few days) in these metazoans allows, for example, developmental neurotoxicity and teratogenicity to be studied on the timescales compatible with HTS (6).

Biophysical and analytical measurements adapted to HTS systems have broadened the sophistication of assays formats, as witnessed by the rapid evolution of electrophysiological assays to measure ion channel function (109). The high sensitivity and multiparametric output inherent in flow cytometry (110) has also been applied to HTS and provides a platform for assay multiplexing that uses either beads or intact cells (111). For example, in one study, a three laser flow cytometry system was used to enumerate over 15 different parameters from phosphorylation events within multiple cell types including primary cells to define both pathway and cell-type-specific compounds (112). Such technologies enable the simultaneous correlated measurements of signal transduction networks that provide physiological relevant compound mechanism and selectivity data. Other methods are being developed, based on principles of microcalorimetry, that will enable the mass screening of compounds against targets with unknown function or ligands (113). These advances are but a few of the biological and technological advances that are impacting the future direction and uses of HTS in chemical biology.

Summary and Future Perspective

HTS developed rapidly in the pharmaceutical sector after the molecular biology revolution changed the drug discovery paradigm from a forward to a reverse genetics model (Fig. 3). Technological breakthroughs that allow sensitive in vitro assay designs (64) combined with the explosion in the availability and quality of chemical libraries (Table 1) accelerated the evolution and need for efficient HTS (114). Broad commercialization of HTS platforms and components coupled with an emigration of many "discovery" scientists from industry to academia in recent years has created a unique environment for HTS technology and expertise to enable academic chemical biology. Within a setting less restrained by market pressures, scientists will continue to find novel uses for the experimental power of HTS.
Breaching the ∼90% of the human genome not currently inter-
dicted by small molecules, particularly the “dark matter” of
the genome whose functions remain undefined, is a key challenge
for HTS-assisted chemical biology. The current drug pharma-
copedia targets only ∼330 proteins although as many as 8000
may be involved with disease (115). Aiding to this situation,
polymorphic variations between individuals will demand in-
creased intricacy in chemical genomics research that will fuel
the growing field of pharmacogenomics and ultimately usher
in the era of personalized medicine. Redirecting the fate of
pluripotent stem cells, the holy grail of regenerative medicine,
will likewise find an ally in HTS (116). Understanding the dis-
position and effects of small molecules on biological systems
through research in chemical biology should provide more op-
timal prediction of in vivo efficacy (64) and toxicity (117).
The new HTS approaches are poised to have a transformative
impact on biology and medicine, to remove many historical
limitations to experimental designs, and to bring the promise
of the genomic revolution closer to realization. Finally, the ad-
vent of synthetic biology, which spans from designer receptors
(118, 29) (e.g., Receptor Activated Solely by a Synthetic Lig-
and (RASSL) and Receptor Exclusively Activated by Designer
Drugs (DREADDs)) to genome transplantation (119) to future
man-made life forms, will require novel ligands and designer
drugs to complement and control (or keep in check) this new
biological frontier.

Acknowledgments
This work was supported by the Molecular Libraries Initiative
of the NIH Roadmap for Medical Research and the Intramural
Research Program of the National Human Genome Research
Institute, NIH. We thank Daryl Leja for the illustrations used in
this work.

References
1994;1:11.
2. Spring DR. Chemical genetics to chemical genomics: small
3. Austin CP, Brady LS, Insel TR, Collins FS. NIH Molecular
4. Lazo JS. Roadmap or roadkill: a pharmacologist’s analysis of the
NIH Roadmap for Medical Research and the Intramural
5. Gordon EJ. Small-molecule screening: it takes a village. ACS
A, Smith CS, Inglese J, Porter Cj, Tice RR, Austin CP. Com-
pound cytotoxicity profiling using quantitative high-throughput
8. Krier M, Briti G, Rognan D. Assessing the scaffold diversity of
F, et al. Key aspects of the Novartis compound collection
enhancement project for the compilation of a comprehensive
10. Lopilato CA, Lombardino F, Donnini RW, Ferey PJ. Experimental
and computational approaches to estimate solubility and perme-
bility in drug discovery and development settings. Adv. Drug
Oxidation protocols in chemical biology: small molecules offer big
2. Austin CP, Brady LS, Insel TR, Collins FS. NIH Molecular
3. Gordon EJ. Small-molecule screening: it takes a village. ACS
A, Smith CS, Inglese J, Porter Cj, Tice RR, Austin CP. Com-
pound cytotoxicity profiling using quantitative high-throughput
6. Krier M, Briti G, Rognan D. Assessing the scaffold diversity of
7. Jacoby E, Schuffenhauer A, Popov M, Azzarui K, Havili B,
F, et al. Key aspects of the Novartis compound collection
enhancement project for the compilation of a comprehensive
8. Lopilato CA, Lombardino F, Donnini RW, Ferey PJ. Experimental
and computational approaches to estimate solubility and perme-
bility in drug discovery and development settings. Adv. Drug
Oxidation protocols in chemical biology: small molecules offer big
2. Austin CP, Brady LS, Insel TR, Collins FS. NIH Molecular
3. Gordon EJ. Small-molecule screening: it takes a village. ACS
A, Smith CS, Inglese J, Porter Cj, Tice RR, Austin CP. Com-
pound cytotoxicity profiling using quantitative high-throughput


104. Nomanbhoy TK, Rosenthal J, Alban A, Burbaum JJ. In-
hibitor focusing: direct selection of drug targets from pro-
105. Hall SE. Chemoproteomics-driven drug discovery: addressing
of 3-D cultures for high-throughput screening: the multicellular
107. Reininger-Mack A, Thielcke H, Robboli AA. 3D-siobrid
108. Lee MY, Kunter RA, Sukumaran SM, Hogg MG, Clark DS,
Dondick JS. Three-dimensional cellular microarray for high-
109. Zheng W, Spencer RH, Kiss L. High-throughput assay technolo-
110. Robinson JP. Flow cytometry. In Encyclopedia of Biomaterials
111. Edwards BS, Oprea T, Prossnitz ER, Sklar LA. Flow cytometry
for high-throughput, high-content screening. Curr. Opin. Chem.
112. Krozik PD, Crane JM, Clutter MR, Rosier GP. High-content
single-cell drug screening with phosphospecific flow cytometry.
113. Vedat M, Nissen FF, Allal-Hassani A, Fodorov OY, Finerty PJ
Jr, Wasney GA, Yeung R, Arrowsmith C, Ball Lj, Berglund H, et
al. Chemical screening methods to identify ligands that promote
protein stability, protein crystallization, and structure determina-
114. Repple J. Expanding the HTS paradigm. Drug Discov. Today
115. Landry Y, Gies JP. Drugs and their molecular targets: an updated
EC, Ding S, Schultz PG. Reversine increases the plasticity
117. Caielli DI, Gray DM, Buccher JP. Toxicology. Transforming
118. Armbruster BN, Li X, Pausch M, Héritier S, Roth BL. Evolving
the lock to fit the key to create a family of G protein-coupled
119. Lalarte C, Gras JI, Alperovich N, Pieter R, Parmer PP,
Hutchison CA 3rd, Smith HD, Viénet JC. Genome transplanta-
tion in bacteria: changing one species to another. Science
Microarrays in Chemical Biology

Angela N. Koehler, Broad Institute of Harvard, Cambridge, Massachusetts, and MIT, Cambridge, Massachusetts

doi: 10.1002/9780470048672.wecb645

Advances in both genomics and proteomics have provided researchers with access to large collections of biomolecules, including DNA, proteins, and metabolites. High-throughput methods are needed to study the function and regulation of these biomolecules within complex systems. Microarrays have emerged as a common platform to study biomolecular interactions that involve nucleic acids, proteins, and small molecules. DNA microarrays have revolutionized genomic research by allowing researchers to study gene expression, sequence variation, and transcription factor binding sites on a whole-genome scale. Protein microarrays can be used to study interactions with other proteins, DNA, RNA, and small molecules, including lipids, carbohydrates, and drugs. Protein microarrays can also be used as analytical tools to profile complex protein mixtures, such as fractionated cell lysates, in an effort to study antibody specificity, measure changes in protein abundance, or characterize disease states. Small-molecule microarrays are useful tools for ligand discovery, comparing inhibitor specificity across enzyme classes, high-throughput, cell-based phenotypic screens, and as diagnostic tools for pathogen detection. The microarray approach has been extended to transfected cell microarrays, RNAi living cell microarrays, virus microarrays, and tissue microarrays. This article reviews chemical strategies for making microarrays and applications of microarrays in chemical biology.

As the molecular parts list for cells and organisms continues to grow, a key challenge is to identify the function of each component within the context of the cellular system. Increasingly, researchers are choosing to adopt high-throughput, discovery-oriented studies aimed at systems of biomolecules in addition to traditional reductionist studies of the structure or function of a specific biomolecule. Studies of complex biological systems benefit from large-scale and global analyses as the interacting parts can be analyzed simultaneously, which leads to hypotheses about how the components come together to effect processes such as development, disease, or evolution. Several high-throughput methods have been developed to study the function and regulation of biomolecules, including nucleic acids, proteins, and metabolites. In particular, microarray technology has emerged as a revolutionary platform to analyze all types of biomolecules in a highly parallel fashion.

Biologic Background

Microarrays were introduced in the 1990s for use in genomic studies (1–4). Microscopic features of probe nucleic acids, including oligonucleotides, cDNAs, or PCR products that correspond to predicted or known mRNAs, are deposited on a stable, solid substrate such as glass or silicon. Taking a cue from the Southern blot, the affixed probes are hybridized with cDNA from the sample of interest. Thousands of probes can be accommodated on a single microarray providing an opportunity to interrogate whole genomes (5). DNA microarrays, otherwise known as DNA chips, are valuable tools for gene expression profiling (6), comparative genomic hybridization (6), genotyping (7), chromatin immunoprecipitation (8), and in vitro studies of protein–DNA interactions (9, 10). DNA microarray technology has also had a significant impact on medicine by facilitating connections between physiologic states and gene expression patterns in studies of cancers, disease progression, and cellular responses to toxins or therapeutics (11). DNA microarray technology is now widely accessible to researchers through commercial microarray products and academic genome centers.
The miniaturized and highly parallel microarray format has also become a common platform in applications that involve other types of biomolecules (Table 1). Shortly after the advent of DNA microarrays, small molecules and proteins were captured in the microarray format (12, 13). Microarrays of carbohydrates (14, 15), lipids (16), peptides (17), cells (18), viruses (19), and tissues (20) have also been reported. Why has the microarray format become so attractive for studying diverse types of biomolecules? Assay miniaturization is a key advantage of microarray-based assays. Miniaturization reduces the amount of sample required for the assay and increases the density of samples that may be analyzed simultaneously. Microarrays are also easy to manufacture and store. Most types of microarrays are stable for up to six months under proper storage conditions. Benefiting from advances in automation, replicate arrays can be processed in parallel, which allows researchers an opportunity to evaluate thousands of probes across a variety of sample conditions simultaneously. These advantages make the microarray format attractive to researchers in chemical biology with an interest in studying responses of transcriptomes or proteomes to small molecules. Additionally, microarrays provide an attractive format to screen precious small molecules, such as natural products and compounds coming from combinatorial libraries, against collections of nucleic acids, proteins, or cells. Advances in microarray-related technologies, including the use of epoxide-coated slides and attachment via N-hydroxysuccinimide (NHS) activated esters, this approach obviates the need for additional coupling catalysts. Readily commercially available epoxide-coated glass slides have been used to capture hydrazide-tagged small molecules and carbohydrates (28, 29). Most surface capture methods take advantage of a reactive functional group that is introduced as part of their synthesis and biases the orientation of the small molecule on the surface. Motivated by the need to increase molecular diversity on SMMs, nonselective approaches to capturing compounds have been adopted. Bradner et al. (30) reported the use of an isocyanate-mediated capture strategy to print nearly 10,000 known bioactive small molecules, natural products, and small molecules originating from several diversity-oriented syntheses. Isocyanates react with a variety of nucleophilic functional groups, thereby increasing the number of small molecules, from natural or synthetic sources, that may be printed on a single surface. Kanoh et al. (31) prepared microarrays of approximately 2000 natural products and drugs by photo-cross-linking compounds on trifluoromethylaryldiazarine-coated surfaces. Using this approach, photogenerated carbenes react with the printed compounds in a manner that is independent of the functional group. Both the isocyanate and the photo-cross-linking strategies present the possibility of printed compounds that occupy multiple modes of orientation within a given spot, which effectively increases the number of binding modes that a given probe protein can sample. Several surface-capture methods applied to SMMs have also been used to prepare protein microarrays, including the use of epoxide-coated slides and attachment via Staudinger ligation (32, 33). A different attachment strategy has been developed in the context of fabricating protein microarrays and have been reviewed previously (34, 35). Commonly, aldehyde-coated and epoxy-coated slides have been used to capture covalently proteins in a heterogenous fashion by reacting with amino groups (13, 35). As mentioned, retaining the protein function is a key concern when using covalent chemistry to prepare microarrays, and thus, several groups have pursued noncovalent fabrication of functional protein microarrays.

### Strategies for Preparing Microarrays and Methods for Detecting Interactions

The first step of any microarray experiment involves design and fabrication of the chips containing probe molecules of interest. Immobilization methods should take both orientation of display and molecular stability into account. Stability is an especially important concern in the area of protein microarrays as many applications require that individual proteins retain their conformation and function (23). Most types of arrays are prepared by immobilizing the biomolecules on chemically treated glass microscope slides using either microcontact spotters or piezoelectric deposition. Microcontact lithography, using either masks or optical methods, is another common method for fabricating microarrays and is used routinely to prepare oligonucleotide arrays (3, 24). Several capture strategies have been developed for microarrays and include both covalent attachment of the probes to surface and noncovalent deposition (Table 2). Both approaches have proven useful in making arrays of nucleic acids, small molecules, and proteins.

### Covalent capture strategies

It is common for microarrays to include biomolecules that are attached covalently to a glass or silicon surface. For spotted microarrays, biomolecules are coupled using a variety of attachment chemistries (Table 2). Most advances in attachment chemistry have been driven by the preparation of small-molecule microarrays (SMMs) as small organic molecules tend to have more diverse functional groups than biopolymers. Many attachment chemistries have been reported and reviewed elsewhere (25, 26). Most of these approaches involve mild and selective coupling reactions. The first SMMs made use of a Michael addition reaction that involves molecules containing free thiols printed onto slides coated with vinyl sulfone or maleimide groups (12). Although this strategy proved successful, it was not general as most compound libraries do not contain a high proportion of free thiols. Many more compounds that come from both combinatorial libraries and natural product collections contain amino groups. Although these compounds may be coupled to carboxy-modified glass via amide bond formation, Chang and coworkers (27) chose to print a library of amine-containing molecules on slides coated with N-hydroxysuccinimide (NHS) activated esters. This approach obviates the need for additional coupling catalysts. Readily commercially available epoxide-coated glass slides have been used to capture hydrazide-tagged small molecules and carbohydrates (28, 29). Most surface capture methods take advantage of a reactive functional group that is introduced as part of their synthesis and biases the orientation of the small molecule on the surface. Motivated by the need to increase molecular diversity on SMMs, nonselective approaches to capturing compounds have been adopted. Bradner et al. (30) reported the use of an isocyanate-mediated capture strategy to print nearly 10,000 known bioactive small molecules, natural products, and small molecules originating from several diversity-oriented syntheses. Isocyanates react with a variety of nucleophilic functional groups, thereby increasing the number of small molecules, from natural or synthetic sources, that may be printed on a single surface. Kanoh et al. (31) prepared microarrays of approximately 2000 natural products and drugs by photo-cross-linking compounds on trifluoromethylaryldiazarine-coated surfaces. Using this approach, photogenerated carbenes react with the printed compounds in a manner that is independent of the functional group. Both the isocyanate and the photo-cross-linking strategies present the possibility of printed compounds that occupy multiple modes of orientation within a given spot, which effectively increases the number of binding modes that a given probe protein can sample. Several surface-capture methods applied to SMMs have also been used to prepare protein microarrays, including the use of epoxide-coated slides and attachment via Staudinger ligation (32, 33). A different attachment strategy has been developed in the context of fabricating protein microarrays and have been reviewed previously (34, 35). Commonly, aldehyde-coated and epoxy-coated slides have been used to capture covalently proteins in a heterogenous fashion by reacting with amino groups (13, 35). As mentioned, retaining the protein function is a key concern when using covalent chemistry to prepare microarrays, and thus, several groups have pursued noncovalent fabrication of functional protein microarrays.
Noncovalent microarrays

Noncovalent fabrication methods have been used to make microarrays of nucleic acids, small molecules, and proteins (Table 2). Slides coated with aminosilane or poly-L-lysine have been used to capture randomly oligonucleotides (36), proteins (37), and cells (18) via electrostatic interactions or passive adsorption. Similarly, DNA, proteins, and carbohydrates have been arrayed on nitrocellulose-based substrates (38–40). Homogenous orientation may be achieved using affinity tags. For example, large collections of His-tagged proteins have been printed on nickel-coated slides (41). Alternatively, probe biomolecules can be biotinylated and printed on streptavidin-coated surfaces (42). Taking advantage of the highly specific fluoros affinity interaction, Pohl and coworkers (43) noncovalently captured fluoros-tagged carbohydrates on fluoroalkylsilane-coated slides. Winsinger and coworkers (44) prepared SMMs containing a PNA-encoded tetrapeptide acrylate library via sequence-specific hybridization to an oligonucleotide microarray. This elegant approach allows encoding combinatorial libraries and immobilization on-array via self-assembly. Finally, both microarrays and microdroplets have been adapted to the microarray format in an effort to carry out experiments in solution (45, 46).

Detection methods

A rather important planning step involves choice of detection method. Whether using DNA microarrays in whole-genome expression profiling or using small-molecule microarrays to identify new ligands for a protein target of interest, some method of detection is required to locate probe-sample interactions. Typically, probes are labeled with fluorescent, chemiluminescent, radioactive, or affinity tags. Most applications involve a fluorescent readout and the use of a fluorescent microarray slide scanner. A variety of fluorescent reagents, including reactive dyes, fluor-labeled proteins, and fluor-labeled secondary antibodies, are commercially available and considered safer than radioactive labels. Detection strategies may include combinations of different label types. For example, a microarray may be incubated with biotinylated protein followed by incubation with fluorescently labeled streptavidin. Potential disadvantages of label-based detection include overlabeling of protein samples that results in loss of protein conformation or activity. Care should be taken not to label a critical functional group in a small molecule or protein required to bind its target probe. Label-free approaches are gaining momentum in the field of microarrays (47). In particular, surface plasmon resonance (SPR) detection has been used successfully to detect interactions of proteins applied to small-molecule microarrays (48) and will likely prove useful in high-throughput studies that involve protein microarrays.

Applications of Microarrays in Chemical Biology

Microarrays have served as tools for investigations in chemical biology. DNA microarrays are used to study the effects of small molecules on gene expression (49–51). Small-molecule microarrays are used for ligand discovery and enzyme specificity profiling (14–17, 26) as well as for high-throughput, cell-based phenotypic screens (52) and as diagnostic tools (22, 53). Protein microarrays are used for studying interactions with other proteins, nucleic acids, and small molecules, including lipids, carbohydrates, and drugs (23). Protein microarrays are also used as analytical tools to profile complex protein mixtures, such as fractionated cell lysates, in an effort to study antibody specificity, measure changes in protein abundance, or characterize disease states. The microarray approach has
Table 2: Representative capture strategies for preparing microarrays

<table>
<thead>
<tr>
<th>Surface</th>
<th>Coupling partner</th>
<th>Linkage type</th>
<th>Microarray type</th>
</tr>
</thead>
<tbody>
<tr>
<td>amino</td>
<td>carboxylic acid, activated ester, aldehyde</td>
<td>covalent or electrostatic, amide, Schiff base</td>
<td>nucleic acid, small molecule, protein, cell, virus</td>
</tr>
<tr>
<td>carboxylic acid</td>
<td>amino, alcohol</td>
<td>covalent or electrostatic, amide, ester</td>
<td>small molecule, protein</td>
</tr>
<tr>
<td>activated ester</td>
<td>amine</td>
<td>covalent, amide</td>
<td>nucleic acid, small molecule, protein</td>
</tr>
<tr>
<td>aldehyde</td>
<td>amine</td>
<td>covalent, Schiff base</td>
<td>nucleic acid, small molecule</td>
</tr>
<tr>
<td>isocyanate</td>
<td>amine, thiol, hydroxamic acid, alcohol, carboxylic acid</td>
<td>covalent, urea, carbamoyl, carbamate, carboxylamide</td>
<td>nucleic acid, small molecule</td>
</tr>
<tr>
<td>epoxy</td>
<td>amine, alcohol, hydrazide</td>
<td>covalent, alkylamine, ether</td>
<td>nucleic acid, small molecule, protein</td>
</tr>
<tr>
<td>maleimide</td>
<td>thiol</td>
<td>covalent, thioether</td>
<td>nucleic acid, small molecule, protein</td>
</tr>
<tr>
<td>phosphine</td>
<td>azide</td>
<td>covalent, amide</td>
<td>small molecule, protein</td>
</tr>
<tr>
<td>diazirine</td>
<td>nonspecific</td>
<td>covalent, photoaffinity capture</td>
<td>small molecule</td>
</tr>
<tr>
<td>fluoruous</td>
<td>C$_2$F$_7$ tag</td>
<td>noncovalent, fluorophilic interaction</td>
<td>nucleic acid, small molecule, peptide</td>
</tr>
<tr>
<td>streptavidin</td>
<td>biotin tag</td>
<td>noncovalent, high affinity protein-ligand complex</td>
<td>nucleic acid, small molecule, protein</td>
</tr>
</tbody>
</table>
been extended to transfected cell microarrays (18), RNAi living cell microarrays (52), virus microarrays (19), and tissue microarrays (20), all of which may be future tools for characterizing the effects of small molecules. A brief sampling of representative applications of microarrays in chemical biology is presented.

Microarrays of nucleic acids
DNA microarrays enable genome-wide measurement of transcriptional responses to small molecule treatment. Typically, drugs and bioactive compounds from target-oriented assays or phenotypic cellular assays are probed in an effort to gain insight into mechanism of action (49). While a great deal of information is gained from these experiments, it is still difficult to draw conclusions about mechanism within isolated datasets. Rather, systematic and comparative analyses using databases of transcriptional profiles have proven valuable in this regard. Lamb et al. (50) described The Connectivity Map as a publicly accessible reference collection of transcriptional profiles in human cultured cells treated with small molecules, including FDA-approved drugs and bioactive tool compounds. By comparing the profile of a novel small molecule to those of known therapeutics, compounds may be annotated as resembling compounds from a known therapeutic class or novel in genomic signature. Additionally, the Connectivity Map project aims to describe genomic signatures of physiological and disease states. In combination with pattern-matching tools, the authors make connections between drugs, genes, and diseases. Specifically, genomic signatures were used to recognize drugs with common mechanisms of action, discover unknown mechanisms, and identify new compounds with therapeutic potential. The authors proposed expanding this resource by profiling all FDA-approved drugs and inhibitory RNAi targeting a wide selection of genes in a panel of diverse cell lines. Individual researchers may perform their own Connectivity Map analysis by using a web-based tool (www.broad.mit.edu/ctmap). In an effort to create target identification hypotheses for small molecules positives in phenotypic screens, Butcher et al. (51) used a microarray approach to monitor effects of gene overexpression on yeast growth in the presence of small molecules. A collection of roughly 3900 Saccharomyces cerevisiae strains harboring different overexpression plasmids was monitored for changes in growth in response to treatment with a small molecule. As a proof of concept, the authors identified genes that, when overexpressed, affect yeast growth in the presence of the natural product drug rapamycin. Target of rapamycin protein (TOR) was successfully identified as candidate rapamycin target and several new genes were implicated in the TOR pathway. The authors used the same microarray-based method to identify candidate targets for LY-83583, a suppressor of rapamycin-induced growth inhibition. Finally, Amin and coworkers used DNA microarrays to interrogate sequence specificity of several fluorescent-labeled DNA-binding molecules, including engineered polyaniline molecules and proteins (10). The cognate site identifier (CSI) microarrays allow rapid and unbiased examination of sequence space and the authors propose using the arrays to determine the sequence preferences of all metazoan DNA-binding proteins and DNA-binding small molecules.

Small-molecule microarrays
The primary use of SMMs to date is ligand discovery. Advantages, including throughput and rapid access to SAR, and disadvantages, including false negatives caused by orientation, have been reviewed previously (26). Typically, a protein of interest is incubated with the array and binding interactions are visualized using a fluorescent readout. Fluorescence intensity is used to rank positive interactions. As most SMM readouts do not correlate directly with affinity, secondary assays, such as thermal-shifts or SPR, are used to confirm positives and to study both the kinetics and the affinity of binding (26). Representative interactions discovered using SMMs are shown in Fig. 1. Interactions of varying affinities that involve proteins from different functional classes, including transcription factors, cytokines, and enzymes, have been discovered. For example, uretupamine and haptamide bind and perturb the functions of two yeast proteins involved in transcriptional regulation and nutrient-sensing (55, 56). Several ligands with differing molecular scaffolds have been identified for calmodulin using SMMs. Calmodioxane (57) and calmodophilin (58) are products of diversity-oriented syntheses executed by Schreiber and coworkers. NPC-15437, which is a known inhibitor of protein kinase C, binds to calmodulin preferentially when Ca²⁺ is present in incubation buffer (26). Hsieh-Wilson and coworkers (59) used chondrin sulfates microarrays to identify chondrin sulfate-E tetrasaccharide as a ligand for tumor necrosis factor alpha (TNF-α) and demonstrated that the compound disrupted a cytokine-cell surface receptor interaction. Selective inhibitors of closely related cytokine proteases, catherpin F and catherpin K, were identified using the PNA-encoded tetrapeptide acrylamide microarrays prepared by Urbina et al. (44). Small-molecule ligands have also been identified for human IGF (25), FKBP12 (60), and A urora A kinase (61). SMMs have also been used to identify high-affinity interactions between RNA secondary structure motifs and small molecules (62). Microarrays containing drug-like small molecules and carbohydrates have proven very useful for studying carbohydrate-cell interactions because the surface presentation mimics interactions at cell surfaces (22, 63). Enzyme activity assays may also be performed in the SMM format. Yao and coworkers (64) used nanodroplet SMMs for profiling 400 hydroxamate-containing peptide analogs against two matrix metalloproteases and identified selective inhibitors with IC₅₀ values in the nanomolar range. Finally, Dordick and coworkers (65) disclosed a very exciting new application that involves on-array synthesis of natural product analogs using in vitro metabolic pathway construction. The authors demonstrate that these microarrays serve as a platform for synthesis and screening by identifying three inhibitors of Fyn tyrosine kinase.

Protein microarrays
The first functional protein and peptide arrays were micropatterned using lithographic methods in 1992 (66). Several years later, McDade and Schreiber (13) reported spotted functional protein microarrays capable of detecting protein-protein interactions, protein-ligand interactions, and biochemical activities.
Figure 1  Representative protein–ligand interactions discovered using small-molecule microarrays. (Adapted from Refs. 26, 27, 44, and 55–62.)

Since initial reports, protein microarrays, including peptides, domains, full-length proteins, antibodies, and lysates, have been used in a variety of applications that are reviewed elsewhere (23). MacBeath and coworkers (67) have used protein microarrays to build quantitative protein interaction networks between ErbB receptors with Src homology 2 (SH2) domains and phosphotyrosine binding (PTB) domains. The authors discovered several new binding interactions and proposed that the oncogenic potential of receptor tyrosine kinases may be a function of alterations in binding promiscuity because of changes in EGFR and ErbB2 protein concentrations as both proteins are notably overexpressed in several cancers. The same research group used a similar approach to examine binding of 157 mouse PDZ domains to 217 genome-encoded peptides in an effort to obtain a broad view of selectivity (68). Snyder and coworkers (69) prepared microarrays that contain 282 yeast transcription factors and probed them with fluor-labeled oligonucleotides in an effort to link sequence to binding. Using this approach, the authors defined the binding site of an uncharacterized DNA-binding protein and determined that several of its target genes are involved in stress response and oxidative phosphorylation. The same research group performed biochemical activity analyses on yeast proteome chips to identify in vitro substrates for protein phosphorylation (70). Protein microarrays are also useful tools for target identification (71). As outlined in Fig. 2, Huang et al. (71) prepared biotinylated versions of small molecules known as SMIR3 and SMIR4 that scored as positives in a chemical genetic modifier screen. The labeled compounds were applied to a yeast proteome chip followed by incubation with fluorescently labeled streptavidin. In this fashion, the authors identified lists of candidate protein targets for each compound. One of these proteins, Ybr077cp, was the subject of follow-up studies to confirm the interaction in vivo. Several research groups have used protein and antibody arrays for profiling activities in complex fractions such as whole lysates or proteomic fractions (Fig. 3). Cravatt and coworkers (72) have employed activity-based protein profiling (ABPP) using active-site-directed probes to profile the functional state of enzymes in proteomes. The authors first treated proteome fractions with fluorescent activity-based probes and then applied the fractions to microarrays that contain antibodies against enzymes of interest. Labeled enzymes will be captured so long as a complementary antibody is present. Miah and coworkers (73) developed a method for analyzing the dynamic glycosylation status of cells by profiling differentially fluor-labeled membrane fractions on microarrays that contain lectin proteins. In analogy to gene-expression profiling, the authors propose this method as a means to make comparisons between glycosylation patterns in different cell states. Finally, reverse-phase protein microarrays, where cell lysate samples are...
target identification using yeast proteome microarrays. (a) Using a chemical genetic modifier screen, Huang and coworkers (71) identified small-molecule inhibitors of rapamycin (SMIRs) that suppress the antiproliferative effect of rapamycin in S. cerevisiae. (b) Biotinylated derivatives of SMIR3 and SMIR4 were synthesized for incubation with proteome microarrays in an effort to identify putative protein targets. (c) Binding of biotinylated SMIR3 and SMIR4 to proteins on a yeast proteome chip that contains ∼5,800 proteins was detected using Cy3-labeled streptavidin. Nine proteins bound to SMIR3 with one protein, Tep1p, annotated as a strong binder. Thirty proteins bound to SMIR4 and four proteins were annotated as strong binders, including Rot1p, YBR077wp, YBR193wp, and Mep1p.

Although less developed relative to microarrays of biomolecules, living microarrays will no doubt play an increasing role in chemical biology. Current applications of this technology include high-throughput and localized transfection (18), high-throughput phenotypic screening of small molecules (52), and loss-of-function screens using RNAi (19, 54). The data gained from these types of experiments should be highly complementary to that gained from the effects of evaluating small molecules using the other microarray platforms. To make the most of the various microarray platforms, standardized array data outputs will be required to address issues that surround reliability and reproducibility of data generated in a microarray format (76). A additionally, databases equipped with Web-based analysis tools will be needed to make the large amount of information generated using microarrays available for public scrutiny (77, 78).

References
Microarrays in Chemical Biology


Figure 3 Changes in protein abundance levels in response to alterations of cell state (e.g., disease or treatment with a small molecule) can be monitored using microarrays. (a) By analogy to transcriptional profiling, antibody microarrays may be used to profile fluorescently-labeled protein fractions. In this scheme, proteins enriched in the compound-treated sample, relative to untreated cells, provide red microarray features. Proteins depleted in the compound-treated sample appear as green features. Proteins that do not change in abundance result in yellow features. Proteins that are absent from the samples, resistant to labeling, or overlabeled such that they no longer recognize the printed antibody give no fluorescent signal. (b) Alternatively, protein fractions from cell lysate samples may be printed on microarrays. Replicate arrays are then probed with different fluor-labeled antibodies of interest to monitor changes in protein abundance. In this scheme, lysate features that contain proteins of interest will be fluorescent (red), whereas lysates that do not contain the protein are not fluorescent (black). Fluorescently labeled control proteins, such as BSA or streptavidin, or antibodies are included as reference points in the arrays (green).
Microarrays in Chemical Biology


57. Wong JC, Stenson SM, Louca JB, Hong R, Schreiber SL. Modular synthesis and preliminary biological evaluation of sterocemic- 

58. Barnes-Seeman D, Park SB, Koehler AN, Schreiber SL. Expanding the functional group compatibility of small-molecule microar-


61. Miao H, Tallarico JA, Hayakawa H, Muger K, Duffner JL, Koehler AN, Schreiber SL, Lewis TA. Ring-opening and ring-

62. Kwon SK, Lee MY, Ku B, Sherman DH, Doridy JS. High-
  throughput, microarray-based synthesis of natural product ana-
  logues via in vitro metabolic pathway construction. ACS Chem. 

63. Disney MD, Childs-Disney JL. Using selection to identify and 
  chemical microarray to study the RNA internal loops recognized 

64. Disney MD, Stenger PH. The use of carbohydrate microarrays to study carbohydrate-cell interactions and to detect pathogens. 

65. Wang J, Utsunomichandani M, Sun LP, Yao SQ. Activity-based high-throughput profiling of metalloproteinase inhibitors using small 

66. Brillard S, Perez-Arnaua E, Clark PG, McMinn B, Connolly P, 

  proteome chip for profiling enzyme activities in complex proteomes. J. Am. 

68. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, Allen JE, 
  Zaslavakaia LA, MacBeath G. PDZ domain binding selectivity is 

  interaction network for the ErbB receptors using protein 

70. Liu B, Baban GC. Synthesis of cationic conjugated polymers for use in 

Further Reading

Butcher RA, Schreiber SL. A microarray-based protocol for mon-
  itoring the growth of yeast overexpression strains. Nat. Protoc. 
  2006;1:2344-2352.

Bradner J, McPherson OM, Koehler AN. A method for the covalent 
  capture and screening of diverse small molecules in a microarray 

Chattopadhaya S, Tan LP, Yao SQ. Strategies for site-specific protein 
  blotting using in vivo, in vitro, and cell-free systems. Toward 

Fiegler H, Fardon R, Carpenter NF. Construction and use of spotted 
  large-insert clone DNA microarrays for the detection of genomic copy 

Hill HD, Mulich CA. The biochip code: Detection of the protein 
  and nucleic acid targets using DTT-induced ligand exchange. Nat. 
  Protoc. 2006;1:2324-2336.

Hsi L, Mital K. A lectin microarray approach for the rapid analysis of 

Jing R, Bolshakov V, Flavell AJ. The tagged microarray marker (TAM) 
  method for high-throughput detection of single nucleotide and indel 

  2006;1:2344-2352.

Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation 

Further Reading

Butcher RA, Schreiber SL. A microarray-based protocol for mon-
  itoring the growth of yeast overexpression strains. Nat. Protoc. 
  2006;1:2344-2352.

Bradner J, McPherson OM, Koehler AN. A method for the covalent 
  capture and screening of diverse small molecules in a microarray 

Chattopadhaya S, Tan LP, Yao SQ. Strategies for site-specific protein 
  blotting using in vivo, in vitro, and cell-free systems. Toward 

Fiegler H, Fardon R, Carpenter NF. Construction and use of spotted 
  large-insert clone DNA microarrays for the detection of genomic copy 

Hill HD, Mulich CA. The biochip code: Detection of the protein 
  and nucleic acid targets using DTT-induced ligand exchange. Nat. 
  Protoc. 2006;1:2324-2336.

Hsi L, Mital K. A lectin microarray approach for the rapid analysis of 

Jing R, Bolshakov V, Flavell AJ. The tagged microarray marker (TAM) 
  method for high-throughput detection of single nucleotide and indel 

  2006;1:2344-2352.

Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation 
See Also
Forward Chemical Genetics
Array-Based Techniques for Glycans: Development and Applications
Peptide Combinatorial Libraries
Transcript Profiling
Receptor-Ligand Interactions
Combinatorial Libraries: Overview of Applications in Chemical Biology
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Gerd Nielsen*, Max Stadler*, Henry Jonker, Marco Betz and Harald Schwalbe, Institute for Organic Chemistry and Chemical Biology, Center for Biomolecular Magnetic Resonance, Frankfurt, Germany
doi: 10.1002/9780470048672.wecb393

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful method to determine the structure of biomacromolecules and their complexes in solution. It allows determination of the dynamics of proteins, RNA, DNA, and their complexes at atomic resolution. Therefore, NMR spectroscopy can monitor the often transient weak interactions in the interactome of proteins and the interaction between proteins and small-molecule ligands. In addition, intrinsically unstructured proteins can be investigated, and first reports of structure determination of membrane proteins in the immobilized state (solid state) are developing. This review will introduce the fundamental NMR observables as well as the methods to investigate structure and dynamics, and it will discuss several examples where NMR spectroscopy has provided valuable information in the context of Chemical Biology.

Chemical Biology

Research in Chemical Biology is dedicated to the design and targeted synthesis of novel molecules of small or large molecular weight to investigate how they maintain, modulate, regulate, or interfere with cellular function or even change the function, morphology, and differentiation status of entire cells. The structure-function relationship that leads to the understanding and prediction of cellular function is at the heart of chemical biology. Molecules exert function through their chemical properties, which are determined by the dynamic spatial arrangement of their atoms, their interactions with other molecules, their stabilities against degradation, and their localization within the cell.

Chemical Biology encompasses more aspects than the development of novel high-affinity ligands for targets of pharmaceutical interest alone. However, if the latter is understood as the art of developing novel chemical entities toward new drugs, then Chemical Biology has a tremendous impact on the understanding of key properties of molecules that may induce specific cellular responses or help maintain cellular function.

NMR Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a noninvasive and nondestructive spectroscopic technique that allows determination of the constitution and relative configuration of molecules, the characterization of the dynamic three-dimensional (3D) conformation of molecules, and their interaction with other molecules. NMR spectroscopy detects the characteristics of nuclear spins; the most commonly studied nuclei are the spin-1/2-particles 1H, 13C, 15N, and 31P. NMR observables sensitively depend on their chemical surroundings of individual atoms. Therefore, NMR spectroscopy can derive information about the conformational dynamics and interactions of molecules in solution and at ambient temperature. In addition, thermodynamic and kinetic information about the interaction of molecules can be derived on a per-atom basis.
Biochemical NMR spectroscopy is applicable to both liquid- and solid-state samples. Liquid-state NMR spectroscopy, in which molecules are dissolved in a variety of different solvents and studied at ambient temperatures, is a powerful tool to derive information on the structure of proteins and nucleic acids, as well as their complexes with each other and small molecules, ions, and solvents. Liquid-state NMR can be applied not only to native folded states of proteins, but also to intrinsically unstructured proteins as well as proteins in their unfolded state and under nonphysiological conditions (i.e., in organic solvents).

Figure 1 provides an overview on the number of protein structures determined by liquid-state NMR spectroscopy. NMR spectroscopy can detect the conformational dynamics, which are also referred to as conformational switching of RNA molecules that exist in different stable states. Solid-state NMR spectroscopy investigates molecules as powders or crystals and has become a powerful tool for the investigation of membrane proteins and their complexes with small peptide agonists. In addition, protein amyloid fibers and polymers can be investigated.

Furthermore, NMR spectroscopy can study intrinsically unstructured proteins and noncoding RNAs uniquely, which includes their structural transitions induced when encountering molecular targets. The technique can be applied to systems under defined conditions (in vitro), but proteins can also be investigated within cells (in vivo). NMR can monitor the concentration fluxes of complex mixtures, which are extracted from tissue or within body fluids. These fluxes are important for understanding the metabolic state of an organism (metabonomics). In this context, NMR has been applied to characterize the metabonomics of entire organisms such as Caenorhabditis elegans.

In this review, we present the basic observables and tools in liquid-state NMR spectroscopy followed by examples of the application of NMR in chemical biology. We will focus on the application of NMR spectroscopy to study proteins in solution.

Basic NMR Observables

In NMR spectroscopy, signals of NMR-active nuclei with spin $i$ (e.g., $^1$H, $^{13}$C, $^{15}$N, $^{19}$F, and $^{31}$P) have Lorentzian line shapes with center peak positions referred to as chemical shift $\delta$. Their linewidths at half height $\Gamma_{1/2}$ depend on the overall correlation time for rotational tumbling $\tau_c$ of the molecule that is proportional to the size of the molecule, and on possible local conformational or chemical exchange processes. Figure 2 shows the one-dimensional (1D) proton spectrum of a small molecule, which is an inhibitor of the kinase p38, dissolved in dimethyl sulfoxide (DMSO). The integral of each signal corresponds to the number of different protons that cannot be superimposed by symmetry operations or that are equivalent because of dynamic averaging. Magnetically nonequivalent atoms generate higher-order spectra; this phenomenon will not be discussed here.

Expansions of three regions in the spectrum on the right hand side of (Fig. 2) show how protons display different line shapes according to their chemical environments. The protons in exchange with residual water in the DMSO solvent, such as the urea protons (H1 and H2), exhibit exchange broadened signals (Fig. 2a), whereas nonexchangeable protons (e.g., H3) (Fig. 2b) have sharp signals. The aromatic protons are coupled to each other through scalar $J$-couplings (vide infra), which results in the higher-order spectrum with a complex peak pattern shown in (Fig. 2c).

In the following five paragraphs, the basic NMR observables will be introduced briefly, for a more thorough introduction we recommend “Spin dynamics, basics of nuclear magnetic resonance” by Malcolm H. Levitt (1) or “NMR: the toolkit” by Peter Hore et al (2).

Chemical shift

The most important NMR parameter is the chemical shift $\delta$, which is derived from the Larmor frequency of a given nucleus $\gamma$ that resonates in a magnetic field. The chemical shift is measured in parts per million (ppm) relative to a reference compound.

The Larmor frequency of a given nucleus depends on the gyromagnetic ratio $\gamma$ (a physical property of the nucleus) and the magnetic field. The electron density that surrounds the nucleus results in an additional magnetic field that opposes the external field and thereby slightly alters its Larmor frequency.

\[ \delta = \frac{\gamma B_0}{2 \pi} \]

The chemical shift can be influenced by chemical exchange, molecular dynamics, and other through scalar $J$-couplings (vide infra), which results in the higher-order spectrum with a complex peak pattern shown in (Fig. 2c).

Chemical shift

The most important NMR parameter is the chemical shift $\delta$, which is derived from the Larmor frequency of a given nucleus $\gamma$ that resonates in a magnetic field. The chemical shift is measured in parts per million (ppm) relative to a reference compound.
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

This modulation of the main field is referred to as shielding, and the shift of the NMR frequency leads to characteristic chemical shift values for the different functional groups or certain environments.

The NMR signal of a nucleus will appear at an average chemical shift value. In case of the three protons of a methyl group, such averaging comes about from fast rotation around the CC-bond that connects the methyl group to the rest of the molecule; the rotation leads to identical chemical shifts for the three protons. Slow molecular motions on the NMR timescale (see the section entitled “Dynamic information from line shape analysis”) can lead to one nucleus having two peaks due to two different conformations. On the intermediate timescale, the averaged NMR signal will be broadened and may disappear. Thus, the NMR chemical shift values are very sensitive probes of the local structure and conformation.

Different methods exist that predict protein secondary structure elements from chemical shift values [e.g., Chemical Shift Index (CSI) (3) and Probability-based protein Secondary Structure Identification using combined NMR chemical-shift data (PSSI)] (4). These methods are based on the statistics of database analysis of chemical shift values from a range of peptides and proteins of known secondary structure. Chemical shifts have also been used to predict backbone torsion angles by software such as Torsion Angle Likelihood Obtained from Shifts and Sequence similarity (TALOS) (5).

More recently, software has been developed that predicts 1H, 13C, and 15N chemical shift values of proteins from either 3D structure files, for example, SHIFTS (6), SHIFTX (7), and SPA RTA (8), or from the mere amino acid sequence using SHIFTY (9). First results have been reported on the de novo structure determination of proteins using fragment-based chemical shift predictions and molecular modeling (10, 11).

Spin-spin coupling (J-coupling constant)

The 1J-coupling, which is also known as the indirect spin-spin-coupling or scalar coupling, develops from interactions between electrons and nuclear spins of atoms n bonds apart, where there is a slight energetic preference for the nuclear spin to have the same direction as the nearest electron spin. Briefly, this phenomenon is known as the Pauli Exclusion Principle (stating that two fermions cannot occupy the same quantum chemical state simultaneously) that leads to tiny but measurable energy differences depending on the spin states of covalently bonded atoms. The J-coupling will split the NMR signal of each nucleus into multiplets depending on its number of NMR active neighbors (i.e., the amount of different spin states available within the coupled spin system). Typically, scalar couplings are observed for nuclei that are connected by up to three bonds. In rigid systems such as aromatic rings, however, they may even be observable across up to five bonds. The size of the coupling between two nuclei is termed the J-coupling constant. The multiplicity of a signal is determined by the number of chemically equivalent nuclei that couple to the nucleus of interest by through-bond interactions, usually one, two, or three bonds apart.

Vicinal 1J-coupling constants provide valuable information for the determination of biomacromolecule conformation. The structural information is derived from Karplus equations (12) that provide empirical relationships between dihedral bond angles, and the 1J-coupling. Karplus equations obey the following general formula:

\[ 1J(\phi) = A \cdot \cos^2 \phi + B \cos \phi + C \]

where \( \phi \) defines the torsion angle; and A, B, and C are either empirically calibrated or calculated from quantum chemistry. A given 1J-coupling may correspond to up to four different torsion angles, but this ambiguity can be resolved by measuring...
which are indicative of secondary structure, can be defined and measured as the Nuclear Overhauser Effect (NOE) (16). The NOE effect arises from dipolar relaxation effects. This interaction between dipoles of the internuclear vector in the magnetic field persists and increases with the gyromagnetic ratio ($\gamma$) of the two nuclei each possessing a magnetic moment that creates a local magnetic field that will be sensed by the neighboring nucleus. The interaction strength of two magnetic dipoles depends on the internuclear distance ($d_{AB}$) and the orientation ($\theta_{AB}$) of the internuclear vector that connects A and B and the external magnetic field.

In isotropic solution, the dipolar coupling is averaged to zero through molecular rotational tumbling with a characteristic correlation time $\tau_c$. Yet, on shorter timescales, fixed orientations of the internuclear vector in the magnetic field persist and induce dipolar relaxation effects. This interaction between dipoles causes relaxation by transfer of magnetization between different NMR active nuclei and is called the NOE. The NOE effect is strongest between protons as they have large gyromagnetic ratios, and NOEs are generally observed between protons that are situated within 5 Å of each other.

Partial alignment introduces a residual dipolar coupling on a spin pair like the backbone $^1$H-$^1$C bond. A $^1$H-$^1$C spin pair in isotropic solution will, in a coupled spectrum, yield a doublet with a splitting the size of the $^1$J-coupling. In an anisotropic solution, however, the splitting observed is the sum of the scalar coupling ($^1$J$_{CB}$) and the residual dipolar coupling ($^1$D$_{CH}$). The size of the residual dipolar coupling depends on the orientation of the bond vector relative to the molecular alignment tensor.

One important application of RDC measurements is the structural refinement of biomolecules that consist of several domains that are connected by more or less flexible linkers. Because of the flexibility of the linker and the distance between domains, $^1$J-couplings and NOE restraints will frequently not be sufficient for correct determination of the relative domain orientation. The addition of RDC restraints in structure calculation not only refines the biomolecular structure but also allows the relationship between structure and function to be studied. Interactions with other biomolecules and ligand binding may induce an intramolecular rearrangement of the relative orientation of domains that is detectable through RDC measurements (23, 22).

**Relaxation**

Relaxation, or the return of nuclear spin magnetization to the equilibrium state, occurs essentially by two different physical processes that allow the nuclear spins to exchange energy with their surroundings: one that occurs parallel to the direction of the external magnetic field (longitudinal) and one that is transverse to the external field. The longitudinal relaxation time $T_1$, or spin-lattice relaxation, develops from interactions with neighboring unexcited nuclei (the lattice) that affects the component of magnetization in the direction of the external magnetic field. The transverse relaxation time $T_2$, or spin-spin relaxation, develops from dispersion of magnetization between excited neighboring spins that have magnetic moments orthogonal to the external field. The longitudinal relaxation time in the rotating frame, $T_{1w}$, is the relaxation time of magnetization that has been spin-locked perpendicular to the external field (23, 24).

$T_1$ reports on fast dynamics on a timescale of ps-ns, whereas $T_2$ relaxation depends on both fast and slower dynamics (ps-ns and μs-ms). The experimentally measured $T_2$ relaxation times include an exchange contribution that can be measured by a Carr-Purcell-Meiboom-Gill (CPMG) pulse train (25, 26) or an effective spin-lock field (27-29). The combination of $T_1$ and $T_2$ measurements allows determination of the contribution of chemical exchange to the relaxation time. Furthermore, relaxation dispersion experiments have been developed to measure slow time-scale μs-ms dynamic processes (30-35).

Heteronuclear NOEs (e.g., $^1$H-$^15$N or $^1$H-$^1$C hetNOEs) are obtained by measuring HSQC-type spectra (see the section entitled “Two-dimensional heteronuclear correlation experiments”) with and without proton saturation. The hetNOE is extracted from the difference in the signal amplitude of these measurements and reports on the fast dynamics of the heteronuclear bonds (ps to ms timescale). Maximal hetNOE values are observed when the bond vector tumbles at the same frequency as the entire protein, whereas faster motion with respect to overall tumbling leads to smaller hetNOEs.
The conformational dynamics of a biomolecule can be determined by measuring the relaxation properties of the heteronuclear bond vectors (e.g., the protein backbone $^{1}H$-$^{15}N$ bond). Regions of a protein that are unstructured or flexible (e.g., loops and tails) will typically show larger $T_2$ and $T_1$ values and smaller $T_2^*$ values and hetNOEs than the rigid well structured core of the molecule. Distinct changes in the flexibility of a biomolecular structure caused by ligand binding can therefore be probed using a combination of heteronuclear relaxation experiments.

Relaxation is strongly dependent on molecular motions. The overall random molecular tumbling, which is expressed in the rotational correlation time $\tau_\nu$, governs the overall relaxation process. Larger molecules have slower tumbling motions that lead to higher $\tau_\nu$ values. However, local dynamics and independent domains can modulate the relaxation parameters, which account for differences in their flexibility and mobility.

Typically, relaxation times are interpreted in the framework of a model-free analysis (36-38), and the general order parameter $S_2$ of a given heteronuclear bond can be extracted. $S_2$ defines the spatial restriction on a per-residue basis of a target protein. Unique information can be obtained about the changes that occur to a target molecule on ligand binding by measuring $S_2$ before and after the binding event. The order parameter varies between 0 and 1, which spans from completely unrestricted internal motion of the bond vector ($S_2 = 0$) to complete rigidity ($S_2 = 1$). Dynamics that are equal to or faster than the overall correlation time can be measured by the order parameter (39, 40).

Experimental Techniques

NMR spectroscopy provides important information on molecular structure and dynamics, but certain limitations apply for the study of biomolecules. The sample conditions need to be optimized toward stability at high concentrations without exceeding a salt concentration of roughly 300 mM (leading to increased NMR pulse length). To reduce $N$-H exchange, the solution should be buffered to a pH between 5 and 7. Distinct chemical shifts allow monitoring of individual atom types, but dispersion is rather low (on the sub-ppm scale) and easily produces spectral overlaps especially for larger molecules. The size of the molecular target (e.g., protein or RNA) also restricts NMR experiments because of relaxation effects. It is, however, possible to reduce these relaxation effects greatly by partial or selective deuteration.

To reduce spectral overlap, 1D NMR can be extended to higher dimensional NMR experiments. Two-dimensional (2D) experiments that correlate protons through J-couplings or dipolar couplings include $^1$H-$^1$H COSY (43), $^1$H-$^1$H TOCSY (42), and $^1$H-$^1$H NOESY (43, 44). Particularly helpful are 2D het-teronuclear experiments, which correlate protons with directly bound carbon or nitrogen atoms. Furthermore, a variety of 3D and higher-dimensional experiments has been developed that are essential for assigning large biomolecules (45). In contrast to protons, NMR-active and stable isotopes $^{13}C$ and $^{15}N$ both have low natural abundances, and proteins are therefore typically expressed in Escherichia coli bacteria that grow in a minimal medium that contains only $^{12}$C, $^{13}$N labeled precursors, which allows proteins to be studied in isotope labeled form. Recently, expression media have also been established to express isotope-labeled proteins in sf9 cells (46) or HEK293 cells (47).

Typical protein concentrations are between 0.1-1 mM in 0.3-0.5 mL of buffered solutions that correspond to approximately $10^7$-10^8 molecules per NMR sample for detailed structural studies, but it can be as low as 10^5 for more analytical purposes.

Sensitivity can furthermore be increased up to fourfold by using cryogenic probe technology, in which the radio frequency (RF) transmitter and preamplifier coils are super-cooled to 20K by helium gas that ensures higher signal-to-noise of the electrical signal. Vacuum insulation around the RF coil allows the NMR sample situated only millimeters away from the coil to be measured at ambient temperature.

The development of higher magnetic field strength has also been vital to improve both resolution of NMR spectra (proportional to $B_0$), where $n$ is the dimensionality of the NMR experiment and sensitivity (proportional to $B_0^n$).

A rather new way of overcoming spectral overlap for proteins is selective isotope labeling of one or several amino acid types, which results in less-crowded $^1$H-$^15$N spectra with cross-peaks from the labeled residues only. The same is possible for RNA where nucleotides can be selectively isotope enriched. In addition, methods for segmental labeling of parts of the biomolecules have been developed for proteins (48, 49) and for RNA (50).

One-dimensional NMR spectroscopy

The very high resolution of the chemical shifts gives information on a per-residue basis, and it also leads to spectral crowding in simple experiments. Figure 3 shows three very different 1D proton spectra of the 14-kDa protein called $\alpha$-lactalbumin, in three different states: native folded, molten globule, and unfolded (urea denatured).

The folded state (Fig. 3a) shows sharp lines and a large degree of chemical shift dispersion. The side-chain methyl groups positioned in the interior of the protein can be in close contact with aromatic rings that induce ring current effects. Methyl groups are often shifted toward very low ppm values (between 0.5 and -0.5 ppm). The molten globule state (Fig. 3b) is a partially folded, highly dynamic protein state that contains some stable secondary structure elements, but with a dynamic tertiary structure. The time scale of the interconversion between different states is typically slow, of the order of milliseconds, which leads to significant line broadening. No side-chain methyl groups are as well shielded as in the folded state, and therefore they show chemical shift values larger than 0.5 ppm. The unfolded state (Fig. 3c), but the total lack of secondary structure elements leads to sharper lines than in the molten globule state.

Protons have large gyromagnetic ratios and are therefore highly sensitive NMR nuclei. The high sensitivity in combination with their high abundance in biomolecules has made protons the nuclei of choice in biomolecular NMR. However,
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

3.5 3.0 2.5 2.0 1.5
1H (ppm)
1.0 0.5 0.0
−0.5

Figure 3 1D proton spectra of the side-chain region of α-lactalbumin in (a) folded state, (b) partially folded (molten globule), (c) unfolded.

other nuclei can also be detected directly for specific purposes. 31P is a moderately sensitive NMR nucleus present in the nucleic acid backbone and in the side chains of phosphorylated proteins. Its chemical shift is highly sensitive, and 31P is therefore a useful probe of structural changes.

19F is almost as sensitive as 1H and therefore is well suited for direct detection. However, for biomolecular NMR, it is of limited interest, as it does not occur naturally in biomolecules; instead, it has to be introduced by chemical modification or labeled precursors for biosynthesis. In small drugs, 19F can be used as a metabolic tracer for interaction studies or as a contrast agent in imaging applications. NMR active isotopes 13C and 15N are not only scarce but also have the disadvantage of low NMR sensitivity. Traditionally, they have therefore only been observed indirectly through neighboring protons. Recently, experiments have been developed to detect either 13C or 15N directly both for proteins (51) and RNA (52, 53).

Two-dimensional homonuclear correlation experiments

A 2D NMR experiment consists of a series of 1D experiments, in which the magnetization is transferred from one nucleus to another. This experiment is followed by a delay called the evolution time, where the spins are allowed to precess freely. Incrementing the evolution time in successive experiments results in a 2D spectrum that correlates protons that have exchanged magnetization during the evolution time.

In the 1H-1H COSY (correlation spectroscopy) experiment (41), magnetization is transferred via the J-coupling and shows correlations between protons three bonds apart. The cross peak usually shows a characteristic antiphase-square pattern, but it may be split even more by additional passive couplings that lead to spectral crowding and loss of intensity. For small molecules, the COSY spectrum may suffice for assignment of the proton resonances. For larger molecules, such as peptides, the COSY-experiment is used in combination with other homonuclear experiments, especially because the spectral region of the aliphatic protons is often too crowded to allow unambiguous assignment.

In the 1H-1H-TOCSY (Total Correlation Spectroscopy) experiment (42), magnetization is also transferred via the J-coupling, but an additional isotropic mixing step leads to the correlation of all spins within a given spin system. In the case of a small peptide, the amide backbone protons show cross peaks to all side-chain protons of the same amino acid residue, which facilitates the identification of amino acid type by their characteristic peak patterns and circumvents the problem of overcrowding in the COSY spectrum. The individual amino acid residues (spin systems) are finally connected by distance information obtained from the 1H-1H-NOESY (Nuclear Overhauser Effect Spectroscopy) experiment (44), in which the correlation is transferred via the dipolar coupling. Cross-peaks may be observed for protons that are up to 5 Å apart from each other. For a small peptide (Fig. 4), this information is sufficient for a sequential assignment. A closely related experiment is ROESY (Rotational frame nuclear Overhauser Effect Spectroscopy) (44, 54), which is complementary to the NOESY in respect to its dependency on the correlation time of the molecule studied. With increasing size, the NOESY cross-peaks display a sign change from positive to negative, and the signal intensity is very low close to the transition point. This low NOE signal intensity is often the case for medium-sized peptides in which case the ROESY experiment is preferable. The signal intensity in the ROESY spectrum is also dependent on the size of the molecule, but the sign is always negative.
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Two-dimensional heteronuclear correlation experiments

Introducing a heteronuclear dimension reduces the signal overlap by using the additional chemical shift dispersion of the heteronuclei and facilitates assignment of biomolecules. The Heteronuclear Single Quantum Correlation (HSQC) experiment yields a spectrum that correlates the chemical shift of a $^1$H spin with that of a covalently bound $^{13}$C or $^{15}$N spin (55). In a $^1$H-$^{15}$N HSQC spectrum, every peak represents the correlation of an amide $^1$H-$^{15}$N bond, which shows correlations for both backbone and side-chain amides of proteins and nucleotide imino protons of RNA. Equivalently to the 1D case (Fig. 3), different folding states of a protein will lead to distinctly different $^1$H-$^{15}$N HSQC spectra; these states are shown for $\alpha$-lactalbumin in folded, molten globule and unfolded states in (Fig. 5a-c). Chemical shift dispersion is clearly much larger in the folded state than in the molten globule or unfolded states. This increased chemical shift dispersion facilitates analysis of larger proteins in their native, folded state, as much less signal overlap is observed compared with the unfolded states. The highly flexible molten globule shows significant line broadening of signals; therefore, many peaks are not visible. In the unfolded state, most peaks reappear, but the signal overlap in the proton dimension is significantly worse than in the folded case as all amide protons are equally solvent accessible and therefore experience similar degrees of shielding from the external magnetic field. Contrarily, the nonexchanging $^{15}$N nuclei typically retain high chemical shift dispersion in the unfolded and molten globule states.

Figure 4  Heteronuclear 2D spectra of the 8mer peptide EVVTLYWR in 90 % H$_2$O, 10 % D$_2$O. (a) Representation of coherence transfer pathways for COSY (solid arrows), TOCSY (dotted arrows) and NOESY (dashed arrows) experiments. (b) Section of COSY spectrum displaying the backbone H$^N$-H$^\alpha$ correlations, additionally the side-chain H$^N$-H$^\alpha$ correlations of R8 is visible in the upper right-hand corner. The H$^N$-H$^\alpha$ correlation of E1 cannot be detected because of solvent-exchange broadening of the N-terminal amino group. (c) Section of TOCSY spectrum that displays the correlations of the backbone H$^N$ with all protons within the amino acid side-chain. (d) Section of ROESY spectrum that displays correlations between backbone H$^N$-H$^\alpha$ intraresidual as well as to the neighboring (i-1) amino acid. The cross peaks to the (i-1) amino acid have higher intensities. Thus, a "sequential walk" is possible (arrows) that allows identification of the position of amino acids within the peptide chain. Additionally the H$^\alpha$ of E1 can be assigned (first arrow on the left).
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Similar to the HSQC experiment, multiple quantum coherences can be used to correlate protons with $^{13}$C-coupled heteronuclei. The information content of the Heteronuclear Multiple Quantum Correlation (HMOC) experiment (56) is equivalent to the HSQC, but the sensitivity can be improved in certain cases. Additionally, by proper tuning of delays and phase cycling, it can be transformed into the heteronuclear multiple bond correlation experiment (57–59), which results in correlations between $^1$H- and $^2$H-coupled nuclei.

For large biomolecules, fast relaxation of the transverse magnetization becomes a serious problem that leads to poor spectral resolution with broad or lacking signals. In a coupled $^1$H-$^{15}$N HSQC spectrum, cross-peaks will appear as multiplets, in which the different components of the multiplet have significantly different line widths because of constructive or destructive interactions of the two main relaxation mechanisms: dipole–dipole relaxation and relaxation caused by the chemical anisotropy. The TROSY (60) experiment (Transverse Relaxation Optimized Spectroscopy) overcomes much of the problem of fast transverse relaxation by selecting the one component of the multiplet, in which these relaxation effects almost cancel each other out, which thereby renders spectra with single sharp peaks even for very large biomolecules. The TROSY experiment relies on the fact that the chemical shift anisotropy scales with the magnetic field whereas dipole–dipole relaxation is field independent, and the maximal TROSY effect is only obtained at high magnetic field strength (above 14.1 Tesla), where it has significantly extended the range of biomolecules amenable to NMR spectroscopy.

Applications of NMR Spectroscopy to Chemical Biology

In this article, some examples are shown of the application of NMR spectroscopy in chemical biology. We put emphasis on experiments that characterize interactions of proteins and small molecular weight ligands. These interactions can be mapped either by characterizing the target protein (protein-observed experiments) or the ligand (ligand-observed experiments). Finally, the method of structure calculation based on NMR-derived data is briefly introduced.

This introductory review cannot possibly cover the whole range of techniques in a field as multifaceted as NMR spectroscopy. We therefore refer to review articles published during the last decade (63–67).

Protein-observed experiments

Binding of a ligand will cause diverse effects on a protein, many of which can be probed by NMR spectroscopy of isotopically enriched protein. Depending on the used isotopes and the labeling pattern, such studies can be both costly and time consuming. Nevertheless, protein-observed experiments offer more detailed information about the binding event than what can be obtained from ligand-observed methods. Especially, they map out the dynamic and often remote (e.g., allosteric) response of the protein receptor on binding of a small molecule. Changes in the chemical environment of each nucleus are indicated by chemical shift perturbations, either directly induced by ligand binding or via spectral crowding and allow for new assignment strategies for proteins (45) and nucleic acids (62). Experiments such as TOCSY-HSQC and NOESY-HSQC combine homonuclear and heteronuclear techniques, whereas triple-resonance experiments such as HNCO, HNCA, or HNCACB (Fig. 6) correlate three different NMR-active nuclei (e.g., along a protein backbone and are therefore used in suitable combinations to assign proteins) (45).

Multidimensional NMR experiments

Two-dimensional experiments may be extended to higher-dimensional experiments by including additional incremented evolution times into the pulse sequence separated by mixing sequences. The basic building blocks for these experiments are usually 2D homonuclear techniques like TOCSY (42) and NOESY (44) or the heteronuclear INEPT-transfer experiment (Insensitive Nuclei Enhanced by Polarization Transfer) (61). Combinations of these experiments alleviate the problem of spectral crowding and allow for new assignment strategies for proteins (45) and nucleic acids (62). Experiments such as TOCSY-HSQC and NOESY-HSQC combine homonuclear and heteronuclear techniques, whereas triple-resonance experiments such as HNCO, HNCA, or HNCACB (Fig. 6) correlate three different NMR-active nuclei (e.g., along a protein backbone and are therefore used in suitable combinations to assign proteins) (45).
induction of structural changes. This method allows immediate distinction of different binding sites and is a particular advantage of NMR screening experiments that is not possible in most, if not all, alternative-screening techniques. Hydrogen-deuterium exchange experiments provide information about solvent accessibility and participation in hydrogen-bonding networks. Changes in the line shape of NMR signals may elucidate folding intermediates and pathways, whereas relaxation studies probe thermodynamic and kinetic properties.

Chemical shift perturbations (CSPs)

The location of a ligand-binding site can be determined for small to medium-sized proteins that are uniformly 15N labeled and where the amide signals have been assigned by NMR spectroscopy. 1H-15N HSQC spectra are recorded in the absence and presence of a ligand (NMR titration experiments), and the observed changes in the amide chemical shift or signal intensity of the amino acid residues that bind the ligand indicate the location of the binding pocket.

This approach can be extended to larger targets by the TROSY experiment in combination with different labeling strategies (e.g., selective labeling of specific amino acids or deuteration). Deuteration replaces the nonexchangeable 1H atoms by 2H, which thereby reduces T2 relaxation significantly (68–70). To map a ligand-binding site by CSPs, it is not necessary to have a complete assignment of the protein. Per-atom information about binding events can be gained for large, only partially assigned proteins, in which structure calculation based on NMR data is not yet possible.

Arginine side chains are often involved in intermolecular contacts, therefore assignment of the 1H-15N side-chain groups allows mapping of ATP binding on kinases to be studied by CSPs as well as binding of RNA or DNA (71).

The relation between ligand binding and CSPs is not only dependent on the proximity of the ligand to the residues that show chemical shift changes. Conformational rearrangements of the protein/biomolecule as well as allosteric effects may cause significant CSPs of residues distant from the binding pocket. Reliable determination of the binding pocket is possible by comparing the CSPs induced by a series of structurally similar ligands (72).

An interesting example of the application of CSP studies is the ribosomal L11 protein (Fig. 7). Complex formation with its natural RNA substrate results in very large CSPs of the amide resonances that correspond to the binding surface on the C-terminal domain of the L11 protein. Addition of the antibiotic thiostrepton induces tighter binding of the N-terminal domain to the RNA, which again results in significant chemical shift changes (73).
A nuclear magnetic resonance (NMR) spectroscopy: overview of applications in chemical biology

Another interesting example of the effect of ligand binding on protein backbone $^{1}H-^{15}N$ signals is the p38$\alpha$ kinase. Certain ligands can bind not only to the ATP binding site but also to an allosteric hydrophobic pocket situated in an Asp-Phe-Gly (DFG) loop. This loop is highly conserved among kinases, and ligands that bind this allosteric site are called DFG-out ligands, as they require the kinase to undergo a conformational change. X-ray structures of p38$\alpha$ show that the Phe side chain of the DFG motive is displaced by 10 Å when bound to DFG-out ligands compared with the apo state (74, 75). This displacement is not observed when the kinase is complexed with ligands that only bind to the ATP binding site (DFG-in ligands) (76).

Investigations of this DFG-in/out phenomenon by NMR spectroscopy reveal dynamics that are not obtainable from static X-ray crystallographic studies (77). The kinase in its free form is in a slow DFG-in/out equilibrium. This equilibrium is not disturbed by a DFG-in ligand, whereas a DFG-out ligand disrupts the motion forcing the kinase into the DFG-out conformation. Cross peaks that develop from residues in the DFG-loop disappear because of conformational interconversion on the intermediate NMR timescale that leads to extensive line broadening. This mechanism is illustrated in (Fig. 8) of two $^{1}H-^{15}N$ TROSY spectra of p38 selectively labeled with $^{15}N$ phenylalanine. The first spectrum of the apo state contains cross peaks from 12 of 13 Phe residues. The missing peak corresponds to the Phe residue in the DFG-loop and only appears during addition of a DFG-out ligand as shown in the second spectrum, which proves that the conformational equilibrium of the free kinase is disrupted by this allosteric ligand.

This DFG-in/out phenomenon was first observed for the Abelson kinase where the ligand gleevec (78, 79) was found to bind both the ATP binding site and the allosteric binding pocket in the DFG-loop. This mechanism is observed not only for the Abelson and p38 kinases (74, 80, 81) but also for Raf (82) and KDR (83).

The lack of signals caused by the flexible nature of the kinase in the free state is observed for most of the activation loop as well as other highly conserved domains (84). The same pattern of missing peaks from these domains is also seen for the protein kinase A (PKA) (85).

SAR by NMR
To reduce the effort of chemical synthesis in drug discovery, the structure-activity relationship (SAR) method uses results from the chemical shift perturbation experiments (SAR by NMR). Optimizing the binding properties of initial hits from a high-throughput screening process without additional knowledge of the exact binding site and orientation is challenging. The SAR by NMR approach, which was developed by Stephen W. Fesik of Abbott Labs, is based on enhancing binding properties of small molecules to the protein surface by linking two active compounds together. First, smaller fragments are screened for binding, and initial hits are optimized toward stronger binding. Second, two compounds that bind to close, but not overlapping, sites on the protein surface are connected by a linker of appropriate length (Fig. 9). Strong dual-site binders can be generated from two initial relatively weak single-site binders with interaction sites close together on the protein surface by connecting them with a linker of suitable length.
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Figure 8 1H-15N TROSY spectra of p38α: (a) In the free state and (b) After binding a DFG-out ligand. The ligand structure is depicted in the upper left corner of the bound spectrum. The phenylalanine residue 169 situated in the DFG loop appears during binding. (c) Schematic representation of the effect of DFG-in ligand (SB203580) versus DFG-out ligand on the conformational exchange of p38 (1P38), the activation loop is indicated by an arrow.

Figure 9 Flow diagram of the SAR by NMR procedure. In case of completely independent binding of two ligands: Left: primary screening can be performed individually for each ligand. Ligands binding to two independent sites are linked, which enhances ligand-binding affinity. Right: for allosteric binders, screening for a second ligand should be done in the presence of a primary ligand.

Optimizing the linker is important because the free energy of binding for the composite ligand ΔG is not just the sum of the binding enthalpies for the individual ligands A and B (ΔH_A and ΔH_B, respectively), but it includes an intrinsic entropic penalty term ΔS depending on rigidity and orientation of the linker. The origin of this entropic loss is the conformational adaptation the ligand undergoes during binding.

Mechanistically, inhibition must not necessarily block the active site itself, but it can exert allosteric effects on the substrate-binding pocket, which thereby enhances or suppresses enzymatic activity. Additional considerations regarding enzymatic reactions are discussed in Reference 86. SAR by NMR has been successfully applied to various systems (i.e., for disrupting intracellular protein-protein binding (87) as well as cytokine-receptor interaction (88)). High-affinity enzyme inhibitors have been developed by this technique (e.g., for the metalloproteinase Stromelysin (89) and the protein tyrosine phosphatase 1B (90)).

Hydrogen/deuterium (H/D) exchange

Protein hydrogen atoms bound to N, O, or S are in constant exchange with solvent protons, and therefore they can be exchanged readily by deuterium. The exchange reaction can be followed by measuring a series of 1H-15N correlation spectra (e.g., HSQC, HMBC, or TROSY) after the solvent exchange. Residues that are buried in the core of the protein are not as solvent exposed as residues on the surface of the protein and will therefore exchange more slowly. Equally, protons that are
involved in hydrogen bonding are not as prone to exchange as non-hydrogen-bonded protons.

One of the most elegant experiments that uses the H/D exchange phenomenon is the characterization of protein folding by Miranker et al. (91). In addition, useful information can be obtained about protein-protein interactions, in which residues at the interaction surface will be solvent accessible before the complexation but not after. Domain reorientation of a protein during binding of a ligand can result in some amino acid residues being buried in the interior of the protein and therefore showing significantly different exchange times before and after binding.

Changes in secondary structure of the target protein during ligand binding can also be extracted from H/D exchange measurements as secondary structure elements, such as α-helices and β-sheets, which are highly stabilized by hydrogen bonds. This is not the case for the more flexible loop regions or domains.

Experimentally, H/D exchange for a stable globular protein sample can normally be performed by lyophilizing the sample and redissolving it in the same amount of D_2O as there was H_2O before lyophilizing to ensure that the buffer concentration remains.

The effect of this exchange reaction on a protein proton in the binding pocket will depend on the difference in Larmor frequency (chemical shift) between the free and bound form compared with the exchange rate. When exchange rate and difference in Larmor frequency are comparable, the reaction is in the intermediate exchange regime, and useful information can be obtained from line shape analysis. Slow intermediate exchange in which the exchange rate is slower than the difference in Larmor frequencies leads to cross-peaks from both the free and bound state. Fast intermediate exchange, however, produces only one single peak at the average chemical shift value. When the exchange rate and frequency difference are equal (the crossover point), the signal is broadened in between the two states.

Proteins at ambient temperature exhibit conformational dynamics on a wide range of time-scales (Fig. 10). And, studies of such dynamics have been shown to be particularly important to understand the pharmacological properties of kinases (vide infra).

Line shape analysis is used to study exchange processes as described by the relative population of the different states (mole fractions), line widths of the signals, and their frequency separation Δν (93). It is assumed that the lifetimes are significantly longer than the rotational correlation time τ_ω, and that spin-spin couplings do not interfere.

Starting from the simplest case of slow exchange between the free and ligand-bound form of an NMR signal from a biomacromolecule, this exchange situation will be indicated by two distinct peaks that do not change position but only vary in intensity. More complex exchange mechanisms may lead to overlaying changes in line shapes and can be simulated by models that are more complex. A nice example is the introduction of an additional fast exchange step between the free form and an intermediate form that subsequently is in slow exchange with the ligand-bound form. In this case, the peak that corresponds to the ligand-bound form will only increase in intensity during titration with the ligand. The peak of the free form, however, will not only decrease in intensity, but also change position toward the averaged peak position of free form and intermediate.

Even more elaborate peak patterns will develop if additional intermediates are in fast exchange with the free form or if alternative ligand-bound forms are present.

The simulated ^1H-line shapes of a ^15N HSQC spectrum of a protein signal during titration with an inhibitor are shown in (Fig. 11). The expected line shapes for three different binding mechanisms of increasing complexity have been simulated for a protein concentration of 0.1 mM. The ligand concentration increases from 0 to 0.1 mM (blue through red lines), and the ligand has a binding affinity of 1 μM. The line shapes for a direct key-lock mechanism are shown in (Fig. 11a). This mechanism is the simplest conceivable mechanism and the first model to use when fitting experimental NMR line shapes. Even when a simple key-lock mechanism is presumed, the line shape can look different depending on the off-rate of the reaction.

A slow reaction on a high μs time scale or slower leads the sharp peak of the free protein to disappear at the same time as a sharp peak of the complex appears. An off-rate in the μs time regime or faster will result in a line that shifts from its original position to the position of the bound form (Fig. 11a). The peak evolves from its free form to a broadened line shape that sharpens again at the end of the titration. Failure to reproduce the measured line shape by any parameter set of a certain model means that the complexity of the model has to be increased. Figure 11b shows the behavior expected from a single intermediate mechanism, whereas (Fig. 11c) presumes binding of the ligand in two different ways, of which only one species reacts to the complex structure. For each model, the input parameters, such as off-rates, are varied until the resulting line shapes fit the experimental data. In case of line broadening of the amide protons because of exchange with solvent water the line shapes of the nitrogen may also be simulated (93–96).

Thus, it is possible to calculate the off-rates of protein-ligand interactions observed for individual amino acid residues of the protein of a verified reaction mechanism. A nice example that shows various cases was recently published for the Apo-Cellular Retinol Binding Protein (97).
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Figure 10 The different processes amenable to NMR spectroscopy are indicated above the time arrow, below are the typical time windows for different molecular motions and events.

(a) (b) (c)

Figure 11 Simulation of line shapes of a reaction according to the binding model: (a) Two-state reaction. (b) Two-state reaction with one intermediate and (c) Two-state reaction with two possible intermediates. Protein concentration is 0.1 mM. Titration in equal steps until an equimolar amount of ligand is reached (0, 0.01, 0.02, ..., 0.1 mM). The signals of P and PL are separated by $\Delta \nu = 150$ Hz. The sequential steps in the titration change color from black to gray.

An example of dramatic change of line shapes is observed when the p38$\alpha$ kinase is bound to the previously mentioned DFG-out ligand. p38$\alpha$ is activated by dual phosphorylation of two conserved residues in the activation loop. The $^{31}$P spectrum of the apo state (Fig. 12) shows two broad peaks; during addition of a DFG-out ligand, one of these is shifted. Both peaks show significant narrowing, which indicates that ligand binding hinders the flexibility not only of the DFG motif but also of the entire activation loop, as the phosphorylation sites are situated 11 and 13 residues downstream of the amino acid chain from the DFG loop.

Line shape analysis may be complemented or verified by measuring chemical exchange by NMR by relaxation dispersion experiments. Relaxation dispersion is based on measuring a series of CPMG-based relaxation rates at different temperatures. Excited state intermediates in folding reactions or ligand-accessible intermediates can thereby be probed, even when they constitute as little as 1% of the entire population.

Dynamic information from relaxation studies

The heteronuclear relaxation properties as measured by $T_1$, $T_2$, $T_1^*$, and hetNOEs contain valuable information about molecular dynamics of biomolecules.

The relaxation data shown in (Fig. 13a) describe the dynamics of the ribosomal protein L11 before (blue) and after (red) complexation with the GTPase region of 23S rRNA and of the subsequent binding of the antibiotic thiostrepton to the protein-RNA complex (in green). L11 consists of an N- and a C-terminal domain. The latter shows a flexible loop region in the free form around residues 86-96, which is evident by small values of $^{1}H$-$^{15}N$ hetNOEs, large $T_2$ values, and correspondingly low order parameters. During binding to 23S rRNA, this loop region shows a significant reduction of both hetNOEs and $T_2$ relaxation times, which indicates a rigidification of the region. This finding is shown in (Fig. 13a) with blue arrows. The overall correlation times that describe random tumbling of the C- and N-terminal domains are similar in the free form.
of the protein, which indicates that the protein domains tumble together more or less like a rigid body. During addition of RNA, both domains show a reduction of the correlation times, which indicates binding of the RNA to the protein; however, the effect is much more pronounced in the C-terminal domain, which indicates that this region is most tightly bound. When adding thiostrepton to the complex, the $T_1$ relaxation times of the N-terminal domain match the ones of the C-terminal domain, which indicates that the thiostrepton locks the N-terminal domain to the RNA, thus creating a stable protein-RNA-ligand complex. The highly flexible parts of the protein are shown on the structures below the relaxation data (Fig. 12b). The color coding goes from yellow for highly flexible over gray to blue for "nonflexible" regions.

Obtaining such detailed information about the dynamics of a complexation reaction is labor intensive. To obtain these data, it was necessary first to assign the protein in its free form as well as in complex with RNA and RNA-antibiotic by a series of 3D NMR spectra of $^{31}$P, $^{1}H$, $^{13}$C, triple-labeled samples. The labor intensity of dynamic studies like this is, however, offset by the invaluable insights into the detailed dynamic driving forces behind biological processes uniquely obtainable by NMR.

From dynamics to thermodynamics

For a comprehension understanding of ligand binding, the thermodynamics of the association process must be taken into account. On this topic, we follow the line of discussion of Steve Homans (99). The binding constant is related to the Gibbs free energy of binding in the following manner: $\Delta G = -RT \ln K$, where $\Delta G$ is composed of an enthalpic and an entropic term ($\Delta G = \Delta H - T\Delta S$), the former representing the structural features and the latter constituting the changes of dynamic properties of ligand, protein, and solvent. Although the binding event is mainly enthalpy driven, the entropic component should not be underestimated. Using only the enthalpic term as a measure of binding strength will lead to incorrect evaluation of the data. In some cases, the entropic term accounts for a difference in Gibbs energy that will distinguish between a mediocre and a tight binder (about 4 kcal mol$^{-1}$). Microscopic techniques such as isothermal titration calorimetry can determine the overall thermodynamic parameters of binding events, whereas a microscopic view on the mechanism can only be obtained by NMR. Using the relaxation methods described above, a per-residue analysis of protein dynamics is possible. Especially for ligand binding, no other technique will provide such detailed account of the binding mechanism. For the major urinary protein, an interesting effect of ligand binding has been observed (100). The increase in protein backbone flexibility during ligand binding results in an entropic contribution to the Gibbs free energy of binding of a magnitude comparable with other driving components. The elucidation of such processes in microscopic detail that NMR can provide may lead to a reassessment of the governing factors of ligand affinity, as this might well be a general mechanism of small ligand binding.

Not only are protein backbone dynamics accessible to relaxation studies, but also side-chain dynamics can be measured. Different techniques that quantify the relaxation of $^{13}$C or $^1H$ nuclei are also available (101-104). The focus usually lies on the dynamics of methyl groups that require elaborate isotopic labeling schemes to obtain either $^{13}$C$^2$H$^2$O or $^{13}$C$^2$H$^2$O$^2$ isotopomers (105). This finding has enabled determination of the order parameters of methyl group rotation axes. In the case of calmodulin, which is a calcium binding protein, it has been shown that complex formation with a peptide leads to nonuniform entropy changes along the protein backbone and side chains. Whereas the backbone only shows small changes in its motional characteristics, some side chains close to the binding site are restricted in their motion. Simultaneously, the motional entropy of more remote side chains increases, which thereby compensates for the entropy losses at the binding site (106, 107).

A nice example of the power of dynamic NMR spectroscopy is the work by Kem and coworkers (108, 109) on the action between enzymatic activity and the kinetics of structural dynamics in the coupled network of proteins. These NMR methods provide greater insight into the complex thermodynamic mechanisms of molecular recognition, which thereby supplies the field of Chemical Biology with unique and invaluable information.

Ligand-observed experiments

Analysis of binding events by observing the NMR signals of the ligand rather than the biomolecular target is an alternative to protein-observed experiments. This approach is attractive to screen large libraries of compounds in a cheap and efficient manner, because expensive isotopic labeling of target or ligand is not required. It is even possible to screen mixtures of ligands using distinct spectral features of each ligand to identify the biologically active compound. Another advantage of ligand-observed drug screening is that it allows all biological targets, regardless of size, to be studied by NMR spectroscopy. This approach is especially useful when dealing with cells, viruses, or membrane proteins. $\alpha$-Coupled receptors and ion channels are membrane proteins that are highly interesting as drug targets but are difficult to express in high yields, which renders isotopic enrichment of such proteins costly. Furthermore,
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Figure 13 Relaxation parameters of protein-RNA-ligand interactions of the L11 system. Free L11 (blue), L11, and RNA (red) and L11, RNA, and Thiostrepton (green). RNA binding shows effects on all parameters, pronounced for certain residues in the C-terminal domain (arrows), whereas the additional Thiostrepton binding leads to an increase in $T_1$ of the N-terminal domain. Binding of RNA locks the two domains in a fixed orientation relative to each other and strongly restricts the motion of the C-terminal domain. See color insert.

the required presence of phospholipids in samples of membrane proteins increases the size of the complex, which restricts the use of protein-observed measurements even more because of fast $T_2$ relaxation (63). Contrarily, size restrictions do apply to the ligand, as the distinction between ligand and target is based on their difference in molecular weight. Therefore, ligands must be much smaller than the target. The binding event of a ligand to a biological target will transfer magnetization from target to ligand, which alters the NMR spectrum of the ligand. To observe binding of a ligand, it should usually be in medium to fast exchange with the target that corresponds to moderate binders with $K_D$ values in the mM to $\mu$M range.

Saturation transfer difference NMR (STD-NMR)

STD-NMR (110) observes magnetization transfer from protein to ligand molecules that are in fast exchange between the free and bound state. The protein is excited selectively by irradiating at a frequency specific to protein signals, typically the protein side-chain protons. A large excess of ligands ensures that the fraction of ligand molecules bound to the target at any one time is negligible. The transfer of magnetization from a large saturated protein to a bound ligand is, however, very efficient, and a fast $k_{off}$ rate of the ligand ensures that a measurable fraction of the free ligand has been bound to the protein during the course of the experiment. This fraction remains saturated throughout the experiment and does not contribute to the ligand NMR spectrum, which reduces the signal intensity. The fraction of ligand molecules that have been bound to the protein during the experiment can be calculated from the difference of two 1D spectra with and without protein saturation. The large excess of ligands also ensures that the amount of ligand molecules that will bind to the target more than once is negligible. The STD effect of a ligand-binding event is strong when the protein is large, even if the ligand only transiently binds to the protein, as the magnetization transfer depends on the molecular weight of the macromolecular complex and is very efficient for large complexes but inefficient for small molecules. STD-NMR cannot readily distinguish between specific and nonspecific binding events. Therefore, an extension of STD-NMR involves the titration of a protein-inhibitor complex with a test ligand that will compete with the inhibitor in fast exchange for the same receptor-binding site on the protein. This situation
the kinase-reporter-ligand sample decreases the STD reporter signals in comparison with the reference spectrum in the absence of any competitor.

Target protein are accented by a grey box. The test ligand competes with the reporter ligand for the binding site.

(a) Stepwise addition of the test ligand to the NMR sample. The resonances of the reporter ligand deplete faster when titrated with the more potent ligand B as observed from the fitted curve of the titration data (Ligand A: stars, Ligand B: squares).

(b) Analogous procedure with a more potent ligand.

(c) The resonances of the reporter ligand deplete faster when titrated with the more potent ligand B.

The competition between an inhibitor of known binding properties to the target will dominate the spectrum.

In a fast-exchange situation with a large excess of ligands compared to receptors, the $K_D$ value derived from NMR experiments is comparable with the $IC_{50}$ value (i.e., the concentration of the inhibitor required to reduce the target activity by 50%). The $IC_{50}$ values are usually derived from biochemical assays; from an inhibitor of known $IC_{50}$ value, it is therefore possible to derive relative $IC_{50}$ values for compounds in competition with this inhibitor by NMR spectroscopy.

The competition between an inhibitor of known binding properties, which is also called a reporter-ligand, and a ligand for the same binding site is used in NMR reporter assays as illustrated in (Fig. 14). The reporter ligand must be highly soluble and have strong, quantifiable NMR resonances. Competing ligands are titrated to the NMR sample that contains a target protein and reporter ligand, and the reduction of the reporter signals corresponds to the fraction that has been displaced from the binding site by the competing ligand. By selecting a reporter-ligand of low affinity toward the target protein, it is even possible to detect binding of high-affinity or poorly soluble ligands (112-114), as well as ligands in slow exchange with the target (115).

Other methods similar to the STD experiment are WaterLOGSY (Water-Ligand Observed via Gradient Spectroscopy) (116) and NOE pumping (117). WaterLOGSY is based on the transfer of magnetization from biomolecule to ligand via bulk water, whereas NOE pumping relies on first suppressing all ligand signals and then transferring magnetization via an NOE experiment from biomacromolecule to the fraction of ligand molecules that have been bound.

Exchange-transferred NOE spectroscopy (et-NOESY)

Similar to the STD experiment, exchange-transferred NOE spectroscopy (et-NOESY) provides information about a binding event between a small molecule and a high molecular weight biomolecule, even when the biomolecule itself is not amenable to NMR studies because of broad line widths, low solubility, or fast $T_2$ relaxation.

The principle of et-NOESY (118-120) is applied to low-affinity protein-ligand complexes where exchange is fast on the NMR timescale. High molar excess of the ligand is used (typical ligand to receptor site ratios range from 10 to 50); therefore, the resonance line shapes are governed by the properties of the unbound ligand in solution (i.e., sharp signals caused by slow $T_2$ relaxation). The et-NOESY observed, however, are caused by the conformational change of the ligand in the binding pocket. The efficiency of magnetization transfer caused by NOE effects depends on the overall correlation time of the molecular complex and becomes more efficient for larger complexes. For small molecules, the NOE is positive and weak, whereas it is negative and strong for large biomolecular complexes. Therefore, the cross-relaxation measured by the NOE transfer within the ligand when it is bound to the target will dominate the spectrum.

The NOEs from the bound state are transferred into the free state of the ligand when it leaves the binding pocket. To observe NOEs from the bound state, the off-rate ($K_{off}$) of the ligand must be much faster than the cross-relaxation rate, only then is the initial build up of the NOE proportional to the cross-relaxation rate (121). Based on the internuclear distances obtainable from

\[
K_I = \frac{c_I (K_0 - K_D)}{1 + K_0 + K_D}
\]
the NOEs it is possible to deduce the conformation of the ligand when bound to the target.

The inversion of sign between NOEs from small and large molecules simplifies measurement of et-NOEs for binding studies of small ligands < 1 kDa and large targets. For larger ligands, such as ligand-peptides, two et-NOESY spectra are recorded in the presence and absence of target protein. By subtracting the two spectra, it is possible to distinguish between NOEs that develop from target-bound ligands and ligands that have not been bound.

The combination of STD-NMR and et-NOESY experiments provides information on ligand protons in direct contact with the binding pocket as well as the overall ligand structure. The two experiments thus complement each other in elucidating the binding surface of a target of otherwise unknown structure.

**NOEs and isotope-filtered NMR experiments**

In the strong binding regime, the measurable NOEs between the biomolecule and the bound ligand can be exploited to derive the 3D complex structure. A common approach involves uniform labeling of the biomolecule with $^{15}$N and $^{13}$C isotopes, whereas the ligand remains unlabeled (122). To distinguish between the intra-molecular NOEs of the biomolecule, the intra-molecular NOEs of the ligand and the intermolecular NOEs that occur at the binding interface, isotope-filtered NOESY experiments are used. The proton magnetization that originates from either $^{1}$H-$^{13}$C or $^{1}$H-$^{15}$N atom pairs is either filtered out or selected (123, 124).

The 3D $^{13}$C- and $^{15}$N- filtered NOESY experiments are used to identify NOE restraints within the biomolecule itself. Intra-molecular NOE restraints of the ligand can be assigned from the $^{12}$C, $^{15}$N-filtered NOESY experiment, which suppresses the signal of $^{13}$C- and $^{15}$N-attached protons of the biomolecule. The intermolecular NOEs between the ligand and biomolecule can be detected by a combination of selection and filtering out of proton magnetization in the 3D $^{12}$C, $^{15}$N-ω-filtered NOESY $^{1}$H, $^{13}$C HSQC experiment, which observes the interface between the two molecules as observed in the 2D projection in Fig. 15.

The information derived from the analysis of this example was used as distance restraints for calculation of the 3D structure of the complex of calmodulin, a calcium-binding protein, and a peptide ligand. The amino acid sequence of the peptide ligand, C20W, corresponds to the N-terminal part of the calmodulin-binding domain of the plasma membrane calcium pump (125).

**Structure calculation based on NMR data**

To provide new insights in the biological context of biomolecules, obtaining structural information is frequently essential. The tertiary structure of molecules with accuracy at the atom level can be calculated using NMR-derived data on intra-molecular atom distances and angles. Many biomolecules form complexes with each other, and the nature of these complex formations can be deduced from intermolecular distances and determination of the interaction surfaces in combination with computational docking procedures to determine their global structure.

![Figure 15](image_url)

**Figure 15** 2D $^{1}$H, $^{1}$H projection of the $^{12}$C, $^{15}$N-ω-filtered 3D NOESY $^{1}$H, $^{13}$C HSQC spectrum that contains only intermolecular NOEs between the double-labeled protein (Calmodulin, resonance assignment in black) and a bound peptide ligand (in light grey). Only contacts between the protein and the peptide ligand are observed. Assigned cross-peaks are indicated in the spectrum.
Simulated annealing

Traditionally, the structure of biomolecules and macromolecular complexes are solved by NMR using proton-proton NOEs that are translated into distances. In addition to NOEs, the structures may be refined using additional NMR data, such as coupling constants, RDCs, chemical shifts, and hydrogen bonds. Many programs and protocols are available for calculation of NMR structures (e.g., XPLOR-NIH (126), CNS (127), CNX (Accelrys, San Diego, USA), CYANA (128, 129), and ARIA (130)). Structural calculations using simulated annealing (SA) and molecular dynamics (MD) require a forcefield with the molecular topology and parameters to define the atomic masses, bond lengths, charges, angles, planarity, and connectivity information of the entire macromolecule (131). Most biomolecular NMR structures are calculated by SA with MD in either Cartesian or torsion angle coordinate space. The torsion angle dynamics protocol allows simulation of longer time intervals and higher temperatures than dynamics in Cartesian space and is thus very efficient for the calculation of NMR structures (128, 132). The SA algorithm generally starts from an extended conformation of the macromolecule and proceeds from a high temperature stage over slow cooling-down stages to a final energy minimization at low temperature. This protocol is commonly repeated multiple times in an iterative way to recalibrate the NOE-based distances. The final NMR structures are often refined in explicit solvent to improve the structure quality significantly (133).

HADDOCK

The conventional NMR approach for determining biomolecular complex structures is based on a time-consuming process of collecting intermolecular NOEs and complementing RDCs. A newer approach uses the easily obtainable chemical shift perturbations from NMR titration experiments in combination with computational docking. The HADDOCK (High Ambiguity Resolved Docking) approach uses the experimental NMR data of the intermolecular interactions as Ambiguous Interaction Restraints (AIRs) to drive the docking process (134). Residues that show significant chemical shift changes during interaction are designated as active. Residues that show less significant changes and/or are surface neighbors of the active residues are called passive. The AIR is accordingly defined as an ambiguous intermolecular distance between any atom of an active residue and any atom of an active or passive residue of the partner molecule in the complex. The HADDOCK approach uses CNS for structural calculations and protocols derived from ARIA for automation. Generally, the docking protocol starts with calculating many random orientations in which the partner molecules are positioned far away from each other. After a rigid-body docking, the best scoring solutions in terms of intermolecular energies are refined in torsion angle space using a multistage, semi-rigid simulated annealing protocol. Subsequently, the structures are refined in explicit water in Cartesian space. Next to protein-protein docking, HADDOCK has also been applied to model protein-DNA, protein-RNA, protein-oligosaccharide, and protein-ligand complexes (135, 136). The HADDOCK protocols allow the usage of additional NMR data, such as RDCs (137) and diffusion anisotropy data (138). Furthermore, HADDOCK is able to perform solvated docking (139), as water molecules in the interaction interface play an important role in mediating biomolecular interactions. These implementations lead to improved accuracy and quality of the final complex structures. The HADDOCK method can calculate models of multicomponent complexes and can easily be applied to drug design.

RDC-based structure refinement

RDCs are especially useful for determining the relative orientation of different partner molecules in a complex or of different domains within a single macromolecule in a complex. When two domains of a macromolecule are connected by a flexible linker and NOE data between those domains is scarce or lacking, RDCs can be used to determine the exact orientation of the domains with respect to each other. In the case of ribosomal L11 protein, these additional orientational restraints have been used to solve the structure of the protein in its free form as well as in complex with its RNA binding partner and an antibiotic (vide supra) (73, 140). The relative orientation of the two domains that are connected by a short linker and oriented in solution (Fig. 16). Using 3668 distance restraints (26.4 per residue), 77 TALOS-derived dihedral angle restraints, and 102 J-couplings derived from the backbone Hα-HN of the structure of the two domains of the ribosomal L11 protein.

Figure 16. Structure calculation purely based on NOE and dihedral angle restrictions often fails to determine proper domain orientations. By incorporating the long-range orientation parameters that RDCs provide, it is possible to overcome this problem and to define relative orientation of the two domains of the ribosomal L11 protein.
be determined unambiguously because of the limited amount of interdomain NOEs. To calculate the overall structure of the full-length protein, an additional 429 RDCs were measured and incorporated in the structural calculation. This step allowed the exact determination of the relative orientation of both domains (73). This result is in good agreement with the fact that the two domains tumble together as a rigid body, based on the dynamic measurements (Fig. 13).

The same method has been applied to measure the L11 domain orientation when the protein is in complex with its RNA partner or both RNA and thiostrepton antibiotic. The additional RDCs revealed a rearrangement of the N-terminal domain of L11 placing it closer to the RNA after binding of thiostrepton. HADDOCK has been used to calculate a model of the ternary structure of the L11 protein in complex with RNA and antibiotic. Based on the orientational data, the dynamics and the docking model, it seems that thiostrepton locks the domain conformation of L11 in a rigid (inhibitory) state. The antibiotic thiostrepton interferes with the interaction of elongation factors to this L11-RNA region, which has a dramatic effect on the level of protein synthesis by the ribosome.

Oriental restraints, which are derived by paramagnetic labeling of a single domain of a protein, are especially important for the analysis of domain-domain interactions (141). The paramagnetic tag induces anisotropic alignment, which is scaled for the two domains of the protein according to their relative domain mobility.

**LIGDOCK and SOS-NMR**

In the weak binding regime, intermolecular NOEs between biomolecule and ligand are usually not exploitable in protein-observed experiments. Instead, other NMR-derived information is used as an experimental restraint for molecular docking procedures to solve the protein–ligand structure. The readily obtainable 1H chemical shift perturbations (CSPs) of the protein amide resonances caused by ligand binding are often used as experimental restraints. During the molecular dynamics simulation, penalty energies calculated from these experimental restraints together with conventional intermolecular van der Waals energies or electrostatic interactions guide the protein–ligand complex toward an energy minimum. This minimum represents a physically reasonable structure that fulfills a maximum number of restraints introduced by NMR-derived information. Accuracy of the NMR structure, as judged by comparison with observed experiments. Instead, other NMR-derived information is used as an experimental restraint for molecular docking procedures to solve the protein–ligand structure. The readily obtainable 1H chemical shift perturbations (CSPs) of the protein amide resonances caused by ligand binding are often used as experimental restraints. During the molecular dynamics simulation, penalty energies calculated from these experimental restraints together with conventional intermolecular van der Waals energies or electrostatic interactions guide the protein–ligand complex toward an energy minimum. This minimum represents a physically reasonable structure that fulfills a maximum number of restraints introduced by NMR-derived information.

**SOS-NMR** is a method for measuring the L11 conformation of thiostrepton. HADDOCK has been used to calculate a model of the ternary structure of the L11 protein in complex with RNA and antibiotic. Based on the orientational data, the dynamics and the docking model, it seems that thiostrepton locks the domain conformation of L11 in a rigid (inhibitory) state. The antibiotic thiostrepton interferes with the interaction of elongation factors to this L11-RNA region, which has a dramatic effect on the level of protein synthesis by the ribosome.

Oriental restraints, which are derived by paramagnetic labeling of a single domain of a protein, are especially important for the analysis of domain-domain interactions (141). The paramagnetic tag induces anisotropic alignment, which is scaled for the two domains of the protein according to their relative domain mobility.

**LIGDOCK and SOS-NMR**

In the weak binding regime, intermolecular NOEs between biomolecule and ligand are usually not exploitable in protein-observed experiments. Instead, other NMR-derived information is used as an experimental restraint for molecular docking procedures to solve the protein–ligand structure. The readily obtainable 1H chemical shift perturbations (CSPs) of the protein amide resonances caused by ligand binding are often used as experimental restraints. During the molecular dynamics simulation, penalty energies calculated from these experimental restraints together with conventional intermolecular van der Waals energies or electrostatic interactions guide the protein–ligand complex toward an energy minimum. This minimum represents a physically reasonable structure that fulfills a maximum number of restraints introduced by NMR-derived information.
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology

Figure 17. Molecular docking guided by NMR-derived information. (a) The complex structure of PTP1b with a ligand (modeled structure in green/yellow) calculated from CSPs only corresponds well with the X-ray reference structure (blue). (b) The cluster of lowest energy structures calculated for p38α and ligand SB203580 shows two possible orientations (c) that differ in their ability to form a hydrogen bond to the Met109 amide group. Additional information is needed to clarify the binding mode. See color insert.

Figure 18. SOS-NMR using deuterated and selectively protonated protein can resolve ambiguity of ligand orientation. (a) Bottom: 1D 1H NMR reference spectrum that shows the resonances of the ligand SB203580. Middle: 1D 1H STD spectrum recorded in the presence of p38α. Top: 1D 1H STD spectrum recorded of a sample prepared according to the SOS procedure of perdeuterated kinase and unlabeled isoleucine residues. A pronounced STD effect for the H1 and H4 protons and only weak effects for the H6 and H5 protons resolves ambiguity of ligand orientation in the binding pocket. (b) Distance analysis of the reference X-ray structure.

especially mass spectrometry, NMR is a very valuable tool for this task. The role of NMR has been evaluated for analytical strategies in metabolomic research (151) and pharmaceutical discovery (152, 153).

Magnetic resonance imaging has long been used for clinical diagnostic purposes, often in combination with relaxation-enhancing contrast agents. Recent development of lipid nanoparticle systems as carriers of relaxation enhancers to specific targets (154) makes imaging methods applicable to the field of chemical biology.

References


Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology


110. Morris BM. Characterization of ligand binding by saturation transfer difference NMR. Spectrosc. 1999;1764-1786.
Nuclear Magnetic Resonance (NMR) Spectroscopy: Overview of Applications in Chemical Biology


Supported Lipid Bilayers: Development and Applications in Chemical Biology

Volker Kiesling, Marta K. Domanska, David Murray, Chen Wan and Lukas K. Tamm,
University of Virginia, Charlottesville, Virginia

doi: 10.1002/9780470048672.wecb663

Lipid bilayers supported on solid substrates were developed almost a quarter century ago as a new model membrane system to study fundamental properties of biological membranes and their constituent lipid and protein molecules, as well as for numerous practical applications. In this review, we summarize the development of supported bilayer-based model membrane systems of increasing complexity while always keeping in mind their primary purpose, which is to reproduce the complex fluid structure of biological membranes to mimic cell membrane-based molecular processes as close to physiological reality as possible. This process must include maintaining the structure and fluidity of the lipid bilayer, as well as the structure and function of reconstituted membrane proteins. We start the review with a general description of a contemporary picture of a generic biological membrane and proceed with an account of how supported membranes recapture different physiological aspects of biological membranes, such as membrane fusion, cell adhesion, or membrane permeability. Supported membranes are particularly good models for reproducing the fluidity of cell membranes even when different liquid lipid phases coexist. They are also superb models for measuring ligand-receptor interactions between membrane-bound receptors and soluble or membrane-bound ligands. The chemical biology of these and many other lipid–lipid, lipid–protein, and protein–protein interactions that have been investigated in supported bilayers are summarized in several sections that form the main body of this review. The article closes with summaries of practical procedures to prepare and characterize supported bilayers.

Since the lipid bilayer was accepted as the basic principle of design for biological membranes more than 40 years ago, researchers have developed model systems to study numerous aspects of biomembrane structure and function. Large multilamellar liposomes have been used frequently in early X-ray diffraction and calorimetric studies that characterize the structure and thermodynamic properties of lipid bilayers. Sonication of these structures leads to small unilamellar vesicles, which are excellent model systems for numerous spectroscopic and binding studies with model membranes. Other experiments require larger or even “giant” unilamellar vesicles, produced by extrusion through filters or swelling from solid lipid deposits with or without applied electric fields. For some applications, such as recording electrical currents across membranes or measuring the mobility and lateral organization of membrane constituents, planar bilayer geometry is advantageous. To this end, single unsupported planar bilayers are suspended with or without solvent in a small orifice in a Teflon septum that separates two chambers that are electrically connected to equipment that can record currents from single molecular channels (1, 2). A wealth of information on the molecular properties of ion channels, toxins, and other membrane-interactive proteins and substances has been obtained using such planar bilayer recordings.
A different planar model membrane design was developed in the early 1980s mainly from research conducted in Harden McConnell’s laboratory at Stanford University—the supported lipid bilayer. The biological motivation driving these early studies was to provide planar model membranes for observation by epi-fluorescence microscopy and to use them as surrogate targets for studying molecular interactions in immunological synapses between antigen-presenting cells and T-cell receptor-bearing lymphocytes. A precursor of the supported lipid bilayer has been the supported lipid monolayer, in which a lipid monolayer is deposited on a long-chain alkylated substrate (3). This early design using silane chemistry for alkylation was later superseded with sulfur chemistry on slides with thin evaporated gold surfaces. However, because these systems mimic only half a bilayer and cannot accommodate integral membrane proteins, which provide so many key functions to biological membranes, they are not considered in this article.

In a quest to produce a model system that more faithfully resembles biological membranes and that also allows the functional reconstitution of integral membrane proteins, we have developed the supported bilayer as a new model membrane system (4, 5). Supported bilayers have since become widely used in many different areas of chemical, biological, and materials research. A plethora of fundamental studies of bilayer and membrane properties to their phase behavior; membrane protein binding; membrane protein structure; electrophoresis and microanalysis of membrane components; functional studies of membrane receptors, pores and channels; interactions with components of bound cells, viruses, and cellular organelles; and so on. This article is organized in four main sections: We will first briefly review current concepts of cell membranes and how supported bilayers can be used as models for cell membranes. We will then describe biophysical properties and chemical interactions that govern the structure, stability, and use of supported bilayer membranes in biological analysis. The article will close with two sections on procedures that are commonly used to prepare and characterize supported bilayers.

**Supported Bilayers as Models for Cell Membranes**

Biological membranes are formed by a lipid bilayer with embedded proteins. The bilayer-forming lipids have polar headgroups that face the aqueous surroundings of the membrane as well as two hydrophobic tails that face the interior of the membrane. A key feature of cell membranes is that they are both fluid but still highly ordered in the membrane plane, as has been captured in the early conception of the fluid mosaic model of biological membranes (6). Lipids move laterally at a fast rate with a lateral diffusion coefficient on the order of 1 μm²/s, but they cross the membrane by flip-flop only once every few hours (7). The slow kinetic rate of lipid flip-flop allows the cell to establish asymmetry across the bilayer, which, with the input of energy, can be maintained for the entire lifetime of a cell. Membrane proteins come in different varieties. Some have sequences that cross the membrane, and therefore they are called integral membrane proteins. Other membrane proteins, such as the peripheral membrane proteins, are attached to the membrane surface either electrostatically or by means of covalent linkage to single or multiple lipid tails. A third class of membrane proteins dip into only one leaflet of the bilayer, where they are stabilized by hydrophobic interactions with the lipids in the bilayer. Regardless of the specific nature of membrane attachment and incorporation, membrane proteins can diffuse relatively rapidly in the plane of fluid membranes. However, they never cross the lipid bilayer except during biogenesis, which requires specialized machines called translocons that translate newly synthesized polypeptide chains across cell membranes in a complicated and still poorly understood energy-requiring process (8).

The notion of a biological membrane represented by a uniform sea of lipids with proteins freely floating in this environment has been significantly revised in the last decade or so (Fig. 1a) (9). It has become increasingly clear that biological membranes are laterally structured even if they are still predominantly fluid with regard to in-plane diffusion of protein and lipid components. First and foremost, membranes are highly crowded with proteins. Although the crowding varies greatly between different biological membranes (e.g., being very high for energy-converting membranes in mitochondria and photosynthetic organelles and rather low for membranes of myelin sheaths that wrap around and electrically insulate neurons), it is clear that membrane proteins do not act independently in membranes but often form clusters of interacting species. In a typical average cell membrane, about a quarter to half of the total cross-sectional area may be occupied by protein. If one considers proteins that have larger extramembranous domains than transmembrane domain cross-sections, the area fraction covered with protein may reach up to one third of the entire membrane surface. Thus, membrane proteins and lipids are both solvents and solutes at the same time. An alloy of hydrophobic proteins and lipids may describe the true nature of biological membranes more accurately than a dilute two-dimensional solution of proteins in a vast sea of lipids. Second, unlike commonly prepared model membranes, biological membranes contain thousands of different lipid species (10). The lipid species distinguish themselves by many different headgroup classes and chain compositions as well as an expansive combinatorial diversity of these structural elements. A very large amount of data has been accumulated over the past four decades on the structures, energetics, and phase properties of pure single and two- or three-component lipid bilayers as summarized in two monographs (11, 12). We know from this data that high- and low-melting lipids tend to separate into different domains than transmembrane domain cross-sections, the area fraction covered with protein may reach up to one third of the entire membrane surface. Thus, membrane proteins and lipids are both solvents and solutes at the same time. An alloy of hydrophobic proteins and lipids may describe the true nature of biological membranes more accurately than a dilute two-dimensional solution of proteins in a vast sea of lipids.

Even though cholesterol (substituted by other steroids in plants) is chemically classified as a lipid, the molecule is better grouped into a special category, which is distinct from membrane proteins and (phospho- and sphingo-) lipids, when discussing its role in biomembrane structure and function. Cholesterol constitutes between 25% and 40% of the total lipid plus cholesterol fraction of most typical cell membranes. These numbers translate into a 33–66% fraction of the noncholesterolic lipids, or considering its smaller size, about a 10–20%
...are referred to as liquid-ordered (lo) phases. Cholesterol-rich phases exhibit more chain order and there-...
environments. Another area where supported bilayers offer distinct advantages is studies of the kinetics of ligand binding to membrane-bound receptors (22). Supported membranes have also been used successfully to measure protein-mediated fusion of vesicles, for example in virus entry (23) or exocytosis (24).

On the analytical practical side, there is much interest in using supported membranes with incorporated ion channels as biosensors for various analytes (25–27). Another interesting application is to separate membrane components in situ by electrophoresis in supported membranes (28, 29). Such techniques could have a significant future impact on the proteomics of membrane proteins, that is, an area that is currently underdeveloped because of difficulties with appropriately separating membrane proteins in complex mixtures by more standard techniques. Finally, supported membranes continue to provide very interesting engineered substrates that mimic natural cell surfaces to trigger the differentiation of adjacent cells such as in the previously mentioned immunological synapses (30).

**Interactions Between Lipid and Protein Components in Supported Bilayers**

A key feature of biological membranes is their graded fluidity. Any useful membrane model system must strive to preserve the characteristic fluidity of lipid and protein components. This concept clearly has been a great challenge in this field. Although it is relatively easy to maintain the fluidity of the lipid components and peripherally attached protein components as demonstrated already in the earliest publications on supported bilayers (5, 31), to achieve the same result with integral membrane proteins has proven to be much more difficult. It is known that the gap between supported membranes and solid glass or quartz supports is on the order of 1 to 2 nm (32, 33), and this gap is filled with a layer of water that is sufficient to lubricate the lower leaflet of the bilayer to permit rapid lipid diffusion (Fig. 1b). However, this gap is not large enough to accommodate integral membrane proteins with significant extramembrane domains. Such proteins are generally immobilized by interaction with the glass support. Much research activity has been directed in the last decade to uncouple such unwanted non-physiological interactions, by increasing the gap distance with intercalated water-soluble polymers, either by simple physisorption or covalently attached to one or both surfaces by forming tethers between the substrate and the membrane (Fig. 1c). In the following section, we summarize strategies that have been successful for creating polymer-supported bilayers. The subsequent sections proceed in turn to brief summaries of lipid-lipid interactions, lipid-protein interactions, and protein-protein interactions in supported bilayers, and then to studies of membrane fusion and cell adhesion using supported membranes as a model for one membrane in these membrane-membrane interactions. The final sections summarize recent advances in patterning supported membranes and using supported membranes as novel analytical tools in chemical biology. Depending on the particular application, bilayers directly supported on solid substrates or (tethered) polymer-supported bilayers are preferred as will be mentioned in each case.

**Polymer-supported bilayers with and without tethers**

The idea of using polymer-supported bilayers has been around for more than a decade (34), but it became practical for chemical and biological applications only more recently. Early versions have used relatively short distances to link the monolayer to the solid substrate and thereby increase their durability for practical applications (25, 35). Because these approaches do not increase the gap distance between substrate and membrane and therefore have not been used to reconstitute integral membrane proteins functionally, they will not be discussed here.

Our group has developed a polyethylene glycol (PEG)-based polymer support for bilayers on glass, quartz, or oxidized silicon (36). PEG that consists of 77 subunits bridges a phospholipid on one end and a silane group on the other end, which allows the covalent tethering of the lipopolymer to free silanols on the silicon dioxide surface. In all, 25% of the reconstituted integral membrane protein cytochrome b diffused freely in supported bilayers that contained 3% of the lipopolymer in the proximal leaflet. Integral membrane SNARE proteins were 80% mobile with a diffusion coefficient of 0.8 μm²/s in this system (37). The distance between the substrate and membrane increased from 2 to 4 nm without and with the polymer, respectively (33). A similar approach was introduced by Naumann’s group, who used 3-octadecylamine(poly(ethyleneoxazoline)-8988) and octadecyl(poly(ethyleneoxazoline)-5822) with 85 and 56 monomer units, respectively, as lipopolymers in their bilayers (38). In this case, the lipopolymers were photo-cross-linked to the glass via a benzophenone silane photocoupling agent. The tethered polymer obstructs lipid diffusion at high concentrations but not at low concentrations (39). The increased stability of the monolayer that faces the substrate allows the preparation of asymmetric supported bilayers that contain separated regions of liquid-ordered and liquid-disordered phases. Lipid raft mixtures in the substrate-proximal layer induce registered raft domains in the distal layer (40, 41). Sackmann’s group used lipopolymers that contain polymerized 2-methyl-2-oxazoline of different lengths to integrate the large transmembrane cell receptor integrin αβ (42). In all, 20% of the integrins were mobile with a relatively low diffusion coefficient of 0.03 μm²/s. A different approach to form a cushion between the solid support and membrane uses the Langmuir-Blodgett technique to transfer several layers of triglycylsilyl cellulose onto a glass substrate. When supported bilayers were formed on this substrate by direct vesicle fusion, 25% of the reconstituted integrins were mobile with a diffusion coefficient of 0.6 μm²/s (43).

**Studies of raft-like lipid domains in supported bilayers**

Substantial literature has been published on studies of lipid domains in supported bilayers. Many have investigated lipid mixtures with coexisting gel and liquid-crystalline phases by atomic force (AFM), spinfluorescence, and near-field fluorescence microscopy (NSOM) (see, for example, References (44–45)).
Other studies have focused on (“raft”) lipid mixtures that form coexisting I_d and I_u phase domains in the presence of cholesterol (19,48–50). AFM studies are usually carried out on mica substrates, which have the advantage of being atomically flat but are less hydrated than glass or quartz substrates that are commonly used in fluorescence microscopic studies. Although higher resolution is achieved by AFM than by optical microscopy, the bilayers are more tightly coupled to the substrate and are not always fluid on mica. Fluorescence has the advantage that different molecular fluorescent dyes can be used to probe different aspects of complex lipid bilayers, such as diffusion, order, and phase partitioning, but it has the obvious disadvantage that the behavior of the probe is observed, which may or may not reflect the behavior of the host lipids accurately. Judicious choices must be made when selecting lipid probes for different purposes.

Typical “raft” lipid mixtures contain about equimolar sphingomyelin, phosphatidylycholine, and cholesterol. Studies have shown that the area of I_d phases in such bilayers depends linearly on the cholesterol concentration and that a percolation threshold exists at about 25 to 30 mol % cholesterol, where I_d domains become connected and I_u domains become disconnected (19). Because this mixture is at about physiological concentrations of cholesterol, cells may use shifts in cholesterol concentration as a switch to connect different groups of membrane proteins and thereby regulate their function.

An exciting new development in supported bilayer technology is that bilayers with asymmetric lipid distributions can be prepared. Although it is not so difficult for gel phase lipids with very low diffusion coefficients, it is much more challenging with fluid phase bilayers. However, these challenges have been solved recently (18). It is now possible to study phase coupling across the mid-plane of bilayers with coexisting I_d and I_u phases (40, 41). The observation that I_u phases can be induced in lipid mixtures that mimic the inner leaflet of cell membranes that do not exhibit such phases on their own has important potential consequences on signal transduction in cell membranes. Although they do not prove the raft-hypothesis of signal transduction in cells, these experiments provide the first experimental evidence that fluid lipid bilayers with physiologically relevant lipid compositions can transmit signals across the mid-plane of membranes by inducing new lipid phases on the other side.

Protein–lipid interactions in supported bilayers

Protein–lipid interactions and particularly peptide–lipid interactions have been studied in supported bilayers by attenuated total reflection (ATR)/FTIR spectroscopy. A slightly dated, but still valid comprehensive review on this method applied to supported bilayers has been published (20). Because IR light probes the vibrational properties of different classes of covalent bonds, this method is useful to examine lipids, peptides, and interactions between the two in the same sample. The most common parameter for assessing lipid structure and order is to study the stretching vibrations of the lipid acyl chains, for example as a function of peptide concentration or temperature. Such studies have lead to the conclusion that fusion peptides from viruses increase the lipid chain order of fluid phase bilayers and that this property correlates with the biological activity of the viral peptide sequences (51). Another prominent band in the FTIR spectra of lipids develops from the ester carbonyl vibration. This band is sensitive to the hydrogen-bonded structure and thus hydration properties of the bilayer interface. The carbonyl ester band has been shown to change during interaction with some fusion peptides (52). Lipids also have an effect on peptide structure. For example, influenza and HIV fusion peptides are induced to form α-helices in bilayers at low concentration, but they transition into β-sheets at higher concentration (53, 54).

Structural transitions are readily followed by monitoring the amide I and II IR bands when the system is changed from an H2O to a D2O buffer (56). Binding of proteins to supported bilayers may be studied quantitatively by total internal reflection fluorescence microscopy (TIRFM), as shown, for example, in binding studies of antibodies to lipid haptens (57, 58). Fast kinetics of antibody fragment binding and unbinding to supported bilayers have been studied in detail by combining fluorescence correlation spectroscopy with TIRFM (59).

A different aspect of lipid–protein interaction that is conveniently studied in supported bilayers is the lateral diffusion of proteins and lipids and their influence on each other. The regulatory lipid phosphatidylinositol-biphosphate (PIP2) slows the diffusion of syntaxin in supported bilayers (37). Conversely, increasing syntaxin concentrations decrease the diffusion of PIP2 to a lesser extent that of phosphatidylserine. In another system, in which the transmembrane domain of the fibroblast growth factor receptor was incorporated into supported bilayers, lipid and protein diffusion were measured (60). Although protein diffusion was slow (0.006 μm2/s), lipid diffusion was fast (2.6 μm2/s).

Protein–protein interactions in supported bilayers

Much current interest exists in lateral interactions between two or more integral membrane proteins. Several membrane-bound receptors are activated by such interactions during ligand binding. In addition, the interaction between individual transmembrane helices is key to membrane protein folding and has thus received a lot of attention. Although these kinds of interactions are more often studied in nonsupported model membranes, the Hristova group has recently used supported bilayer formats to examine helix–helix interactions between the transmembrane domains of the human fibroblast growth factor receptor (21). Fluorescence resonance energy transfer measurements of fluorescein- and rhodamine-labeled peptides in supported bilayers show that these transmembrane domains dimerize with a sequence-specific dimerization energy of −3.6 kcal/mol. Meaningful measurements of such interaction energies require that the proteins are laterally mobile in the supported membranes, which was verified in this system (60).
Membrane fusion with supported bilayers

Membrane fusion is a key process in cell biology that takes center stage in membrane biogenesis, fertilization, virus entry, and other events. Our group began studying membrane fusion in supported bilayers 14 years ago. Influenza virus hemagglutinin (HA) (i.e., the protein that mediates virus entry into cells by membrane fusion) was reconstituted into supported bilayers. The physiological pH-dependent fusion of liposomes to the plasma membrane of neurites has been studied in supported bilayers. Several fusion-related structural transformations of HA were recorded by FTIR spectroscopy (61, 62). The fusion of single virions to supported bilayers has also been reported recently (63).

Recently, SNARE-mediated fusion (i.e., the process that leads to neurotransmitter release in synaptic transmission) has been reconstituted in a supported bilayer format in four different laboratories including our own. The general design of these experiments is illustrated in Fig. 2. Fix et al. (24) observed that 15% of the docked vesicles fused within 15 seconds, which yielded a fusion rate of ~7 x 10^{-5} s^{-1} in POPC bilayers at a cocrystallized syntaxin1/SNAP25 protein/lipid concentration of 1:300. The fusion probability increased 40-fold after Ca^{2+} addition. Bowen et al. (64) observed thermally induced fusion with a rate of 0.07 s^{-1} within 120 sec after triggering. Liu et al. (65) showed that 77% of the reconstituted synaptobrevin vesicles fused within 100 ms after docking rate, 40 s^{-1}) to syntaxin or syntaxin/ SNAP25 complexes in POPC/DOPS bilayers at a protein/lipid ratio of 1:30,000. The results of these three groups are different from each other, and each of these experiments had peculiar aspects which may be caused by the immobile reconstitution of the SNARE proteins in the supported bilayers (66). The independence of fusion on the presence of SNAP25 in the reconstitutions of Bowen and Liu is puzzling because it contradicts in vivo as well as in vitro fusion results. Preliminary work from our laboratory is in much better agreement with the biological literature on this process.

Cell adhesion and signaling in supported bilayers

Supported bilayers have been used since their inception as surrogate cell surfaces to stimulate immune cells in artificially created immunological synapses. For example, Brian and McConnell (67) have reconstituted major histocompatibility complex (MHC) proteins into supported bilayers and used these planar membranes to stimulate cytotoxic T-cells via specific T-cell receptor interaction. A nitorgen presentation and signaling through the T-cell receptor have been studied in this system in several follow-up papers as summarized by Watts and McConnell (68). The approach of these early studies (and several others that followed) has experienced a recent renaissance because of improved imaging technologies and because we have learned in the interim how to manipulate supported membranes to form specific controllable spatial patterns (see the next section). These newer approaches are exciting because they have allowed investigators to manipulate spatial patterns in the immunological synapse and thereby ask biologically important new questions with regard to the mechanism of cell-induced cell signaling. Mossman et al. (30) have used laterally constrained bilayers to induce novel patterns of T-cell receptor and intracellular adhesion molecule (ICAM) distribution in MHC-stimulated cells. The authors used a GPI-linked version of the MHC in the supported bilayers to circumvent problems of lateral mobility of the native integral membrane protein in un-cushioned bilayers. Wu et al. (69) demonstrated the compartmentalization of IgE-receptors in rat basophilic leukemia cells when they stimulated these cells with lipid-hapten bound IgE in nano-fabricated bilayer patches.

A dhesion of different immune cells to one another or to epithelial cells has also been studied using planar bilayer models. For example, lymphocyte function-associated protein-1 (LFA-1) promotes cell adhesion in inflammation (i.e., a reaction that can be mimicked by binding to purified ICAM-1 in supported membranes (70)). Similarly, purified LFA-3 reconstituted into supported bilayers mediates efficient CD2-dependent adhesion and differentiation of lymphoblasts (71). This work was followed by a study in which transmembrane domain-anchored and GPI-anchored isoforms of LFA-3 were compared (72). Because this research occurred before the introduction of polymer cushions and because the bilayers were formed by the simple vesicle fusion technique, the transmembrane domain isoform was immobile, whereas the GPI isoform was partially mobile. By comparing results with these two isoforms at different protein densities in the supported bilayer, the authors showed that diffusible proteins at a sufficient minimal density in the supported membrane were required to form strong cell adhesion contacts in this system.

In another study, epithelial cells were bound via their integrin receptors to supported bilayers that presented the
Supported Lipid Bilayers: Development and Applications in Chemical Biology

RGD-peptide, which is the classic integrin ligand (73). The cells spread on RGD presenting membranes but not on control membranes that lacked this peptide. In a similar approach, a laminin-derived peptide was presented on supported bilayers and shown to mediate the spreading and partial differentiation of neuronal subventricular zone progenitor cells (74). The authors observed a strong nonlinear relationship between surface concentrations of the peptides and conclude that this approach may provide novel conditions for growing stem cells with only a limited and controlled amount of differentiation induction.

### Patterning supported bilayers

To use supported bilayers as platforms for screening assays, different approaches have been taken to pattern the membrane on the surface. One goal is to observe the membrane interaction of substances simultaneously with supported bilayers of different composition. Bilayers can be subdivided into different areas by diffusion barriers erected on the substrate or by blotting patches of membrane with polydimethylsiloxane (PDMS) stamps (75). A combination of these techniques may be used to form patterned membranes of different composition. Barriers may be formed from metal or metal oxides by standard lithographic techniques or by stamping proteins onto the substrate (76). Originally, homogeneous supported membranes have been patterned photolithographically using deep-UV light (77). Using photolithography in combination with a polymer lift-off technique, patterned lipid bilayers with patches measuring 1 to 76 µm may be formed (78). A rather approach uses the flow of vesicle suspensions within microfluidic devices to deposit membranes of different compositions at the same time (75). Supported bilayers of different compositions on the same substrate are conveniently addressed with different solutions of analytes flowed through micro-channels in PDMS (79).

Using supported bilayers as platforms for chemical sensing and analysis

Because of their planar geometry, supported bilayers are predestined for biological and chemical sensor applications. The basic concept is to couple the high specificity and sensitivity of molecular membrane receptors to substrates that integrate optical-electronic circuits (25–27). In combination with the described patterning techniques, supported membranes should provide a nearly ideal physiological environment for high-throughput sensing with biological membrane receptors or channels in chip-based arrays. However, despite their great promise, economically viable commercial applications have so far not yet been used because of a series of obstacles that still must be overcome. Nanoscopic and microscopic defects in fluid supported lipid bilayers most often lead to low electrical resistance and therefore make them unsuitable for high-sensitivity electronic detection. In some approaches, this problem has been overcome by linking the lipid bilayer with very short and dense tethers to the substrate. However, doing so prevents the incorporation of larger membrane receptors and channels.

Another approach is to prepare black lipid membranes with solvent over microscopic or nanoscopic holes in the substrate (80-82), or by using giant vesicles that adhere to the substrate (83). Although high-resistance seals have been obtained, long-term stability has not yet been achieved with these systems. In cases where highly insulating membranes are established, the conductivity of ligand-gated ion channels or larger pores may be recorded (25, 84). Alternatively, the conductivity may be probed by measuring the impedance (26) or with metal-free field-effect transistors that have the advantage of avoiding electrochemical perturbations (83). Optical detection is suitable for applications such as immunoassays when the molecules of interest can be addressed by fluorescent- or gold-labeled antibodies (79). As is true for electronic biosensors, optical biosensors designed for routine practical applications must be robust and stable for a long time. Lippolymers that stabilize supported bilayers against air exposure are also helpful in this case (85).

### Procedures to Prepare Supported Bilayers

Supported lipid bilayers with or without reconstituted membrane proteins are prepared by one of three methods described below and illustrated in Fig. 3. Methods details can be found in Reference 86.

**Langmuir-Blodgett/Langmuir-Schafer (LB/LS) technique (4, 5)**

This technique is historically the first method to prepare supported bilayers. A lipid monolayer is spread from a desired lipid solution in organic solvent onto a water surface in a Langmuir trough. After evaporation of the solvent, the monolayer is compressed slowly to reach a surface pressure of 32 mN/m (thought to be the equivalence pressure of a bilayer) and equilibrated. A carefully cleaned hydrophilic substrate (glass, quartz, oxidized silicon, etc.) is then rapidly submerged into the trough and slowly withdrawn with a dipper mechanism while a constant surface pressure is maintained. This step transfers a single monolayer of lipids known as the LB layer onto the substrate. A second monolayer known as the LS layer is then spread and compressed on the trough in the same fashion. The LB-coated substrate is attached to a suctioning tip, and its face is gently lowered to contact the LS monolayer at the air/water interface for a few seconds. To complete the bilayer, the slide is then pushed through the interface and placed on a cover slip fitted with two spacers of water-resistant double-sided tape that had been previously placed at the bottom of the trough. After removal of the supported bilayer sandwich from the trough, the water between the surfaces may be exchanged by flow-through with any desired buffer while always maintaining full hydration of the bilayer. Some investigators have prepared peptide-containing supported bilayers by placing peptides in the first or second monolayer.

**Vesicle fusion (VF) technique (67)**

This technique is the simplest method for forming supported bilayers. Much literature is available on the mechanism and
Figure 3 Methods for supported bilayer formation and membrane protein reconstitution. (a) and (b) LB/LS method. A lipid monolayer is spread at the air-water interface of a Langmuir trough and transferred to a solid substrate while keeping the surface pressure constant. A second monolayer is transferred by horizontal apposition of the first transferred monolayer and collection of a counter-piece with spacers. (c) Direct VF method. Membrane vesicles are flown into a chamber with a clean surface substrate on top. After about an hour of incubation, they form a supported bilayer on the substrate and excess vesicles are flushed out. (d) LB/VF method. The procedures depicted in panels (a) and (c) are combined leading to an asymmetric bilayer with an asymmetric protein distribution. Although this method can also be performed without a polymer, it is shown here with the polymer transferred during the LB step.

kinetics of vesicle spreading on hydrophilic substrates, which is not reviewed here. Although it is simple, the method tends to result in bilayers with more defects, and the orientation of membrane proteins cannot be controlled in this method. Small or large unilamellar pure lipid vesicles or proteoliposomes are prepared by standard liposome preparation or membrane protein reconstitution methods. A clean hydrophilic substrate is placed in a flow-through chamber, and the vesicles are injected and incubated with the surface for 30–60 minutes. Excess vesicles are washed out by extensive rinsing with a buffer. This method has also been used to make polymer-supported bilayers, in which case the surface of the support is pre-treated with the polymer using conditions that depend on the particular polymer being used.

Langmuir-Blodgett/vesicle fusion (LB/VF) technique (87)

This method is a combination of the other two methods. In our opinion, it is the most gentle method to reconstitute membrane proteins into supported bilayers and to prepare supported bilayers with fragile coexisting liquid phases of lipids. A LB monolayer is prepared on a hydrophilic substrate as described above. To prepare tethered-polymer supported bilayers, a suitable lipopolymer may be included at a concentration of a few mol % at this stage. With some lipopolymers, it is necessary to cure the slide (by light, temperature, zero humidity, etc.) at this stage. The monolayer-coated slide is then placed in a custom-built flow-through chamber, and vesicles or proteoliposomes are injected and incubated with the surface for 30–60 minutes (pure lipid vesicles) or 60–120 minutes (proteoliposomes). Excess vesicles are washed out by extensive rinsing with buffer. Because the second monolayer is completely vesicle derived and because membrane proteins are introduced as proteoliposomes only in the second step, they tend to be unidirectionally oriented in the supported bilayer as can be verified with quenching antifluorophore antibodies (23, 37) or by FLC microscopy (33).

Procedures to Characterize Supported Bilayers

Microscopy

The simplest way to examine the quality and integrity of supported membranes is to include a fluorescent lipid probe such as n’-tetrabenzoxa-diazol (NBD)-PE or rhodamine-PE and to look for
their uniform appearance on a standard epifluorescence microscope. Many artifacts can be detected readily and eliminated with this very simple test. Higher-resolution images can be obtained by AFM or NSOM. These techniques are useful to detect small defects in the 10 to 500 nm range that might escape detection by standard wide-field optical microscopy. Because supported bilayers are only stable under water, these imaging modalities must be carried out under water. AFM probes height profiles, but NSOM and epifluorescence microscopy permit the labeling of specific chemical structures or physical properties of the structures by using different fluorescent probes. It is often necessary to discriminate surface from bulk fluorescence, for example when measuring the binding of fluorescent ligands to supported membranes. This result is conveniently achieved with TIRFM, which has a typical 1/e illumination depth of 50 to 100 nm from the surface, which depends on refractive indices, angle of incidence, and wavelength of light. Brewster angle and surface plasmon microscopies are also surface-selective optical imaging techniques, which do not require the fluorescent labeling of the membrane, but rather depend on lateral changes of refractive index in the sample.

Lateral diffusion

A next and very important level of characterization of supported bilayers is the measurement of the lateral mobility of their constituents. Because biological membranes and their proper function are defined by their fluidity, to recreate this characteristic is imperative for biology-motivated work. In addition, lateral diffusion measurements on supported bilayers can easily detect many artifacts that may go unnoticed by simple microscopic inspection. For example, deposited membranes may look completely uniform but may not show any long-range lateral diffusion when vesicles or membrane fragments that are smaller than the resolution of the light microscope are densely packed on the substrate surface. Two techniques are common to determine the diffusion of fluorescently labeled lipids or proteins in supported bilayers: FRAP and SPT. The lateral diffusion coefficient and fraction of mobile molecules are obtained from either measurement.

In FRAP, a brief pulse of intense laser light is used to photobleach fluorophores partially in a small area of the sample. The recovery of fluorescence caused by diffusion of labeled molecules into the bleached area is then observed over time, while care is taken to minimize additional photobleaching. In spot photobleaching the light is focused to a circular spot, which reflects the Gaussian beam profile of the laser. During recovery, the half-width of the bleached area decreases whereas the intensity increases. The diffusion coefficient and mobile fraction are extracted from the time course and the amplitude, respectively, of the recorded recovery curve (88). In a variant called periodic pattern photobleaching, the bleach pulse projects a stripe pattern of a Ronchi ruling onto the sample, which permits integration over a larger area and therefore an increased signal/noise (89).

SPT is the preferred technique when more detail on different populations of moving particles in a heterogeneous system is required. Although the information content of SPT is much higher than that of FRAP, it is much more demanding on instrumentation and statistical evaluation procedures. SPT has been introduced to the characterization of supported bilayers in 1995 (90) and has been used frequently since then. In practice, the technique is best used in combination with TIRF microscopy and high-sensitivity charge-coupled devices. Typical labeling ratios of the lipid bilayer are 1:100 fluorescent probes:lipids. The reconstructed particle trajectories and appropriate statistics can be analyzed to distinguish between diffusion, anomalous diffusion, confined diffusion, and directed motions.

Bilayer structure

Neutron reflectivity has been used to characterize the transverse organization of supported bilayers structurally (32, 91). The lateral structure of lipid bilayers on solid supports may also be characterized by grazing incidence X-ray diffraction, although this technique has so far been mainly used on monolayers at the air-water interface. Vibrational spectroscopies open interesting windows to look at details of lipid structure in supported bilayers. FTIR spectroscopy has several bands that are characteristic of the state of lipid order and hydration in supported bilayers (20). An interesting relatively recent method to study the organization of supported bilayers, particularly with respect to their asymmetry, is sum frequency vibrational spectroscopy (SFVS) (92). Signals develop in this nonlinear form of vibrational spectroscopy only when symmetry is broken (i.e., when the type, structure, and number of lipids are unequal across the mid-plane of the supported bilayer). A complementary technique to study lipid asymmetry, or more specifically fluorescent probe asymmetry, is FLIC microscopy, in which the average distance of fluorophores from an oxidized silicon mirror surface is measured interferometrically (18).

Protein secondary structure and orientation

Two methods have been used to determine the secondary structure of supported bilayers: polarized ATR-FTIR spectroscopy and oriented CD spectroscopy. SFVS may also be applied to study peptide and protein structures in supported bilayers. Polarized ATR-FTIR spectroscopy is sensitive enough that high-quality spectra can be obtained from a single bilayer. Beta-sheet structures are readily distinguished from alpha-helical and random structures, and the orientations of alpha-helices are determined from the linear dichroism of the peptide amide I bands (20). Multiple stacks of supported bilayers have to be used to gain enough sensitivity to determine the structure and orientation of alpha-helices in lipid bilayers by oriented CD spectroscopy (60, 93).

Conductance and impedance

Pure lipid membranes are electrical insulators with a specific capacitance of ~1 µF/cm², which separate two electrolytic compartments. The conductance of biological membranes is mainly determined by highly specialized proteins that act as ion channels. For supported membranes to mimic the electrical properties of a biological membrane, it is necessary to measure its electrical characteristics. Even very small defects that are not
visible by microscopy increase the conductance significantly. If the membrane is supported by a conductive substrate like gold or indium tin oxide, then it is possible to measure the impedance by applying A.C. voltages with frequencies up to 100 kHz and by measuring the magnitude and phase of the current. The conductance and capacitance of the supported membrane is determined from the evaluation of the entire electrical circuit, which includes resistance and capacitance of the electrode and if necessary alternative electrical paths around the membrane (94). A similar approach uses metal–field effect transistors that probe the electrical potential at the transistor gates that face the cleft beneath the supported membrane. The electrical membrane parameters are extracted from the voltage transfer from the electrolyte bath to the transistor gates (83).

Conclusion

In conclusion, supported bilayers have evolved into a reliable model membrane system since their first inception almost a quarter century ago. Numerous basic research questions regarding the structure and function of biological membranes and applications that range from biosensing to proteomic analyses have been addressed with this system. We anticipate more growth and an even more prominent role of this tool in basic and applied membrane research in the decades to come.

Acknowledgments

The work in our laboratory was supported by grants from the National Institutes of Health.

References


See Also
Membranes, Fluidity of
Membrane Proteins, Properties of
Membrane Fusion, Mechanisms of
Lipid Bilayers, Properties of
Lipid Rafts
Biosensors
Bioorganic Interfaces
Glycan arrays are powerful tools for high throughput analysis of carbohydrate–macromolecule interactions. A glycan array is composed of many different carbohydrate structures immobilized on a solid support in an orderly arrangement. This review describes the challenges and considerations in the development of a glycan array. Various array formats, immobilization techniques, and assay systems are discussed. In addition, several interesting applications of glycan array technology are described, such as assessing the specificities of carbohydrate-binding lectins and antibodies, profiling antiglycan antibodies in serum as biomarkers of disease, and evaluating carbohydrate-dependent cell binding.

DNA and protein microarray technology has revolutionized the way scientists study complex biological processes. These arrays consist of thousands of nucleic acids or proteins immobilized on a solid support. The array format allows one to rapidly evaluate interactions with a large number of molecules simultaneously. For example, one can examine the expression profiles of thousands of genes in a single experiment.

Glycan arrays are an equally powerful technology for the evaluation of carbohydrate–macromolecule interactions. Analogous to DNA and protein arrays, glycan arrays contain many different carbohydrates affixed to a solid support. This review will focus on development strategies, challenges, and applications of glycan arrays. Several other reviews have been published over the last few years (1–4).

Biological Background

Carbohydrates, biopolymers composed of monosaccharide units, play a central role in a wide range of biological processes such as protein folding, inflammation, and development. In addition, glycans undergo dramatic changes in expression during the onset and progression of many diseases such as rheumatoid arthritis and cancer. Unfortunately, progress toward defining the specific roles of most carbohydrates and understanding the relationships between structure and function has been frustratingly slow. Furthermore, efforts to exploit altered expression for therapeutic benefit have only been successful in a limited number of cases.

Molecular recognition is a fundamental element for both basic and applied carbohydrate research (see cross references: Glycan–glycan interactions, Glycan–protein interactions, and Sugar–lectin interactions in cell adhesion). Many important biological processes involve specific interactions between a carbohydrate-binding protein (lectin) and a glycan. For example, one early stage of inflammation involves interactions of selectins with carbohydrate ligands. Carbohydrate–protein interactions are also directly involved in diseases. For example, many pathogens such as the influenza virus, E. histolytica, and H. pylori bind carbohydrates on the surface of host cells as a key step of infection (5). For many other carbohydrate-binding proteins, however, the biological functions are still unknown. For example, most lectins used routinely as research reagents are isolated from plants. Although used in the laboratory for many years, the biological roles of many of these proteins are not well understood. As a result, there has been significant interest in identifying natural and unnatural ligands that can be used to modulate the activity of carbohydrate-binding proteins. Lectins and glycan-binding antibodies are also used extensively as research tools, diagnostics, and therapeutic agents. Information on the specificities of these proteins is critical for interpreting results and selecting the best clinical candidates.

Although analysis of carbohydrate–macromolecule interactions is crucial for glycochemistry, it remains a challenging area of science. First, carbohydrates can be exceedingly difficult to obtain, especially in homogeneous form. With limited access to the ligands, one cannot easily assess recognition. Second, traditional methods used to evaluate carbohydrate–protein interactions such as monosaccharide and oligosaccharide inhibition studies, isothermal calorimetry (ITC), surface plasmon resonance (SPR), and enzyme-linked lectin assays (ELLAS) can be labor intensive or require large quantities of each carbohydrate.
A as a result, these methods are not well suited for high throughput evaluations. Finally, one must consider the issue of valency. In most cases, interactions between a single carbohydrate ligand and a single binding domain of a protein (referred to as monovalent binding) are very weak. The affinity of monovalent binding events is typically too low to withstand the washing involved in common biochemical assays such as enzyme-linked immunosorbent assays (ELISAs), Western blots, and immunohistochemical staining. However, most carbohydrate-binding proteins possess two or more binding sites or assemble into functional units with multiple binding sites. As a result, they can simultaneously bind two or more carbohydrate ligands (referred to as polyanalytic binding) leading to a high overall affinity or “avidity.” A key feature that probe interactions between carbohydrates and proteins must account for the unique aspects of polyanalytic recognition. For example, the ligands should be presented in a polyanalytic context. In addition, the spacing and orientation of ligands will affect the ability to form a polyanalytic complex.

As a result of the difficulties mentioned above, researchers have been developing alternative methods to study carbohydrate-protein interactions. Array technology is specifically designed for high throughput evaluations of molecular recognition events. Miniaturization facilitates the process by allowing minimal use of reagents and other hard-to-obtain components. In the following sections, the development and application of glycan array technology are described.

Fabrication of Glycan Arrays

Fabrication of a glycan array involves several interrelated factors: choosing a solid support format, choosing a method to attach glycans to that solid support, and obtaining appropriately functionalized carbohydrates for immobilization.

Array formats

A recent trend of modern glycan arrays can be traced back to the method of separating a complex glycolipid mixture using thin-layer chromatography, then probing the “array” with a carbohydrate-binding protein (6). Recently, most glycan arrays are created on either microtiter plates or glass microscope slides (see Fig. 1).

The microtiter plate array format involves immobilizing carbohydrates in wells of 96-, 384-, or 1536-well microtiter plates (see Fig. 3a). Each carbohydrate component is spatially separated from other components within the plate. Two of the primary advantages of a microtiter plate format are cost and simplicity. Carbohydrates can be distributed into wells using multichannel pipetters, and assay results can be measured using standard plate readers. Thus, the equipment and supplies needed for the array are relatively inexpensive and common. However, microtiter plates generally require larger amounts of each carbohydrate and can accommodate a smaller total number of components per support unit.

An alternative to the microtiter plate format is a glass microscope slide. Slides can come in various layouts. The array developed by Glycominds, Ltd. contains 200 microwell’s with a single carbohydrate in each well (7). The miniaturized wells use a smaller amount of material than a microtiter plate but retain spatial separation of components (see Fig. 3b). A second format involves spotting components directly onto the slides. One of the key differences is that all the carbohydrates are in the same “well” (see Fig. 3c). Consequently, recognition of each component is compared under identical assay conditions and well-to-well variation is minimized. In addition, microarray printers used for producing DNA microarrays can be used to print very small features (50–200-µm spots) with high precision allowing for tens of thousands of spots on each slide.

As a result, much smaller quantities of material are required and the total capacity is considerably higher than a microtiter plate. However, a microarray printer and high resolution scanner are required. A glass slide can also be modified to create 2–16 macrowells on a glass slide (see Fig. 3d). In this format, an entire array is printed in each well.

Several three-dimensional (3-D) approaches to glycan arrays have also been published, including fiber optic bead-based glycan arrays (see Fig. 3e) (5) and hydrogel arrays (9).

Methods for immobilizing carbohydrates

The methods of attachment to the solid support can be divided into two broad categories: covalent and noncovalent. Several factors should be considered. First, the ligands should be stable to immobilization, storage, and assay conditions. Second, the immobilized ligands should be accessible to the receptors, which are usually in solution. Third, the method should be efficient.

One of the most straightforward methods for immobilizing ligands is noncovalent attachment (4, 9–18). This method involves adsorption onto surfaces using noncovalent forces such as hydrophobic interactions, charge-charge interactions, and charge-dipole interactions (see Fig. 2a). A key feature is that the immobilized ligands must withstand routine screening and assay conditions. In general, lower molecular weight hydrophilic molecules, such as monosaccharides and oligosaccharides, show poor retention on solid supports when directly attached by noncovalent methods. In contrast, polysaccharides, glycoproteins, and glycolipids are well retained and can be directly arrayed on modified solid supports.

In one of the first reports of a carbohydrate microarray, underivatized polysaccharides and glycoproteins were immobilized directly on nitrocellulose-coated glass slides (10). In a further demonstration of the feasibility of noncovalent immobilization of underivatized glycans, Willaet et al. reported microarrays of glycoproteins, proteoglycans, and polysaccharides on oxidized black polystyrene slides (11). Modified neutral and anionic dextran polysaccharides have also been synthesized and printed on semicarbazide-coated glass slides (12).

One strategy for noncovalent attachment of smaller monosaccharides and oligosaccharides involves coupling them to a lipid tail or other hydrophobic moiety, which provides suitable retention on a solid support. This strategy was exemplified by arrays of synthetic and natural oligosaccharides as neoglycolipids or glycolipidoligosaccharide derivatives affixed to nitrocellulose-coated slides (13) and microtiter plates (14). Alternative approaches include covalent attachment to solid supports using a
Array-Based Techniques for Glycans

Figure 1 Examples of formats that have been used for glycan arrays.

- (a) Microtiter plates: 1 Sugar/Well
- (b) Micro-Well Glass Slide: 1 Sugar/Well
- (c) Glass Slide: Complete Array/Slide
- (d) Glass Slide with Macro- Wells: Complete Array/Well
- (e) Fiber Optic Bundle: 1 sugar/bead

Cu(I)-catalyzed 1,3-dipolar cycloaddition (15); reductive amiation reactions between underivatized oligosaccharides and immobilized aminolipids on microtiter plates (16); and urea formation between amino-derivatized sugars and tetradecyl isocyanate adsorbed on microtiter plates (17). Formation of hydrogel microarrays of amino-saccharides and polyacrylamide glycoconjugates has been reported. Here, glycans were printed on hydrophobic glass surfaces with polymerizable monomers, and then polymerization was induced by photoactivation to form 3-D gel elements comprising the glycans (4, 9).

An elegant alternative approach for noncovalent interaction relies on fluorous-fluorous interactions. A glycan array of monosaccharides and disaccharides bearing anomeric fluorous tags was noncovalently immobilized on fluorous-derivatized glass slides (19, 20). The attachment method is compatible with a wide range of functional groups and has been successfully used to probe carbohydrate-protein interactions.

Glycan arrays have also been developed to take advantage of the noncovalent, but extraordinarily strong, biotin-streptavidin interaction. The Consortium for Functional Glycomics developed a glycan array consisting of biotinylated oligosaccharides and amino acid glycoconjugates immobilized on streptavidin-coated 384-well microtiter plates (21, 22). Another biotin-streptavidin “sandwich” glycan microarray was produced by self-assembly of biotinylated alkylithiols on a gold substrate, followed by coating with streptavidin and printing of biotinylated polyacrylamide glycoconjugates (23).

An alternative to noncovalent approaches is covalent attachment of ligands to the surface. Methods for covalent attachment can be broadly categorized as follows: 1) condensation with the aldehyde of a reducing sugar, 2) nucelophilic addition to or displacement of a group on an activated surface or sugar, 3) cycloaddition, and 4) insertion reactions (see Fig. 2b). Some derivatization methods, for example, condensation and reductive amination, lead to ring opening of the reducing end sugar, whereas other methods result in indeterminate anomeric configurations, depending on the reaction conditions and point of attachment. The linker used to connect the sugar to the surface is an important consideration because it can affect recognition and accessibility of the ligands. In choosing an ideal linker, the issues to consider include the chemical nature and composition (hydrophobic vs. hydrophilic), stability, length, and flexibility of the linker.

Unmodified sugars have been attached to various hydrazide- and amino-derivatized slides by condensation or by reductive amination, which proceeds by reduction of the intermediate Schiff base (see Fig. 2b.1). This strategy was used for immobilization of free oligosaccharides onto hydrazide-derivatized plates (24) and various unmodified carbohydrates (monosaccharides, disaccharides, and polysaccharides) on hydrazide-coated glass slides (25). Later, immobilization of free oligosaccharides on hydrazide-derivatized self-assembled monolayer (SAM) of gold-coated glass slides was also reported (26). The formation of oximes between reducing end aldehydes of sugars and amino-oxy groups has been explored in the fabrication of glycans arrays. This method was explored for the immobilization of monosaccharides, disaccharides, and oligosaccharides on amino-oxy glass slides and successfully used for making a functional glycan array of oligosaccharides on amino-oxy-functionalized glass slides (27). However, stronger signals were observed for hydrazide-coated slides in comparison with amino-oxy-coated slides (25). Several glycans...
### Figure 2: Immobilization strategies. (a) A schematic representation of noncovalent adsorption. The lipid tail of a neo-glycolipid adheres to a hydrophobic surface via noncovalent interactions. (b) Strategies used for covalent immobilization of glycans.

Arrays have also been reported where the simple condensation of aldehydes and amines was used to covalently attach the saccharides. Deaminated heparin oligosaccharides bearing aldehyde groups were attached to amine-coated glass slides as Schiff bases without further reduction (28), while aldehyde and amino functionalized monosaccharides and disaccharides were immobilized on amine- and aldehyde-coated slides, respectively (29, 30). It should be noted that reductive amination results in opening of the reducing end sugar whereas condensation can lead to changes in anomeric configuration or mixtures of anomers at the reducing end.

An alternative strategy used in the covalent attachment of glycans to solid supports is nucleophilic addition or displacement. In this method, the decision about where to install the nucleophile and the electrophile—on the sugar or on the solid support—rests with the researcher, but could be influenced by the ease of installation and availability of derivatized reaction partners. Glycan arrays involving the Michael addition of maleimide-linked carbohydrates and thiol-coated glass slides (see Fig. 2b.2) have been created and used to probe lectin–carbohydrate interactions (31, 32). The transposition of the reacting functional groups has also been reported where glycans arrays were fabricated using thiol-linked sugars on maleimide-functionalized glass slides (33, 34) and on self-assembled monolayers presenting maleimide groups (35).

Epoxide-opening reactions have also been used to covalently attach carbohydrates, glycoproteins, and neoglycoconjugates to glass slides (see Fig. 2b.3). One report demonstrated the feasibility of immobilizing hydrazide-functionalized small molecules, including monosaccharides, on epoxide-coated slides and further showed that hydrazide-containing ligands react more rapidly with the epoxide-coated surface than thiols and amines (36). The use of epoxide-coated slides is relatively common for the fabrication of protein microarrays. One advantage of using epoxide-coated surfaces is the potential indiscriminate reactivity of epoxides to nucleophiles such as amines, thiols, and alcohols. However, it is important to note that the reactivity of these nucleophiles often depend on the pH conditions during printing. The development of glycan arrays of neoglycoconjugates and glycoproteins on epoxide-coated slides has been described (37, 38). In this report, neoglycoconjugates were synthesized by conjugating to lysine residues on serum albumin through activated carboxyl-bearing linkers, and then printed directly onto epoxide-functionalized glass slides.

The facile reaction of amines with N-hydroxysuccinimide (NHS)-activated carboxyl groups, and the concomitant formation of a stable amide bond, has been exploited in the fabrication of glycan arrays (see Fig. 2b.4). This method has been used for covalent attachment of amine-presenting glycans ranging in size from monosaccharides to glycoproteins on NHS-activated slides (39, 40). In one report, oligosaccharides were derivatized with photocleavable amine-bearing linkers used to attach the ligands to a porous silicon surface presenting NHS groups (41). Oligosaccharides labeled with fluorescent 2,6-diaminopyridine have also been printed on NHS-activated glass slides (42). An array of monosaccharides and oligosaccharides was created by the reaction of p-aminophenyl-glycosides with cyanochrome-activated linkers on wells of microtiter plates (7).

In the fabrication of bead-based glycan arrays, glycopeptides and bovine serum albumin neoglycoproteins were conjugated to the carboxyl-presenting beads using water-soluble carbodiimides. The beads were spatially arrayed into microwells for
screening with lectins (8) or assayed for lectin binding as a suspension (43).

Covalent immobilization of glycans by cycloaddition (see Fig. 2b.5 and insertion reactions (see Fig. 2b.6 and 2b.7) has been used in the creation of arrays (44–47). Cyclopenta-diene-carbohydrates were arrayed on self-assembled monolayers alkanethiols presenting benzocyclooctene through pentacyclic organometallic links (44). An efficient Diels–Alder reaction between the modified glycan and the support and the reduction of non-specific interactions by the use of the glycol linker are highlights of this report. One attachment method that portends the development of functional group-independent reactions where any underivatized molecule can be immobilized is the use of insertion reactions. In one report, various glycoproteins were printed on polymeric surfaces presenting arylfluoromethyl disorbidines. The surfaces were photoactivated resulting in formation of a covalent bond between the solid support and the printed ligand by carbene C–H insertion (46). In an alternative formal insertion strategy, glycans arrays were created by printing underivatized carbohydrates on SAM’s presenting phthalimide groups, which, on photoactivation, undergo hydrogen atom abstraction and recombination of radicals to form a covalent bond (47).

Obtaining glycans for an array

Preferably, one would like access to a representative set of naturally occurring monosaccharides and oligosaccharides, polysaccharides, glycopeptides, and glycolipids and their synthetic analogs. However, obtaining a diverse set of carbohydrates in a format compatible with the immobilization strategy remains a major challenge for glycan array development. Current arrays contain around 30–300 different components (see Table 2). This level of diversity can provide useful information but represents only a small fraction of the total diversity of carbohydrate structures found in nature. Including glycans from animals, plants, and bacteria, there are over 10,000 different structures that have been identified and characterized to date. Therefore, one of the principal challenges for the future advancement of glycan array technology is developing strategies to obtain much larger sets of carbohydrate ligands.

One method that has been used to acquire carbohydrates is isolation and purification from natural sources such as human or animal tissue, milk, urine, plants, and bacteria (see cross reference: Isolation of glycans). A access to homogenous carbohydrate structures can be challenging due to the difficulties in separation of complex mixtures, identification of carbohydrate(s) contained within each fraction, and preparation of sufficient quantities from the limited amounts present in a particular sample. A fortuitously, mixtures of unknown composition can be used to survey a broad repertoire of the glycome. On identification of a mixture containing one or more members with interesting receptor-binding properties, the mixture can then be deconvoluted by further fractionation and separation by routine analytical techniques (13).

A second method that has been invaluable for accessing glycans is chemical synthesis (see cross references: Glycan synthesis, key reaction of; Glycan synthesis, key strategies for; Glycan synthesis, protection and deprotection steps of; and Glycopeptides and glycoproteins, synthesis of). This approach can provide both natural and unnatural structures that can be very useful for probing relationships between structure and function. Although there have been tremendous advances in our ability to synthesize carbohydrates both in solution and on solid phase, carbohydrate chemistry remains a challenging area of organic chemistry. Progress on the synthetically tractable carbohydrate libraries (48–50), automated carbohydrate synthesis (51), and one-pot multicomponent synthetic strategies (41, 52, 53) will substantially aid glycan array development.

Enzymes can be powerful tools for obtaining carbohydrates as well (54). Chemoenzymatic synthesis of oligosaccharides involves the use of glycosyltransferases or glycosidases for the regio- and stereo-selective formation (typically between unprotected reaction partners) or hydrolysis of glycosidic bonds, respectively. In some cases, one reaction partner is obtained by traditional solution or solid-phase chemical synthesis, followed by enzymatic glycosylation. In other cases, synthesis begins with enzymatic cleavage and continues with more traditional chemical synthesis methods.

Evaluation of Binding

Once a glycan array is constructed, one needs to develop a method to evaluate binding of proteins, cells, viruses, or other macromolecules to the array. The general process involves carrying out the assay, detecting signals on the array, and analyzing processing the results. Although the overall process appears straightforward, assay development can be very challenging. Importantly, every step in the process from fabrication to assay conditions to detection must be successful to produce a signal. As a result, it can be difficult to determine which step or steps need to be optimized, especially when no signal is observed. Therefore, assay development typically requires systematic variation of many parameters. Additional considerations include sensitivity, signal-to-noise ratios, reproducibility, flexibility, and dynamic range.

The vast majority of array assays have used fluorescence as the detection method due to the high sensitivity and the availability of detectors such as fluorescent plate readers for microtiter plates and high resolution DNA microarray scanners for glass microarray slides. Therefore, the following discussion will focus on assay development coupled with fluorescent detection.

Non-specific adsorption

One important general consideration is non-specific binding. Proteins and other macromolecules can adsorb to surfaces leading to high background signals and low signal-to-noise ratios. One common approach to prevent non-specific adsorption is to cover or block any "sticky" surface areas with a noninterfering receptor-binding properties, the mixture can then be deconvoluted by further fractionation and separation by routine analytical techniques (13).

A second method that has been invaluable for accessing glycans is chemical synthesis (see cross references: Glycan synthesis, key reaction of; Glycan synthesis, key strategies for; Glycan synthesis, protection and deprotection steps of; and Glycopeptides and glycoproteins, synthesis of). This approach can provide both natural and unnatural structures that can be very useful for probing relationships between structure and function. Although there have been tremendous advances in our ability to synthesize carbohydrates both in solution and on solid phase, carbohydrate chemistry remains a challenging area of organic chemistry. Progress on the synthetically tractable carbohydrate libraries (48–50), automated carbohydrate synthesis (51), and one-pot multicomponent synthetic strategies (41, 52, 53) will substantially aid glycan array development.

Enzymes can be powerful tools for obtaining carbohydrates as well (54). Chemoenzymatic synthesis of oligosaccharides involves the use of glycosyltransferases or glycosidases for the regio- and stereo-selective formation (typically between unprotected reaction partners) or hydrolysis of glycosidic bonds, respectively. In some cases, one reaction partner is obtained by traditional solution or solid-phase chemical synthesis, followed by enzymatic glycosylation. In other cases, synthesis begins with enzymatic cleavage and continues with more traditional chemical synthesis methods.

Evaluation of Binding

Once a glycan array is constructed, one needs to develop a method to evaluate binding of proteins, cells, viruses, or other macromolecules to the array. The general process involves carrying out the assay, detecting signals on the array, and analyzing processing the results. Although the overall process appears straightforward, assay development can be very challenging. Importantly, every step in the process from fabrication to assay conditions to detection must be successful to produce a signal. As a result, it can be difficult to determine which step or steps need to be optimized, especially when no signal is observed. Therefore, assay development typically requires systematic variation of many parameters. Additional considerations include sensitivity, signal-to-noise ratios, reproducibility, flexibility, and dynamic range.

The vast majority of array assays have used fluorescence as the detection method due to the high sensitivity and the availability of detectors such as fluorescent plate readers for microtiter plates and high resolution DNA microarray scanners for glass microarray slides. Therefore, the following discussion will focus on assay development coupled with fluorescent detection.

Non-specific adsorption

One important general consideration is non-specific binding. Proteins and other macromolecules can adsorb to surfaces leading to high background signals and low signal-to-noise ratios. One common approach to prevent non-specific adsorption is to cover or block any "sticky" surface areas with a noninterfering receptor-binding properties, the mixture can then be deconvoluted by further fractionation and separation by routine analytical techniques (13).
### Array-Based Techniques for Glycans

#### Table 1: Summary of glycan arrays

<table>
<thead>
<tr>
<th>Group or Company</th>
<th>Composition</th>
<th>Format(s)</th>
<th>Number of Components</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortium for Functional Glycomics</td>
<td>Oligosaccharides; neoglycoconjugates</td>
<td>Plates; glass slides</td>
<td>264</td>
<td>[21, 22, 40, 64, 65, 68, 70-72, 74-76, 78-81, 97, 98]</td>
</tr>
<tr>
<td>Feizi, T.</td>
<td>Neoglycolipids</td>
<td>Glass slides</td>
<td>190</td>
<td>[13, 67, 69, 73, 77]</td>
</tr>
<tr>
<td>Gildersleeve, J.</td>
<td>Neoglycoproteins; glycoproteins</td>
<td>16-well glass slides</td>
<td>73</td>
<td>[37, 38]</td>
</tr>
<tr>
<td>Glycominds</td>
<td>Oligosaccharides</td>
<td>200 microwells</td>
<td>37</td>
<td>[7, 66, 85, 90-93, 95]</td>
</tr>
<tr>
<td>Hsieh-Wilson, L.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>4</td>
<td>[29]</td>
</tr>
<tr>
<td>Koberstein, J.</td>
<td>Oligosaccharides; polysaccharides</td>
<td>Glass slides</td>
<td>10</td>
<td>[47]</td>
</tr>
<tr>
<td>Ligier, F.</td>
<td>Monosaccharides</td>
<td>Glass slides</td>
<td>2</td>
<td>[82]</td>
</tr>
<tr>
<td>Makower, J.</td>
<td>Lipopolysaccharides</td>
<td>Glass slides</td>
<td>9</td>
<td>[88]</td>
</tr>
<tr>
<td>Merchant, D.</td>
<td>Polysaccharides</td>
<td>Glass slides</td>
<td>51</td>
<td>[12]</td>
</tr>
<tr>
<td>Mrak, M.</td>
<td>Monosaccharides</td>
<td>Glass slides</td>
<td>10</td>
<td>[35, 44]</td>
</tr>
<tr>
<td>Pohl, N.</td>
<td>Oligo-, disaccharides</td>
<td>Glass slides</td>
<td>8</td>
<td>[19, 20]</td>
</tr>
<tr>
<td>Rubina, A.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>6</td>
<td>[9]</td>
</tr>
<tr>
<td>Ruhl, S.</td>
<td>Neoglycoconjugates; glycoproteins</td>
<td>Nitrocellulose membrane</td>
<td>28</td>
<td>[96]</td>
</tr>
<tr>
<td>Schmidt, R.</td>
<td>Monosaccharides</td>
<td>Glass slides</td>
<td>7</td>
<td>[30]</td>
</tr>
<tr>
<td>Seeburger, P.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>14</td>
<td>[8, 28, 34, 39, 57-59, 83, 94]</td>
</tr>
<tr>
<td>Shin, I.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>22</td>
<td>[25, 31, 32, 36]</td>
</tr>
<tr>
<td>Spranger, N.</td>
<td>Oligosaccharides; neoglycoconjugates; glycoproteins</td>
<td>Glass slides</td>
<td>7</td>
<td>[46]</td>
</tr>
<tr>
<td>Sugahara, K.</td>
<td>Neoglycolipids</td>
<td>Nitrocellulose membrane</td>
<td>14</td>
<td>[87]</td>
</tr>
<tr>
<td>Uzawa, H.</td>
<td>Anionic glycopolymers</td>
<td>Gold surface</td>
<td>3</td>
<td>[18]</td>
</tr>
<tr>
<td>Wang, D.</td>
<td>Neoglycoproteins; glycosaminoglycans; neoglycoconjugates</td>
<td>Glass slides</td>
<td>51</td>
<td>[10, 47, 89]</td>
</tr>
<tr>
<td>Wang, P.</td>
<td>Neoglycolipids</td>
<td>Plates</td>
<td>4</td>
<td>[16]</td>
</tr>
<tr>
<td>Willats, W.</td>
<td>Polysaccharides; proteoglycans; neoglycoconjugates</td>
<td>Black polystyrene slides</td>
<td>23</td>
<td>[11]</td>
</tr>
<tr>
<td>Wong, C.H.</td>
<td>Oligosaccharides; neoglycolipids</td>
<td>Glass slides</td>
<td>6</td>
<td>[14, 15, 17, 41, 45, 84, 86, 99, 100]</td>
</tr>
<tr>
<td>Yamamoto, K.</td>
<td>Acrylamide</td>
<td>Fiber optic</td>
<td>12</td>
<td>[43]</td>
</tr>
<tr>
<td>Zhi, Z.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>3</td>
<td>[26]</td>
</tr>
<tr>
<td>Zhou, J.</td>
<td>Oligosaccharides</td>
<td>Glass slides</td>
<td>10</td>
<td>[27]</td>
</tr>
</tbody>
</table>

Largest published array:

#### Directly labeled receptors

The most straightforward method for generating a fluorescent signal is directly labeling the protein or macromolecule of interest with a fluorophore (see Fig. 3a). Prior to incubating on the array, the receptor is conjugated to a fluorophore such as fluorescein, Cy3, or Cy5. One typically obtains 1-5 molecules of fluorophore per molecule of receptor. With a limited number of fluorophore molecules, this approach generally has the lowest sensitivity. In addition, covalent labeling of a protein can sometimes interfere with the functional properties of a protein. Nevertheless, this approach has been used successfully in many reports (19, 20, 32-34, 42, 46, 56-59).
Labeled secondary reagents

An alternative to direct labeling of a receptor involves the use of a secondary reagent that specifically binds the primary receptor of interest (see Fig. 3b). For example, a mouse monoclonal antibody can be detected with polyclonal anti-mouse antibody secondary reagents. In general, the array is incubated in the presence of the primary receptor such as a lectin or antibody. After washing away unbound receptor, the array is incubated with the secondary reagent. To generate a signal, the secondary reagent may be labeled with a fluorophore. One advantage of this method is that it is possible to get multiple secondary reagents binding to each molecule of primary receptor leading to increased sensitivity. Alternately, the secondary reagent may be conjugated to an enzyme such as horseradish peroxidase or alkaline phosphatase. To generate a signal, the array is incubated with a substrate that can be converted into a fluorescent product thus generating a large increase in fluorescent signal. Alternatively, a solution-phase fluorescent substrate is converted into a reactive intermediate that can covalently attach to the surface. In this case, soluble substrate is washed away and immobilized fluorophores are detected. The primary advantage of the enzyme-linked approach is sensitivity. As each molecule of enzyme can generate many molecules of fluorescent product, the signal is amplified. Moreover, the signal can be enhanced by increasing the length of time the substrate is incubated with the enzyme. In addition to enhanced sensitivity, use of a secondary reagent allows detection of receptors in the presence of complex mixtures of other proteins due to the specific binding between the primary and secondary proteins. For example, human serum contains a wide variety of different proteins and other components. With the aid of a secondary reagent, glycan-binding antibodies present in human serum can be profiled as biomarkers of disease (see below). Naturally, this approach is limited to situations where a secondary reagent is available.

Other methods of detection

Although labeled receptors and secondary reagents can be very useful, they are not suitable for all applications. Attachment of a label can disrupt binding, and secondary reagents are not available in all cases. Label-free assay systems have been used in protein microarrays to overcome these problems and are potentially useful in glycan array technology. Surface plasmon resonance imaging (60), mass spectrometry (61), oblique-incidence reflectivity difference (OIR-D) scanning microscope (62), and...
Kelvin nanoprobe (63) are some of the tools that have been adopted for label-free assay systems.

**Processing of assay signals**

Processing of signals and analysis of large datasets can present a major challenge for array technology. Several software programs have been used in combination with glycan arrays such as GenePix, ImageJ, QuantArray, ArrayWinX, and ScanRay. As glycan arrays increase in size, the need for bioinformatics tools will increase significantly.

**Applications of Glycan Arrays**

Glycan arrays have been used to evaluate the specificity of carbohydrate-binding antibodies and lectins used as research tools, diagnostics, and therapeutic agents. In addition, they provide useful information for increasing our basic understanding of molecular recognition of carbohydrates. Glycan arrays are also emerging as useful tools for biomarker discovery.

**Identification and characterization of carbohydrate-binding proteins**

Glycan arrays have been used to evaluate the specificity of numerous plant, animal, and microbial lectins.

Plant lectins are used routinely as research tools, diagnostic agents, and therapeutic agents. Plant lectins are frequently used as an initial screen of a glycan array to verify that the assay works. Numerous papers also describe the use of glycan arrays for more comprehensive analysis of plant lectin specificity (38, 40, 43, 64, 65). Although these proteins have been studied for years, array screening has uncovered novel binding properties illustrating the utility of high throughput, unbiased screening.

In addition to published results, the Consortium for Functional Glycomics lists results for many lectins and other proteins on its website.

A notable lectin plays an important role in many biological processes, and information on ligand binding is essential for development of agonists/antagonists. One interesting example is DC-SIGN, a calcium-dependent (C-type) lectin expressed on dendritic cells. DC-SIGN is known to bind HIV and facilitate infection of T cells. DC-SIGN has been profiled on glycan arrays and found to bind blood group antigens, ganglioside terminal sugars, and mannose oligosaccharides (22, 66). The related proteins SIGN-R1 through SIGN-R8 have also been evaluated (67, 68). Dectin-1, another C-type lectin-like receptor on leukocytes that mediates phagocytosis and inflammatory mediator production in innate immunity to fungal pathogens, was profiled using a glycolipid microarray (69). It was shown to bind clusters of 11-13 gluco-oligosaccharides. The scavenger receptor C-type lectin (SRCL) was profiled on a glycan array and was found to bind Le^b^ and related triosaccharides (70). A glycan array was also used to evaluate binding properties of the rat alaglyco-protein receptor, Kupffer cell receptor, macrophage-galactose lectin, and human scavenger receptor C-type lectin (71, 72).

Based on their galactose-binding preferences, the lectins were classified into two categories: one group selective for Lewis antigens and the other showed broad selectivity for various Gal/GalNAc-containing residues. A glycan array consisting of Le^b^, SL^b^, and their sulfated analogs was used to profile siglecs (73). This analysis revealed that the sulfite groups on the carbohydrate epitopes act as modulators of siglec binding.

Through the use of a glycan array, mouse Siglec-F and human Siglec-8 (21) expressed on esoinophils were also shown to recognize 6-sulfogalactose Le^b^ (74). A polysaccharide microarray was used to study interactions between dextran polysaccharides and platelet-derived growth factor BB isofrom (12). A chondroitin sulfate microarray was used as a screening tool to identify an antagonist for a therapeutically important proinflammatory cytokine, tumor necrosis factor-α (29). Another animal lectin profiled with glycan arrays include galectin-1 (75), galectin-4 (76), mannose 6-phosphate receptor (77), dectin-2 (78), Manilla clamt lectin (79), and Langerin (67).

Bacterial, viral, and microbial lectins are a third class of important proteins that have been screened on glycan arrays.

One of the most interesting applications of glycan arrays involved profiling hemagglutinins from pathogenic strains of influenza A virus including the pandemic 1918 strain (80, 81). Avian viruses were found to preferentially bind sialic acid with alpha 2-3 linkages while human viruses bound alpha 2-6 linkages. The difference in specificity provides a potential explanation for the variations in infectivity and pathogenicity. A group of bacterial toxins and cells, Salmonella typhimurium, Listeria Monocytogenes, Escherichia coli, staphylococcal enterotoxin B, cholera toxin, and tetanus toxin, has also been profiled on a glycan array (82). Several lectins with potent antiHIV activity such as cyanovirin, scytovirin, and microcuvirin have been evaluated on glycan arrays (58, 83).

Another important application of glycan arrays is assessing the specificity of glycan-binding antibodies (37, 58, 84-87). Carbohydrate-binding antibodies are important research tools, diagnostics, and therapeutic agents. Information on specificity is critical for the proper interpretation of results or selection of clinical candidates. For example, antibody 2G12 is known to neutralize a broad range HIV-1 isolates in a carbohydrate-dependent manner. Profiling using a glycan array illustrated that the peptide backbone is not required for binding and that the antibody has a preference for Manα1-2Man linkages (58, 84). A glycan array comprising the tumor antigen Globo H and its fragments was used to profile two monoclonal antibodies against Globo H and serum from breast cancer patients (86). This array showed that the fucose residue was essential for recognition by the monoclonal antibodies but not for the serum antibodies. In another glycan array study, various antibodies and lectins that have been used for decades to monitor the expression of the tumor-associated Tn antigen were found to cross-react with other human glycans (37). The results suggested that information on the expression of the Tn antigen used for the development of diagnostics and vaccines may be inaccurate. The antimalaria antibody MG96 was evaluated in an effort...
to identify suitable carbohydrate antigens for a malaria vaccine (85).

**Profiling serum antibodies as biomarkers for disease**

Glycan arrays have also been used to measure levels of antiacarbohydrate antibodies in patient serum as diagnostic or prognostic markers (10, 88–93). A glycan array was used to measure the antiacarbohydrate antibody levels in eight Hodgkin’s lymphoma patients (92). From the study, elevated antibody levels against four carbohydrate epitopes were detected in the lymphoma patients. Blood samples from multiple sclerosis and Crohn’s disease patients were screened on a glycan array for elevated antibody expression (93). In another report, a glycan array was used to profile 167 multiple sclerosis patient sera (93). From this study, antibodies to Glc1–4GlcNAC were identified as novel biomarkers for relapsing/remitting multiple sclerosis. A set of 72 Crohn’s disease patient sera was also profiled using a microarray (94). It was reported that antichitobioside and antiaminobiose antibodies are novel serologic markers associated with Crohn’s disease.

**Characterization of cell and virus binding**

Cell–cell recognition can be a complex process involving multiple interactions between different proteins and glycans. In some cases, the specific proteins involved are not known or are not functional on their own. In addition, cell binding can trigger signaling cascades or other responses. As a result, evaluation of binding of whole cells to glycans can be very useful and several reports have demonstrated successful screening of cells on glycan arrays. Examples include screening of E. coli strains (25, 94), chicken hepatocytes and CD4+ human T-cells (95), and Helicobacter pylori (96). Viruses and viral capsids have also been screen on glycan arrays (97, 98).

**Identification of enzyme substrates and inhibitors**

Glycan arrays have also been used to identify inhibitors of enzymes. Inhibitors for the enzyme fucosyltransferases were discovered from a library of triazole-containing compounds using a glycan array (99). Due to the importance of aminoglycosides as therapeutic agents and the attractiveness of RNA as a drug target, a number of aminoglycoside–glycan arrays have been developed. One aminoglycoside array was used for the high throughput analysis of glycan–RNA interactions (100). The molecular recognition between oligonucleotide mimics of aminoglycoside-binding sites in the bacterial 30S ribosome and various aminoglycoside–ribosomes and various aminoglycoside–antibiotics has been studied by the use of a glycan array (101). Another aminoglycoside–membrane microarray was developed to assess the potency of these glycans as improved antibiotics (59). The interactions between aminoglycosides and two aminoglycoside acetyltransferases implicated in antibiotic resistance were analyzed. The array analysis showed that some of the aminoglycoside mimetics inhibit the aminoglycoside acetyltransferases, and thus have potential as improved antibiotics.

**Conclusions**

Glycan arrays have emerged as powerful tools for analysis of carbohydrate–macromolecule interactions. As research tools, they are being used to identify ligands for glycan-binding proteins and evaluate specificity of carbohydrate–binding proteins. In a clinical setting, glycan arrays hold enormous potential for the development of diagnostic and therapeutic agents. For example, several studies have demonstrated the utility of glycan arrays for profiling antiglycan antibodies in serum as diagnostic markers of disease. As the technology matures, the range of applications and capabilities will continue to expand and glycan array analysis is anticipated to become routine in chemistry, biology, and clinical laboratories.

**References**


65. Liu SY, Wong CH, Piek B, Gus O, Kumay KB, Lee B. DC-SIGN binds to HIV-1 glycoprotein 120 in a distinct but overlapping fashion compared with ICAM-1 and ICAM-3 in Caco-2 cells. J. Biol. Chem. 2006;281:10297-10305.


See Also

Glycan Synthesis, Key Reactions of
Glycan Synthesis, Key Strategies for
Glycopeptides and Glycoproteins, Synthesis of
Isolation of Glycans
Sugar-Lectin Interactions in Cell Adhesion

Glycan-Glycan Interactions
Glycan-Protein Interactions
Glycan-Glycan Interactions
Glycan-Protein Interactions
Glycan Sequencing and Data Integration for Glycomics

Rahul Raman and Ram Sasisekharan, Department of Biological Engineering, Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts
doi: 10.1002/9780470048672.wecb247

Databases and informatics tools are central to rapid acceleration in the progress of the genomics and proteomics fields. Glycomics, an emerging paradigm in the age of postgenomics, involves understanding the structure–function relationship of glycans in fundamental biological processes. In comparison with genomics and proteomics, the advancement of glycomics has faced unique challenges in the pursuit of developing tools, biological readouts, and an informatics platform to investigate glycan structure–function relationships. These challenges are because of the structural diversity and heterogeneity in glycans that develops from their complex non-template biosynthesis. This review focuses on areas of technologies for sequencing glycans and the importance of developing a bioinformatics platform to integrate the diverse data sets generated using the different technologies to allow a systems approach to glycan structure–function relationships. Although aspects of glycan analysis and integration of glycomics data sets have been covered in earlier reviews, this review brings these aspects together for both linear and branched glycans.

Glycosylation, or the attachment of complex carbohydrates (glycans) to proteins, is the most extensive and complex form of posttranslational modification and is required for the functional diversity to generate extensive phenotypes from a limited genotype. Glycans can be classified broadly as linear and branched, based on their chemical structure. Branched glycans (Fig. 1) are present as N-linked and O-linked glycosylation on glycoproteins or on glycolipids (1, 2). The majority of linear sugars are glycosaminoglycans (GAGs) that contain long polymers of sulfated disaccharide repeat units that are O-linked with a core protein to form a proteoglycan aggregate (3) (Fig. 1). The known intracellular roles of protein glycosylation include facilitating and stabilizing the protein folding process and targeting proteins to various intracellular compartments (1, 2). The glycans present on the glycoproteins and proteoglycans at the cell surface and the cell–extracellular matrix (ECM) interface are instead in an environment of many proteins including growth factors, cytokines, immune receptors, enzymes, and others. The ubiquitous distribution of glycans at the cell–ECM interface results in their involvement in the dynamic interplay between the cell and its environment. Glycans thus play a central role in fundamental biological processes such as development, cancer, infection, and immunity (4–8). The numerous biological roles of glycans are attributed to their specific interactions with a variety of proteins. Therefore, glycans modulate protein activity at the cell–ECM interface.

Informatics approaches are emerging as critical tools to accelerate understanding of the structure–function relationships of genes and proteins in fundamental biological processes. Such approaches involve storing, integrating, manipulating, analyzing, and mining sequence, structure, and biological-function information on genes and proteins at the molecular, cellular, tissue, and higher levels. The need for informatics tools continues to grow in the face of rapid data evolution from single-gene or protein level to complex high-throughput data sets on multiple genes or proteins at multiple levels (molecular, cellular, etc.). In the case of glycans, comprehending structure–function (or sequence–activity) relationships presents unique challenges when compared with genes and proteins. The biosynthesis of glycans is a non-template-driven process that involves the coordinated expression of several glycosyltransferases (and sulfotransferases in the case of GAGs), some of which have additional tissue-specific isoforms (1–3, 5). The complex biosynthesis and the lack of a proofreading machinery lead to an inherent heterogeneity and large diversity of glycan structures. Owing
Glycan Sequencing and Data Integration for Glycomics

Figure 1  Linear and branched glycans. Shown in figure are the different classes of glycan in their symbol nomenclature. The directionality of the glycans is from the nonreducing end at the top to the reducing end at the bottom with the arrows indicating the extension at the nonreducing side. The links between monosaccharides contain the anomeric configuration of the monosaccharide (α-alpha and β-beta) and the oxygen atom in the reducing end monosaccharide to which it is linked. The common terminal sugars in N-linked and O-linked branched glycans are shown in a dotted box (α3/6 indicates α3 or α6 link for the sialic acid). In the case of GAGs, representative chains of heparan sulfate (HS), chondroitin sulfate (CS), and dermatan sulfate (DS) are shown.

to their mode of biosynthesis, ubiquitous subcellular distribution, and the glycoprotein diversity that develops from one or more glycosylation sites, glycans additionally always need to be considered as a heterogeneous mixture of different chemical structures when isolated from cells and tissues. These challenges have hampered the development of analytical techniques that accurately can define the chemical structures on the ensemble of glycans isolated from glycoproteins, cells, or tissues. An understanding of the biochemical basis for glycan-protein interactions is complicated more by the multivalency and graded affinity that involve an ensemble of glycans making multiple contacts with multivalent protein binding sites (3, 9, 10).

In light of the above complexities, it is evident that developing an informatics platform is critical to decode the structure-function relationships of glycans. Informatics-based approaches are especially key to decode the sequence of glycans because no automated sequencing strategies are available similar to those that are found for DNA and proteins. Additionally, a need has developed to cut across diverse data, ranging from glycan biosynthesis, glycan characterization, and glycan-protein interactions, to take an integrated system or glycomics approach to delineate glycan structure-function relationships. This review is divided into two parts to provide a perspective on the development and application of informatics tools for glycans. The first part covers the critically needed informatics tools that sequence glycans to decode their sequence diversity. The second part covers the development of an informatics platform that bridges multiple data sets collected using different technologies to provide a systems framework for understanding glycan structure-function. These arguments are illustrated using examples of both linear and branched sugars.

Decoding Glycan Diversity: Strategies for Sequencing Glycans

The complete set of glycans in a given cell (analogous to the genome or proteome of a cell) includes different types of glycoconjugates such as glycolipids, glycoinositolphospholipids, glycans attached to proteins in the form of glycoproteins, peptidoglycans, and proteoglycans. Decoding the complete chemical structure (analogous to the primary sequence of DNA and protein) of each of these glycans in the entire repertoire mentioned presents a formidable challenge. Many efforts have focused on the isolation and characterization of glycans from glycoproteins (for branched glycans) and proteoglycans (for linear GAGs).

Isolation of glycans from glycoproteins

Two distinct approaches characterize glycans on glycoproteins. The first approach involves releasing the glycans and characterizing the structures of the entire ensemble of these released glycans. The second approach instead fragments the entire glycoprotein to glycopeptides and characterizes the peptide and glycan component of each glycopeptide. The second approach involves characterization of glycoproteins in terms of the distinct glycan structures at each of their glycosylation sites. Glycans can be released from glycoproteins by using either chemical methods, such as hydrazinolysis (N-linked)
and β-elimination (O-linked), or enzymatic methods that use PNGase F or A (N-linked). Currently, no generic enzymatic methods are available to release O-linked glycans. The enzy-

The next level of characterization involves matching the frag-

mentation patterns of the parent ion from the MS² (and possibly higher order MS³, MS⁴, ..., MSⁿ) data to reference data sets and de-

termining the sequence of glycans (Table 1) (17–28). It should be

noted that a trade-off occurs between the accurate characteriza-
tion of the exact glycan structure and high-throughput analysis of several glycans in a given sample. In the case of a typical high-throughput data set such as MALDI-MS profile of glycans isolated from cells/tissues, a set of compositions and glycan topologies are assigned to a given mass peak based on biosyn-

thetic constraints. The GlycoMod program (17) available on the Ev- 

PAzy website provides an automatic calculation of possible monosaccharide compositions for a given parent ion mass. The 

Carbonator software (18) provides an automated annotation of a MALDI-MS glycan profile with glycan topologies selected from a library of topologies generated using biosynthesis rules. These programs provide a quick read of the graphical represen-
tation of glycan topologies for an entire repertoire of glycans in a biological sample.
## Table 1: Resources for sequencing glycans

<table>
<thead>
<tr>
<th>Analytical Tool/software</th>
<th>Reference/online Access</th>
<th>Overall logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cartoonist</td>
<td>Goldberg et al. [18]</td>
<td>Automated annotation of MALDI-MS using a library of glycan topologies derived from glycan biosynthesis rules. Applicable mainly to mammalian N-linked glycans.</td>
</tr>
<tr>
<td>GlycoSearch-MS</td>
<td><a href="http://www.dkfz.de/spec/glycosciences.de/sweetdb/ms/">www.dkfz.de/spec/glycosciences.de/sweetdb/ms/</a> [19]</td>
<td>Matches MS² data to a theoretical fragmentation of structures in a glycan structure database.</td>
</tr>
<tr>
<td>Glyco-Peakfinder</td>
<td><a href="http://www.dkfz.de/spec/EuroCarbDB/applications/ms-tools/GlycoPeakfinder/GlycoPeakfinder.action">www.dkfz.de/spec/EuroCarbDB/applications/ms-tools/GlycoPeakfinder/GlycoPeakfinder.action</a> [20]</td>
<td>Automated annotation of MS and MS⁵ fragmentation spectra with different types of fragment ions (glycosidic and cross ring cleavages). The composition of mass ion peaks can be used to search glycan structures database for possible structures.</td>
</tr>
<tr>
<td>GlycanBuilder/GlycoWorkbench</td>
<td><a href="http://www.dkfz-heidelberg.de/spec/EuroCarbDB/GlycoWorkbench/">www.dkfz-heidelberg.de/spec/EuroCarbDB/GlycoWorkbench/</a> [21]</td>
<td>An integrated platform to build glycan structures, simulate their fragmentation, and enable automated annotation of MS profile along with the MS⁵ fragmentation spectra.</td>
</tr>
<tr>
<td>GlycosidIQ</td>
<td><a href="http://www.glycosuite.com">www.glycosuite.com</a> [22]</td>
<td>Matches MS² data to a theoretical fragmentation of a database of reported glycan structures in literature. Not open access and requires subscription.</td>
</tr>
<tr>
<td>Saccharide Topology Analysis Tool STAT</td>
<td>Gaucher et al. [23]</td>
<td>Reference data set containing set of all possible glycan topologies satisfying set of compositions generated based on parent ion mass. Mass of MS⁵ fragmentation pattern matched to fragmentation of reference structures.</td>
</tr>
<tr>
<td>StrOligo</td>
<td>Ethier et al. [24]</td>
<td>Compositional mass predicts possible structures based on known N-linked mammalian biosynthetic pathways, which are theoretically fragmented and matched to experimental data.</td>
</tr>
<tr>
<td>Oligosaccharide Subtree Constraint Algorithm OSCAR</td>
<td>glycome.unh.edu/tools/GlySpy/ Lapadula et al. [25]</td>
<td>Theoretical MS⁵ fragmentation trees are used to predict possible glycan topologies matching compositions for a given mass derived from a fragment composition finder database. The inferential rules to match MS⁵ fragmentation are constructed de novo, so this tool does not match with glycan structures database and thus is better suited to identify novel structures.</td>
</tr>
<tr>
<td>Automated interpretation of HPLC glycan profiles</td>
<td>Rudd et al. [26]</td>
<td>Parent glycan samples are treated with a combination of exoglycosidase specifically to remove selected monosaccharides from nonreducing end. Peak shifts are interpreted based on database of HPLC migration times for known glycan standards.</td>
</tr>
<tr>
<td>CASPER</td>
<td><a href="http://www.casper.org.suse/casper/">www.casper.org.suse/casper/</a></td>
<td>Enables determination of glycan structure based on ¹H or ¹³C NMR chemical shifts, component and link analysis.</td>
</tr>
<tr>
<td>Glycosciences NMR DB</td>
<td><a href="http://www.glycosciences.de/sweetdb/nmr/">www.glycosciences.de/sweetdb/nmr/</a></td>
<td>Database of ¹H and ¹³C NMR shifts of glycans with tools to search glycans using NMR shifts or estimate shift for a given glycan.</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Analytical Tool/software</th>
<th>Reference/online Access</th>
<th>Overall logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEN-MALDI sequencing of GAGs</td>
<td>Venkataraman et al. (27)</td>
<td>Disaccharide composition from capillary electrophoresis (CE) and MALDI-MS profiles of chemical and enzymatic degradation or modification of GAGs are applied as constraints using a computational framework to reduce the entire GAG sequence space for that given composition to converge on the final sequence or a set of sequences that satisfy all the constraints.</td>
</tr>
<tr>
<td>PEN-NMR sequencing of GAGs</td>
<td>Guerini et al. (28)</td>
<td>Similar to PEN-MALDI strategy but uses monosaccharide composition and link abundance from NMR along with disaccharide composition from CE as constraints to converge on the final sequence or set of sequences that satisfy all the constraints.</td>
</tr>
</tbody>
</table>

The above MS methods and data interpretation tools are valuable to obtain rapidly the most probable structures of an entire ensemble of glycans isolated from cells and tissues. Determining the exact structure of each glycan (in a mixture), including the anomeric configuration of each sugar and the specific link ab initio, remains a formidable task. One approach toward this task employs exo-glycosidases that specifically cleave monosaccharides from the nonreducing end of a glycan. The HPLC chromatographic profile of parent glycan sample and the shifts in peaks that result from treatment with the exo-glycosidases are matched with reference chromatographic profiles of known glycan structures using software tools to derive the exact structure of glycans in the sample (26, 31). Nuclear magnetic resonance (NMR) spectroscopy is another valuable tool to obtain link information in glycans (32). The characteristic NMR chemical shifts and coupling constants of various glycans have been compiled in databases to facilitate the assignment of glycan structures based on NMR data (Table 1). Such approaches for fine structure characterization of glycans have limitations when considering biological samples composed of a diverse mixture of glycans. The requirement of high sample amounts (in the case of NMR) and multiple steps of exo-glycosidase treatment for larger glycan mixtures complicate the use of these techniques for high-throughput analysis. In the case of linear sugars such as GAGs, because of the heterogeneity in the chain length and sulfation pattern (Fig. 1), their isolation from biological samples results in a mixture of different chain lengths and compositions. To determine the sequence of individual GAG oligosaccharides in a mixture, it is necessary to subfractionate them based on size and charge and then to purify these fractions to homogeneity. The chain length of a homogeneous GAG oligosaccharide isolated in this manner is typically 4–8 disaccharide repeat units. It is difficult to determine the exact sequence of a GAG chain (particularly for highly sulfated GAGs) that is greater than 10 disaccharide repeat units. Given that GAG–protein binding specificity typically involves chain length of 2–5 disaccharide units, sequencing these chains does provide important information on the structure-function relationships of GAGs (3). The two main areas for generating tools for the characterization of GAGs have been the development of chemical and enzymatic tools for the controlled and specific depolymerization or modification of a GAG chain and the development of analytical tools based on chromatography, electrophoresis, MS, and NMR techniques (see Reference 3 for details). Informatics-based methods have also been developed to capture the information density of GAGs and to enable the application of data from a combination of tools as constraints to sequence GAGs (27, 28). The underlying rationale of these approaches is to begin with a comprehensive master list of GAG sequences that represents the entire sequence space for a given chain length and overall composition (monosaccharide or disaccharide units). The data from a series of GAG depolymerization/modification profiles (mass
spectrometry, electrophoresis, and chromatographic) serve as constraints to reduce this master list in an iterative fashion to the final sequence or set of probable sequences that satisfy all the constraints. These strategies provide an unbiased approach to sequencing GAGs through elimination of sequences from a master list that do not satisfy the experimental constraints.

### Integrated Informatics Platform for Glycans

The type of data sets needed to fully define structure-function relationships of glycans goes beyond the complex analytical data for structural characterization of glycans. Several international collaborative efforts (Table 2) recognize the need for an integrated glycomics approach and are developing novel resources and technologies to better understand glycan structure-function relationships. The data sets and databases developed by these initiatives revolve around three main components that are discussed in detail below (summarized in Table 3): 1) glycan structures, 2) glycan biosynthesis pathways, and 3) glycan-protein interactions.

#### Glycan structures database

The common goal of all the major glycomics initiatives is the development of a database of glycan structures in parallel with the analytical data sets (described above). The common source for the initial set of glycan structures to seed different glycan databases was the CarbBank database. This database was discontinued in the 1990s because of a lack of resources that was required to sustain its development and ensure the quality of glycan structures based on extensive curation. The diversity in the sugar building blocks of glycans, the link between the sugars and two-dimensional topology (degree and extent of branching), offers challenges in the development of standardized data formats to represent glycan structures across different databases. Because of this obstacle, independent data formats were developed in the various international initiatives to encode glycan structures in databases. An important issue develops from the ambiguities in glycan structure that result from incomplete assignment of their exact structure. These ambiguities include unspecified links, anomeric configurations, and uncertainties in the connectivity of a specific sugar (typically terminal sugar) or a set of sugars to a particular two-dimensional glycan topology.

Consequently, different databases employ disparate mechanisms to manage these ambiguities; they represent ambiguities as they are or deconvolute them into a set of likely exact structures. An important feature of the glycan structures databases is the tools to query the database using parameters such as monosaccharide composition, substructure, and mass. Different algorithms have been developed to search, score, and compare glycan structures using tree matching methods, scoring matrices, and dynamic programming tools. The interpretation of the search results from these tools is still demanding primarily because of the lack of standardized definitions of glycan structures across different databases. Querying different glycan structures databases using the same query parameters often gives dramatically different results in this context.

### Glycan biosynthesis pathways

The Consortium for Functional Glycomics (CFG) has led a collaboration with Japanese and European initiatives to develop a comprehensive annotation of the enzymes involved in glycan biosynthesis. Currently, these annotations are disseminated using glycosylation pathway interfaces and glycosyltransferase databases of the CFG, KEGG glycans, Complex Carbohydrates Research Center (CCRC), and CAZy. The CFG has developed customized DNA microarrays that appropriately represent glycan biosynthesis and binding protein genes to study their expression pattern. Hundred of samples representing various tissues and cell types have been analyzed on the CFG glyco-gene microarrays. The focus of experiments includes the analysis of different tissues and binding protein genes to study their expression pattern. A valuable resource for quantitative analysis of gene expression is the quantitative real-time PCR analysis of glycan-related genes from the CCRC.

### Table 2: International Glycomics Initiatives

| Consortium for Functional Glycomics (CFG) (USA) | www.functionalglycomics.org |
| Complex Carbohydrates Research Center (CCRC) (USA) | www.ccrcl.edu |
| LipidMaps Consortium (USA) | www.lipidmaps.org |
| Glycosciences.de (Germany) | www.eurocarb.org.de |
| UK Glycosciences (UK) | www.glycosciences.de |
| Centre for GlycoBiology (Israel) | www.glycoarrays.org.uk |
| Human Disease Glycomics-Proteome Initiative (HGPI) (Japan) | www.hgpi.jp |
| Kyoto Encyclopedia of Genes and Genomes (KEGG) Glycan (Japan) | www.genome.jp/kegg/glycan/ |
### Table 3  Open access online web-based resources for glycomics

<table>
<thead>
<tr>
<th>Glycan Structures and Analytical Data</th>
<th>Glycans from CarbBank, synthesized by CFG, and glycan topologies used for annotation of CFG MALDI-MS data.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG Glycan Structures DB</td>
<td><a href="http://www.functionalglycomics.org/glycomics/molecule.jsp/carbohydrate/">www.functionalglycomics.org/glycomics/molecule.jsp/carbohydrate/</a></td>
</tr>
<tr>
<td>Glycosciences Glycan Structures DB</td>
<td><a href="http://www.glycosciences.de/sweetdb/index.php">www.glycosciences.de/sweetdb/index.php</a></td>
</tr>
<tr>
<td>Glycome DB</td>
<td><a href="http://www.glycome-db.org/about.action">www.glycome-db.org/about.action</a></td>
</tr>
<tr>
<td>KEGG Glycan Structures DB</td>
<td><a href="http://www.genome.jp/kegg/glycan/">www.genome.jp/kegg/glycan/</a></td>
</tr>
<tr>
<td>Bacterial Carbohydrate Structures Database</td>
<td><a href="http://www.glyco.ac.ru/bcddb/start.shtml">www.glyco.ac.ru/bcddb/start.shtml</a></td>
</tr>
<tr>
<td>Glycobase</td>
<td>glycobase.univ-lille1.fr/base/</td>
</tr>
<tr>
<td>Glycoconjuate DB</td>
<td>akashia.sci.hokudai.ac.jp/</td>
</tr>
<tr>
<td>CFG MALDI-MS Glycan Profile of Cells and Tissues</td>
<td><a href="http://www.functionalglycomics.org/glycomics/publicdata/glycoprofiling.jsp">www.functionalglycomics.org/glycomics/publicdata/glycoprofiling.jsp</a></td>
</tr>
<tr>
<td>SUGABASE NMR DB</td>
<td>boc.chem.uu.nl/sugabase/sugabase.html</td>
</tr>
<tr>
<td>CCRC NMR and MS Data</td>
<td><a href="http://www.ccrc.uga.edu/specdb/specdbframe.html">www.ccrc.uga.edu/specdb/specdbframe.html</a></td>
</tr>
</tbody>
</table>

## Glycan Binding Proteins (GBPs) and Glycan-Protein Interactions

<table>
<thead>
<tr>
<th>Glycan Binding Proteins (GBPs) and Glycan-Protein Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG GBP DB</td>
</tr>
<tr>
<td>Thorkild's Lectin Page</td>
</tr>
<tr>
<td>CFG Glycan Array Screening Data</td>
</tr>
<tr>
<td>Lectines 3D</td>
</tr>
<tr>
<td>GlycoEpitope DB</td>
</tr>
<tr>
<td>SugarBindDB</td>
</tr>
</tbody>
</table>

## Glycan Biosynthesis and Biology

<table>
<thead>
<tr>
<th>Glycan Biosynthesis and Biology</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFG GT DB</td>
</tr>
<tr>
<td>KEGG GT DB</td>
</tr>
</tbody>
</table>

This table includes open access online web-based resources for glycomics. Each resource is associated with a specific type of glycan structure, analytical data, or glycan-protein interactions. Some examples include CFG Glycan Structures Database, Glycosciences Glycan Structures Database, Glycome DB, KEGG Glycan Structures Database, Bacterial Carbohydrate Structures Database, Glycobase, Glycoconjuate DB, CFG MALDI-MS Glycan Profile of Cells and Tissues Database, SUGABASE NMR DB, CCRC NMR and MS Data, CFG GBP DB, Thorkild's Lectin Page, CFG Glycan Array Screening Data, Lectines 3D, GlycoEpitope DB, SugarBindDB, CFG GT DB, KEGG GT DB. These resources provide valuable information for researchers working in the field of glycomics.
The data generated from these analyses have uncovered novel glycan recognition motifs for these diverse range of samples analyzed. The CFG and other groups have also developed specific databases for GBPs and other lectins (from plants and bacteria) that capture the sequence, three-dimensional structure, known glycan binding specificities, and other information from the proteins (Table 3).

Data integration for glycomics

Clearly, a need to cut across multiple data sets to understand the structure-function relationships of glycans is apparent. A critical component that permits this process is the bioinformatics platform used to store, integrate, and process the information generated by the above methods and to disseminate this information in a meaningful fashion via the Internet to the scientific community worldwide (35, 50). To capture complex relationships between diverse data, it is necessary to develop an object-based relational database. Three primary objects are found in glycomics data sets: GBPs, glycan biosynthetic enzymes, and the glycan structures. The different methodologies that generate data sets are organized into secondary and other levels of objects with defined interrelationships and relationships to the primary objects. The overall scheme for data integration in the CFG databases is shown as an example (Fig. 3). The key identifiers to uniquely define the primary objects include the glycan structure represented in the internal database format, primary sequence (or accession numbers such as GenBank and SwissProt) of the GBPs, and glycan biosynthetic enzymes. The relationship of primary to secondary data objects is established using object-based keys such as sample, protocols, and data keys. Unlike in the case of the primary objects, it is challenging to define standardized criteria that can determine the uniqueness of the secondary objects (such as sample information) because of the lack of standardized ontologies to describe sample and protocol information for glycomics. To address this challenge, a key step would require outlining metadata standards for obtaining various glycomics data sets, including mass-spectrometric glycan analysis data and glycan-protein interaction data. It is important, for example, to capture mass spectrometer instrument settings, internal standards, and other experimental conditions used to characterize glycans from a given sample. The blueprint of the object-based relational database is a data model that captures data definitions and interrelationships between the data sets using the specific object-based identifiers, which is quite complex.
Glycan Sequencing and Data Integration for Glycomics

Figure 2  Multivalent glycan–protein interactions. Shown on the left is a schematic of multivalent interactions between branched N-linked and O-linked glycans on glycoproteins with CRDs glycan binding proteins. The binding specificity is critically governed by the chemical structure and three-dimensional topology of the glycan ligand motif and the multivalency. Given that the glycan ligand motif can be present on the same or different glycoproteins, it is important to characterize the ensemble of glycans isolated from a given cell. Shown on the right is a schematic of GAG–protein interactions. GAGs are typically involved in facilitating the assembly of protein–protein complexes, enzyme–inhibitor complexes, and ligand–receptor complexes on the cell surface and thus play important roles in extracellular signaling. The schematic shows an example of HS GAG (H-glucosamine; I–iduronic acid) wherein the chain length, chain conformation, and sulfation pattern governs whether an active or inactive protein–protein complex is assembled.

It is important, therefore, to develop a software architecture that keeps this complexity hidden from the user during the actual data acquisition and dissemination. The three-tier software architecture containing a back-end relational database to store the data and annotate relationships, a middleware application layer that communicates between the database and the user interface, and the top layer that comprises the user interfaces to the database is best suited for this purpose. This software architecture facilitates the easy deposit of data into the database, which is organized automatically into the relational tables by the middleware application layer.

Central to this data integration is an ability to link orthogonal data sets derived from identical, or similar, samples. For example, the gene expression profile of a specific tissue or cell line isolated from a given strain of transgenic mice needs to be automatically associated in the database with orthogonal information such as glycan profile, histological staining, and immunological profile from a similar or identical sample. Such integration would permit researchers to cut across multiple data sets and begin posing questions such as "Does the expression of glycosyltransferase correlate with glycan profile of that tissue?" or "Can the pathological analysis of the tissue be explained on the basis of gene expression profile?" The bioinformatics platform ultimately needs to support computational tools that perform data mining analysis on the large-scale glycomics data sets. The prediction of glycan structures based on gene expression profiles of glycan biosynthetic enzymes (41) is now enabled because of the user-friendly access to diverse data sets via relational databases.

Significance and Future Directions

Glycomics is an emerging field that adds new dimensions to both technology development and obtaining an understanding of fundamental biological processes in this age of postgenomics. The two fundamental questions advancing glycomics are as follows: 1) How do we define glycan diversity in the context of biosynthesis and specificity in modulating a biological process and 2) how do we define specificity in the context of multivalent glycan–protein interactions? It is important to consider these questions in light of the numerous efforts underway to discover novel glycan-based biomarkers and therapeutic targets (51–54). In the case of biomarker discovery, two distinct approaches are found. The glycomics approach begins with the discovery of a specific glycan ligand motif specifically involved in a biologically important glycan–protein interaction. The use of this motif as a biomarker then is investigated by studying the expression of this motif as a part of glycan structure
or glycan biosynthesis enzymes that generate this motif in different biological samples. The other approach is more of a high-throughput analytical approach that relies on the various analytical and informatics tools to characterize glycan structures in different samples and to identify unique markers based on differential expression of specific structures in a given sample. The generation of diverse data sets and their integration via an informatics platform is critical to both approaches. From an informatics perspective, the key issues that need to be addressed are as follows: 1) developing standardization and quality control of the data, 2) developing consistent format for representation and transfer of complex glycan structures across different databases, and 3) developing robust ontologies and structured vocabulary for glycomics (35).

A significant portion of the high-throughput data sets on glycan analysis and glycan-protein interactions are fairly recent, and therefore a critical need is found to standardize the generation of these data so they can be compared across data sets generated by different initiatives (35). Also, awareness is increasing of the need to develop data exchange formats, such as X ML, for consistent description of glycan structures and glycomics data sets across different large-scale glycomics initiatives (19). Leading glycomics initiatives recently agreed to adopt the Glyde-II XML format (52). The next step toward standardization of glycan structure databases is to establish the standards for incorporating glycan structures into a database to develop the glycan database into an international resource similar to GenBank and SwissProt. Collaborative discussions between the large-scale glycomics initiatives to address these issues are notable steps toward advancing glycomics.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences Glue Grant U54 GM62116. The authors thank the CFG members, the members of Bioinformatics Core (B) of CFG, for the work described in this review, and V. Sasisekharan for critical reading of the manuscript.

References


Figure 3 Data integration in glycomics. An example of integrating the diverse data sets generated by CFG to the three primary databases viz glycan structures, GBP, and GT is shown. The arrows indicate the connectivity between the objects using the primary identifiers for genes, proteins, glycan structures, and samples analyzed. This integration enables seamless navigation across these data sets. For example, starting from a glycosylation pathway, it is possible to navigate to a specific glycan structure whose biosynthesis involves a particular enzyme. If a glycan ligand motif present on the glycan array is a part of this glycan structure, it is possible to determine which of the samples analyzed on the array gave good binding signals to this motif. If any of the samples analyzed are mammalian GBP, then it is possible to access an entire portal of information on that particular GBP.
Glycan Sequencing and Data Integration for Glycomics


Differential Scanning Calorimetry to Study Lipids and Lipid Membranes

Ruthven N.A.H. Lewis, David A. Mannock, and Ronald N. McElhaney, University of Alberta, Edmonton, Alberta, Canada

doi: 10.1002/9780470048672.wecb049

Differential scanning calorimetry (DSC) is a relatively rapid direct and nonperturbing thermodynamic technique for studying the thermotropic phase behavior of hydrated lipid dispersions and of reconstituted lipid model and biological membranes. DSC can accurately and reliably determine the temperature, enthalpy, entropy and cooperativity of a wide variety of lipid phase transitions and how these parameters are influenced by variations in hydration, and in the pH and the ionic strength and composition of the aqueous phase. A too, the effects of the presence of membrane-associated sterols, peptides and proteins, as well as toxins, drugs, and other agents, on the thermotropic phase behavior of lipid membranes can be determined. Under appropriate conditions, DSC can also characterize the kinetics of some lipid phase transitions. The thermodynamic data provided by DSC, therefore, can provide valuable information about the phase state and organization of lipid assemblies and about how the structure and physical properties of lipid model and biological membranes are modulated by other membrane constituents and by the environment. However, because DSC is a thermodynamic and not a structural technique, it is most valuable when applied in conjunction with a direct structural technique, such as X-ray diffraction, and with nonperturbing spectroscopic methods, such as nuclear magnetic resonance and Fourier transform infrared spectroscopy.

Lipid Mesomorphic Phase Behavior

Membrane lipids are invariably polymorphic: that is, they can exist in a variety of kinds of organized structures, especially when hydrated. The particular polymorphic form that predominates depends not only on the structure of the lipid molecule itself and on its degree of hydration, but also on such variables as temperature, pressure, ionic strength and pH (see References 11 and 12 and article Lipids, Phase Transitions of). However, under physiologically relevant conditions, most (but not all) membrane lipids exist in the lamellar or bilayer phase. In the lamellar phase, the lipid bilayer is stabilized by hydrophobic interactions between the hydrocarbon chains in the interior of the lipid bilayer and by polar interactions with the glycerol backbone and polar headgroup regions of the host lipid bilayer. In addition to playing a structural role in determining the topology and stabilizing the active conformation of peripheral and integral membrane proteins, the physical properties of the lipid also markedly influence the activity and thus presumably the conformation and dynamics of many membrane proteins (4–8, but see 9). Specifically, the physical state (lamellar gel versus liquid-crystalline), fluidity, hydrophobic thickness, lipid lamellar/nonlamellar phase propensity (lipid shape), surface charge and surface-charge density, as well as various mechanical properties of the lipid bilayer (10), all modulate the thermal stability and activity of many membrane-associated enzymes, transporters and receptors. Therefore, understanding the thermotropic phase behavior and organization and thus the specific functions of the large number of lipid classes and molecular species that comprise biological membranes, remains a major challenge in membrane biology generally. In this brief review, we consider the applications of DSC to lipid model and biological membranes to address in particular the role of lipid fluidity and phase state and to some degree the role of lipid lamellar/nonlamellar phase propensity in membrane structure and function.
and is also the most biologically relevant. This cooperative phase transition involves the conversion of a relatively ordered gel-state bilayer, in which the hydrocarbon chains exist predominantly in their rigid, extended, all-trans conformation, to a relatively disordered liquid-crystalline bilayer, in which the hydrocarbon chains contain several gauche conformers and exhibit greatly increased rates of intramolecular motions. The gel-to-liquid-crystalline phase transition is accompanied by a pronounced lateral expansion and a concomitant decrease in the thickness of the bilayer, as well as by a small increase in the total volume occupied by the lipid molecules. Evidence also shows that the number of water molecules bound to the surfaces of the lipid bilayer increases during hydrocarbon chain melting. Thermodynamically, the gel-to-liquid-crystalline phase transition occurs when the entropic reduction in free energy that results from hydrocarbon chain isomerism counterbalances the decrease in bilayer cohesive energy that results from the lateral expansion and from the energy cost of creating gauche rotational conformers in the hydrocarbon chains.

Gel-to-liquid-crystalline phase transitions can be induced by changes in temperature and hydration, as well as by changes in pressure and in the ionic strength or pH of the aqueous phase. In this article, we will concentrate on thermally induced phase transitions because these have been studied most extensively and are of direct biological relevance, particularly for organisms that cannot regulate their own temperature. However, hydration-induced (lyotropic) and pressure-induced (barotropic) phase transitions also occur, and these may also be biologically relevant under special environmental circumstances. Finally, phase transitions induced by alterations in pH and in the nature and quantity of ions in the aqueous phase that surrounds the bilayer are also possible, and these transitions may also be of importance in living cells. However, a detailed discussion of these types of lipid phase transitions is beyond the limited scope of the current article, and interested readers should consult appropriate reviews for detailed information on this topic.

Pure synthetic lipids often exhibit gel-state polymorphism, and phase transitions between various forms of the gel-state bilayer can occur. Although we will illustrate this behavior for a common phospholipid, dipalmitoylphosphatidylcholine (DPPC), gel-state transitions will not be emphasized here because with only one known exception (13), they do not seem to occur in the heterogeneous collection of lipid molecular species found in biological membranes. Moreover, certain synthetic or naturally occurring lipid species can exist in liquid-crystalline nonlamellar phases, especially three-dimensional reversed cubic and hexagonal phases. Although the actual existence of nonbilayer lipid phases in biological membranes has never been demonstrated under physiological conditions, the propensity to form such phases likely plays major roles in membrane fusion and other processes (see Reference 14). Moreover, evidence suggests that the relative proportion of bilayer-prefering and nonbilayer-prefering lipids may be biosynthetically regulated in response to variations in temperature and membrane lipid fatty acid composition and cholesterol content in some organisms. Thus, lipid species that in isolation may form nonlamellar phases may have important roles to play in the liquid-crystalline bilayers found in essentially all biological membranes. The transitions between lamellar and nonlamellar lipid phases have been reviewed in detail by us and others elsewhere (see References 14 and 15).

Differential Scanning Calorimetry

As mentioned earlier, the technique of DSC has been of primary importance in studies of lipid phase transitions in model and biological membranes (see References 16–18). The principle of DSC is comparatively simple. A sample and an inert reference (i.e., a material of comparable thermal mass that does not undergo a phase transition within the temperature range of interest) are simultaneously heated or cooled at a predetermined constant rate (dT/dt) in an instrument configured to measure the differential rate of heat flow (dE/dt) into the sample relative to that of the inert reference. The temperatures of the sample and reference may either be actively varied by independently controlled units (power compensation calorimetry) or be passively changed through contact with a common heat sink that has a thermal mass that greatly exceeds the combined thermal masses of the sample and reference (heat conduction calorimetry). For our purposes, the sample would normally be a suspension of lipid or membrane in water or an aqueous buffer, and the reference cell would contain the corresponding solvent alone. At temperatures distant from any thermotropic events, the temperatures of the sample and reference cells change linearly with time, and the temperature difference between them remains zero. The instrument thus records a constant difference between the rates of heat flow into the sample and reference cells, which, ideally, is reflected by a straight, horizontal baseline. When the sample undergoes a thermotropic phase transition, a temperature differential between the sample and reference occurs, and the instrument either actively changes the power input to the sample cell to negate the temperature differential (power compensation calorimetry) or passively records the resulting changes in the rate of heat flow into the sample cell until the temperature differential eventually dissipates (heat conduction calorimetry). In both instances, a change develops in the differential rates of heat flow into the sample and reference cells, and either an exothermic or endothermic deviation from the baseline condition occurs. On completion of the thermal event, the instrument either re-establishes its original baseline condition or establishes a new one if a change in the specific heat of the sample has occurred. The output of the instrument is thus a plot of differential heat flow (dE/dt) as a function of temperature in which the intensity of the signal is directly proportional to the scanning rate (dT/dt).

The variation of excess specific heat (dE/dt) with temperature for a simple two-state, first-order endothermic process, such as the gel-to-liquid-crystalline phase transition of a single, highly pure phosphatidylcholine (PC), is illustrated schematically in Fig. 1. From such a DSC trace, several important parameters can be determined directly. The phase transition temperature, usually denoted Tm, is that temperature at which the excess specific heat reaches a maximum. For a symmetrical curve, Tm represents the temperature at which the transition from the...
relationship:

determined for a particular phase transition, the van’t Hoff

technique is not trivial but should be well

enthalpy, 

depressed by DSC, the measurement must be made with

and sample size, and sample equilibration before data acquisition;

Figure 1

Figure 1 The variation of excess specific heat with temperature during a

two-state, endothermic lipid phase transition. The symbols are explained in the text.

gel-to-liquid-crystalline state is one-half complete. However,

for asymmetric traces, which are characteristic of certain pure

phospholipids and many biological membranes, the $T_m$ does

do not represent the midpoint of the phase transition, and a $T_{1/2}$

value may be reported instead. Once normalized with respect

to the scan rate, the peak area under the DSC trace is a direct

measurement of the calorimetrically determined enthalpy of the

transition, $\Delta H_{cal}$, usually expressed in kcal/mol. The area of the

peak can be determined by planimetry or by the cutting

and weighing technique; alternatively, the calorimeter output

can be digitized, and the $T_m$ and $\Delta H_{cal}$ can be calculated by a

computer. Because at the phase transition midpoint temperature

the change in free energy ($\Delta S$) of the system is zero, the entropy

change associated with the transition can be calculated directly

drop from the equation:

$$\Delta S = \Delta H_{cal} / T_m$$

where $\Delta S$ is normally expressed in cal/K·mol$^{-1}$.

The sharpness or cooperativity of the gel-to-liquid-crystalline

phase transition can also be evaluated from the DSC trace.

The sharpness of the phase transition is often expressed as the

temperature width at half-height, $\Delta T_{1/2}$, or as the temperature

difference between the onset or lower boundary of the phase

transition, $T_0$, and the completion or upper boundary, $T_l$, or

$\Delta T = T_l - T_0$. The $\Delta T_{1/2}$ values may range from $0.1$ C for very pure synthetic phospholipids to as much as $10-15$ C for biological membranes. From the $T_m$ and $\Delta T_{1/2}$ values determined for a particular phase transition, the van’t Hoff

enthalpy, $\Delta H_{vH}$, can be approximately determined from the relationship:

$$\Delta H_{vH} \approx 6.9 T_m / \Delta T_{1/2}$$

From the ratio $\Delta H_{m}/\Delta H_{vH}$, the cooperative unit size (CUS) in molecules can be determined. The CUS is a measure of the degree of intermolecular cooperation between phospholipid molecules in a bilayer; for a completely cooperative, first-order

phase transition of an absolutely pure substance, this ratio

should approach infinity, whereas for a completely noncoopera
tive process, this ratio should approach unity. Although the

absolute CUS values determined should be regarded as tentative,
because this parameter is markedly sensitive to the presence

of impurities and may be limited by instrumental parameters,
carefully determined CUS values can be useful in assessing the

purity of synthetic phospholipids and in quantitating the degree

of cooperativity of lipid phase transitions.

It must be stressed that the thermodynamic parameters de

rived from DSC measurements will be valid only if measure

tments are performed under conditions where the instrument

response is true to the properties of the sample (so-called

high-fidelity DSC) and if the thermotropic process being stud

ied is at equilibrium throughout the measurement. In practice,

this statement means that the measurement must be made with

a high-sensitivity instrument operating with a modest thermal

load at scan rates that are slow relative to the thermal time con

stant of the instrument and to both the width and half-life of the

thermotropic process under investigation. For processes such as

the gel-to-liquid-crystalline phase transition of certain single,

pure synthetic phosphatidylcholines, this condition is rarely a

problem because such processes are usually rapid enough to

be effectively free of kinetic limitations even at moderate scan

rates. However, processes such as the pretransition and the

subtransition of synthetic saturated phosphatidylcholines are

known to be kinetically limited at all temperatures and scan rates

at which calorimetric observation is feasible (16–18). For such

processes, their thermodynamic parameters cannot be reliably

measured by DSC.

Another aspect of the thermodynamic equilibrium problem

that should be considered is the question of whether the system

is at equilibrium before the calorimetric scan is initiated. In

many DSC studies of model and biological membranes, the

sample is placed in the calorimeter and cooled fairly rapidly

to a low temperature, and a calorimetric heating scan then is

begun relatively quickly. Because, as was mentioned above,

the kinetics of lipid phase transitions in complex systems are

not well studied and because the rates of reversible lipid phase

transitions are generally considerably slower when proceeding

from a higher-temperature to a lower-temperature state than the

reverse, the possibility exists that the system under study may

not be at thermodynamic equilibrium when the calorimetric run

is begun. This possibility can be the case even if no exothermic

events are observed during heating. For this reason, it is always

advisable to cool the sample slowly and to investigate the effect

of variations in the "annealing" time at low temperatures on the

DSC results obtained.

We stress here that although DSC is in principle a rela

tively straightforward physical technique, its theoretical ther

modynamical and kinetic basis is not trivial but should be well

understood as it applies to equilibrium and nonequilibrium ther

motropic lipid phase transitions of various types and to either

heat conduction or power compensation instruments. Moreover,
some care must be taken in sample preparation, selection of

sample size, and sample equilibration before data acquisition;

in the choice of suitable scan rates, starting temperatures, and

ending temperatures during data acquisition; and in the anal

ysis and interpretation of the DSC thermograms obtained. An

adequate treatment of these issues is not possible in this brief

section.
induced rotational excitations of the hydrocarbon chains. Thus, the DPPC molecules are progressively overcome by thermally interchain interactions that favor a crystalline-like packing and its motion is also severely restricted. The polar headgroup contains only a few bound water molecules, rotation about their long axes is very severely restricted. The hydrocarbon chains that are tilted slightly with respect to the interfacial area. Thus, the Lβ phase is less tightly packed and much more strongly hydrated than is the Lγ phase. The Lβ′ phase of DPPC forms very slowly when cooling to low temperatures and requires about 3.5 days at 0°C to fully form; if a heating run on a DPPC sample not annealed at low temperature is performed, no subtransition will be detected, which indicates the importance of lipid samples being at thermal equilibrium before analysis by DSC. In fact, the DPPC subtransition was discovered by Stutevant and coworkers (20) when a DPPC sample was inadvertently cooled and left over the weekend before a DSC heating scan was initiated. A dilation increases in temperature result in a marked increase in the long-axis rotational rates of the hydrocarbon chains, and at the pretransition temperature, the Lγ phase converts to the so-called ripple or Pβ phase. In the Pβ phase, the extended hydrocarbon chains seem to remain tilted with respect to the normal to the local bilayer plane but behave as if they are packed in a distorted orthorhombic lattice, and undergo relatively slow, restricted rotational motion about their long axes. The polar headgroup now contains about 15–18 waters of hydration and exhibits slow, hindered rotation on the NMR timescale.

### Table 1: Thermodynamic characteristics of the three phase transitions exhibited by multilamellar aqueous suspensions of dipalmitoylphosphatidylcholine after annealing (–3–5 days) at 0–4°C before heating

<table>
<thead>
<tr>
<th>Transition type</th>
<th>T(°C)</th>
<th>ΔT1/2(°C)</th>
<th>ΔHΔ(k cal/mol)</th>
<th>ΔS(cal/k mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtransition</td>
<td>16.5*</td>
<td>3.0*</td>
<td>3.5</td>
<td>11.7</td>
</tr>
<tr>
<td>Pretransition</td>
<td>34.2</td>
<td>1.8*</td>
<td>1.1</td>
<td>3.6</td>
</tr>
<tr>
<td>Main Transition</td>
<td>41.4</td>
<td>0.1</td>
<td>7.8</td>
<td>24.8</td>
</tr>
</tbody>
</table>

*The phase transition temperature of the subtransition and the ΔT1/2 values of the subtransition and pretransition are overestimated because of kinetic limitations, even at the lowest heating scan rates feasible. The data listed were obtained from Reference 41.

### Figure 2: A typical high-sensitivity DSC heating thermogram of a multilamellar, aqueous suspension of DPPC which has been annealed at 0–4°C for 3–5 days prior to commencement of heating. The subtransition, pretransition and main phase transition temperatures are denoted by Ts, Tp and Tm respectively.
are rotationally symmetric, packing into a hexagonal lattice. The cross-section areas of the hydrocarbon chain thus show a small increase at the pretransition temperature. The interfacial area increases much more substantially, however, because of the displacement of each lipid molecule along its long axis with respect to its neighbor. The increased area occupied by the polar headgroups allows them to rotate almost freely, although the degree of hydration does not seem to change. In contrast to the Lc and Lp phases, the bilayer is no longer planar but exists as a series of periodic, quasi-lamellar segments.

With increasing temperature, the formation of gauche rotational conformers in the hydrocarbon chain becomes increasingly favorable until at the gel-to-liquid–crystalline phase transition temperature, chain melting occurs. Spectroscopic and thermodynamic studies have shown that the hydrocarbon chains of DPPC in the melted or Lp phase contain about four gauche bonds per chain, mostly, but not entirely, in the form of kink (gauche-trans-gauche) sequences. As the melting of the hydrocarbon chains produces a marked increase in cross-section area and effectively shortens the length of the chains, the bilayer expands laterally and thins at the main phase transition. Although the hydrocarbon chains exhibit rapid flexing and rotation in the Lc phase, they are on average oriented normally to the bilayer plane and pack in a loose hexagonal lattice. This increase in the cross-section area per molecule results in an increase in the area available to the polar headgroup, with the result that rotational motion becomes fast on the NMR timescale, and the hydration at the bilayer interface increases, in part because of the partial exposure of more deeply located polar residues, such as the carbonyl oxygens of the fatty acyl chains, to the aqueous phase.

The pattern of thermostropic phase behavior exhibited by an aqueous dispersion of any lipid molecular species will vary considerably, depending on the length and structure of the hydrocarbon chains, the structure and charge of the polar headgroup, the nature of the linker (ester or ether) of the hydrocarbon chains to the glycerol or sphingosine backbone, and other chemical features of the lipid under study. Also, the degree of hydration and the pH and ionic composition of the aqueous phase can affect lipid thermostropic phase behavior profoundly. However, even a cursory discussion of this topic is beyond the scope of this article, and the reader is referred to recent reviews for more detailed information (16–18).

Phospholipid mixtures

Although studies of the thermostropic phase behavior of single-component multilamellar phospholipid vesicles are necessary and valuable, these systems are not realistic models for biological membranes that normally contain at least several different types of phospholipids and a variety of fatty acyl chains. As a first step toward understanding the interactions of both the polar and apolar portions of different lipids in mixtures, DSC studies of various binary and ternary phospholipid systems have been carried out. Phase diagrams can be constructed by specifying the onset and completion temperatures for the phase transition of a series of mixtures and by an inspection of the shapes of the calorimetric traces. A comparison of the observed transition curves with the theoretical curves supports a literal interpretation of the phase diagrams obtained by DSC. For a summary of the first high-sensitivity DSC studies of binary phospholipid–cholesterol mixtures and a description of how phase diagrams can be constructed from DSC data, the reader is referred to an early review by Mabrey and Sturtevant (21); for a compilation of the results of later DSC and other studies on other phospholipid mixtures, the reader is referred to Marnas (22).

The effect of cholesterol

The occurrence of cholesterol and related sterols in the membranes of eukaryotic cells has prompted many investigations of the effect of cholesterol on the thermotropic phase behavior of phospholipids (see References 23–25). Studies using calorimetric and other physical techniques have established that cholesterol can have profound effects on the physical properties of phospholipid bilayers and plays an important role in controlling the fluidity of biological membranes. Cholesterol induces an "intermediate state" in phospholipid molecules with which it interacts and, thus, increases the fluidity of the hydrocarbon chains below and decreases the fluidity above the gel-to-liquid–crystalline phase transition temperature. The reader should consult some recent reviews for a more detailed treatment of cholesterol incorporation on the structure and organization of lipid bilayers (23–25).

Recent high-sensitivity DSC studies of cholesterol–DPPC interactions have revealed a complex picture of cholesterol/DPPC interactions (26). At cholesterol concentrations from 0 to 20–25 mol %, the DSC endotherm consists of two components (see Fig. 3). The sharp component exhibits a phase transition temperature and cooperativity only slightly reduced from those of the pure phospholipid, and the enthalpy of this component decreases linearly with increasing cholesterol content, becoming zero at 20–25 mol %. In contrast, the broad component exhibits a progressively increasing phase transition temperature and enthalpy with a progressively decreasing cooperativity over this same range of cholesterol content. A bove cholesterol levels of 20–25 mol %, the broad component becomes progressively less cooperative, the phase transition midpoint temperature continues to increase, and the transition enthalpy continues to decrease, eventually approaching zero only at cholesterol concentrations near 50 mol %. These results suggest that at low cholesterol concentrations, cholesterol-poor and cholesterol-rich domains coexist, with the former decreasing in proportion to the latter as cholesterol concentrations increase. In fact, a cardinal point in the cholesterol/DPPC phase diagram at about 22 mol % had been predicted from the earlier model-building studies, which calculated that the cholesterol molecule could interact with a maximum of 7 adjacent phospholipid hydrocarbon chains (or 3.5 phospholipid molecules) and thus that free phospholipid would exist only at cholesterol concentrations below this value. This model also explains the decreasing enthalpy of the broad component observed above 22 mol % cholesterol because an increasing proportion of phospholipid molecules would interact with more than one cholesterol molecule rather than with the more flexible hydrocarbon chains of adjacent phospholipids and, thus, progressively decrease and eventually abolish the cooperative chain-melting phase transition.
The effect of cholesterol on the thermotropic behavior of PC bilayer also varies significantly with the structure, particularly the degree of unsaturation, of the hydrocarbon chains, and with more highly unsaturated PCs exhibiting a reduced miscibility with cholesterol and other sterols. Moreover, the structure of the lipid polar headgroup is also important in determining the effect of cholesterol on the host lipid, as is the structure of the sterol molecule itself. For more information on the application of DSC to the biologically important area of lipid–sterol interactions, the reader is referred to recent reviews (23–25).

The effect of small molecules

Several lipid-soluble small molecules, including drugs like tranquilizers, antidepressants, narcotics, and anaesthetics, produce biological effects in living cells. Although some of these compounds are known to produce their characteristic effects by interacting with specific membrane proteins, others seem to interact rather nonspecifically with the lipid bilayer of many biological membranes. The effect on the gel-to-liquid-crystalline phase transition profile of synthetic PCs of over 100 hydrophobic small molecules that produce biological effects has been studied by DSC (27). At least four different types of modified transition profiles can be distinguished: In so-called type C profiles, the addition of the additive shifts T_m usually (but not always) to a lower temperature while having little or no effect on the cooperativity ($\Delta T_{1/2}$) or $\Delta H_{cal}$ of the transition; other physical evidence suggests that additives that produce this behavior are usually localized in the central region of the bilayer, which interacts primarily with the C9–C16 methylene region of the phospholipid hydrocarbon chains. Type A profiles are characterized by a shift in $T_m$, usually to a lower temperature, an increase in $\Delta T_{1/2}$, and a relatively unaffected $\Delta H_{cal}$ during the addition of the appropriate small molecules; these additives seem to be partially buried in the hydrocarbon core of the bilayer, which interacts primarily with the C2–C8 methylene region of the hydrocarbon chains. In type B profiles, a shoulder emerges on the main transition, the area of which increases in conjunction with a corresponding decrease in the area of the original peak as the concentration of additive increases. The total area of both peaks is relatively unchanged, at least at low additive concentrations. Additives that produce type B profiles generally reside at the hydrophobic–hydrophilic interface of the bilayer and interact primarily with the glycerol backbone of the phospholipid molecules. Finally, type D profiles exhibit a discrete new peak that grows in area at the expense of the parent peak as the additive concentration increases; normally, however, the final $\Delta H_{cal}$ and $\Delta T_{1/2}$ values of the new and
original peaks are not greatly different. Type D additives usually seem to be located at the bilayer surface and interact with the phospholipid headgroup. Although this classification is useful, not all small molecules produce one of these four types of DSC profiles. Whether a consistent relationship exists between the type of transition profile produced by a small molecule and its physiological effects remains to be determined.

The effect of transmembrane peptides

DSC has been used to great effect to study the effect of the incorporation of α-helical transmembrane peptides on the thermotropic phase behavior of various phospholipid bilayers. Because most integral membrane proteins contain one or more α-helical transmembrane segments, such studies are relevant to the mechanisms by which the physical properties of the membrane lipid bilayer modulate the structure and activity of such proteins. In this regard, several investigations have been carried out using DSC and many other physical techniques to understand how the presence of transmembrane peptides affects the organization and dynamics of the host lipid bilayer and vice versa. Such studies have examined the effects of systematic variations in the length and structure of model α-helical transmembrane peptides on lipid bilayer organization and dynamics, and how the effects of such peptides are themselves affected by the hydrophobic thickness and chemical composition of the host phospholipid bilayer. These important studies are ongoing, and the reader should consult recent reviews for more information.

The effect of membrane antimicrobial peptides

DSC has also been used to study the effects of a wide variety of antimicrobial peptides on the thermotropic phase behavior of different lipid bilayers. These studies again are highly biologically relevant because the primary mode of action of most antimicrobial peptides is the perturbation and permeabilization of the lipid bilayers of the target membrane, and these agents have considerable promise as antibiotics, especially to treat multiple drug-resistant pathogenic bacteria. Again, the reader should consult recent reviews for more information on this topic.

The effect of membrane proteins

Because of their obvious relevance to biological membranes, the effect of several peptides and proteins on the thermotropic phase behavior of single synthetic phospholipids or phospholipid mixtures has been studied by many groups (see 16, 17). It was originally proposed by Papahadjopoulos et al. (32) that polypeptides and proteins could be considered as belonging to one of three types according to their characteristic effects on phospholipid gel-to-liquid–crystalline phase transitions. Type 1 proteins typically produce no change or a modest increase in Tm, a slight increase or no change in ΔT1/2, and an appreciable and progressive increase in ΔHcal as the amount of protein added is increased. These proteins normally do not expand phospholipid monolayers nor alter the permeability of phospholipid vesicles into which they are incorporated. Type 1 proteins are “hydrophilic” proteins that are thought to interact with the phospholipid bilayer exclusively by electrostatic forces and, as such, normally show stronger effects on the phase transitions of charged rather than zwitterionic phospholipids. Type 2 proteins produce a decrease in Tm, an increase in ΔT1/2, and a considerable and progressive decrease in ΔHcal; phospholipid monolayers are typically expanded by such proteins, and these proteins normally increase the permeability of phospholipid vesicles. These proteins, which are also hydrophilic, are believed to interact with phospholipid bilayers by a combination of electrostatic and hydrophobic forces, initially adsorbing to the charged polar headgroups of the phospholipids and subsequently partially penetrating the hydrophilic–hydrophobic interface of the bilayer to interact with a portion of the hydrocarbon chains. Finally, type 3 proteins usually have little effect on the Tm or ΔT1/2 of the phospholipid phase transition, but ΔHcal decreases linearly with protein concentration. Type 3 proteins are “hydrophobic” proteins that markedly expand phospholipid monolayers and increase the permeability of phospholipid vesicles. These proteins are thought to penetrate deeply into or to span the hydrophobic core of anionic or zwitterionic lipid bilayers and, thus, to interact strongly with the phospholipid fatty acyl chains and essentially to remove them from participation in the cooperative chain-melting transition. It should be noted, however, that some type 3 proteins may also interact electrostatically with phospholipid polar headgroups, particularly with those bearing a net negative charge.

The results of more recent DSC and other studies of lipid-protein model membranes clearly indicate that the classification scheme originally proposed is not completely appropriate for naturally occurring membrane proteins (see Reference 17). Thus, none of the water-soluble, peripheral membrane-associated proteins studied thus far exhibit classical type 1 behavior (no change or a modest increase in Tm, a slight increase in ΔT1/2, and an increase in ΔH of the chain-melting transition). Therefore, it seems doubtful whether natural membrane proteins ever interact with phospholipid bilayers exclusively by electrostatic interactions. However, a few examples of membrane proteins do exhibit more-or-less-classical type 2 behavior. These examples include the myelin basic protein and cytochrome c, all of which usually reduce the Tm, increase the ΔT1/2, and substantially reduce the ΔH of the chain-melting transition of anionic phospholipids. Strictly speaking, few if any membrane proteins actually exhibit classical type behavior as originally defined (no change in the Tm or ΔT1/2 and a progressive linear reduction in the ΔH of both neutral and anionic phospholipid phase transitions with increasing protein concentration). This is because, with the advent of high-sensitivity calorimeters and the availability of pure phospholipids, it has become clear that all integral membrane proteins reduce the cooperativity of gel-to-liquid-crystalline phase transitions, as indeed would be expected from basic thermodynamic principles. Moreover, some type 3 proteins exhibit a nonlinear decrease in ΔH with changes in protein levels, whereas others can produce at least moderate shifts in the Tm of phospholipid phase transitions. However, if we relax the original type 3 criteria...
somewhat, then several integral, transmembrane proteins can be said to exhibit "modified" type 3 behavior.

The classification scheme of Papahadjopoulos et al. (32), appropriately modified for type 3 proteins, is still of some use in studies of lipid–protein interactions, although some proteins, at least under certain conditions, do not fall neatly into any of these three categories. It seems that all naturally occurring membrane proteins studied to date interact with lipid bilayers by both hydrophobic and electrostatic interactions and that different membrane proteins differ only in the specific types and relative magnitudes of these two general classes of interactions. It is also clear that the behavior exhibited by any particular membrane protein can depend on its conformation, method of reconstitution, and relative concentration, as well as on the polar headgroup and fatty acid composition of the lipid bilayer with which it is interacting (see Reference 17).

Although DSC and other physical techniques have made considerable contributions to the elucidation of the nature of lipid–protein interactions, several outstanding questions remain. For example, it remains to be definitively determined whether some integral, transmembrane proteins completely abolish the cooperative gel-to-liquid–crystalline phase transition of lipids with which they are in direct contact or whether only a partial abolition of this transition occurs, as is suggested by the studies of the interactions of the model transmembrane peptides with phospholipid bilayers (see above). The mechanism by which some integral, transmembrane proteins perturb the phase behavior of very large numbers of phospholipids also remains to be determined. Finally, the molecular basis of the complex and unusual behavior of proteins such as the concanavalin A receptor and the Acholeplasma laidlawii B ATPase is still obscure (see Reference 17).
strong upward shift in the Lα phase transition temperature that this membrane lipid strongly stabilizes the Lα phase and destabilizes the H2 phase of DEPE, indicating that it has a conical shape with a large polar headgroup volume relative to the volume of the hydrocarbon chains. In fact, this lipid forms a normal micellar phase in water in isolation from other membrane lipids (31).

A. laidlawii is notable in this regard is the phase preference of glycerylphosphoryl(diacylglycerol, which strongly destabilizes the reversed hexagonal phase and which actually prefers the normal micellar phase in isolation (36) (see Fig. 4). The presence of normal, lamellar, and reversed phase-prefering lipids in a single membrane has important implications for understanding the physical basis of lipid organization and biosynthetic regulation in this organism and possibly in other organisms. We also showed that the characteristic effect of the individual A. laidlawii membrane lipids on the lamellar/reversed hexagonal phase transition temperature of the phosphatidylethanolamine matrix is not well correlated with their polar headgroup intrinsic volumes. This result indicates that the effective cross-section area of the polar headgroups of these lipid species must be strongly influenced by factors such as charge, hydration, orientation, and motional freedom as well as by intrinsic headgroup volume.

The approach discussed above to determine quantitatively the effect of various membrane phospho- and glycolipids on the lamellar/nonlamellar phase behavior of a host phospholipid/ethanolamine or similar matrix has been applied to determine the relative shape, and thus the effect on the monolayer curvature of the host bilayer, of several agents, including sterols, peptides, detergents, and drugs (see Reference 14). Such studies can be very useful in providing insight into the function and mechanism of action of these agents on biological membranes.

DSC Studies of Biological Membranes

The A. laidlawii membrane was used by Stein and colleagues (37) to show for the first time that biological membranes can undergo a gel-to-liquid-crystalline lipid phase transition similar to that previously reported for lamellar phospholipid-water systems. These workers demonstrated that when whole cells or isolated membranes are analyzed by DSC, two relatively broad endothermic transitions are observed on the initial heating scan. The lower-temperature transition is fully reversible, varies markedly in position with changes in the length and degree of unsaturation of the membrane lipid fatty acyl chains, is broadened and eventually abolished by cholesterol incorporation, and exhibits a transition enthalpy characteristic of the mixed-acid synthetic phospholipids. Moreover, an endothermic transition that has essentially identical properties is observed for the protein-free total membrane lipid extract dispersed in excess water or aqueous buffer, which indicates that the presence of membrane proteins has little effect on the thermotropic phase behavior of most membrane lipids. In contrast, the higher-temperature transition is irreversible, is independent on membrane lipid fatty acid composition or cholesterol content, and is absent in total membrane lipid extracts, which indicates that the higher temperature transition results from an irreversible thermal denaturation of the membrane proteins. A comparison of the enthalpies of transition of the lipids in the membrane and in water dispersions indicates that at least 75% of the total membrane lipids participate in this transition. Evidence was also presented that the lipids must be predominantly in the fluid state to support normal growth. These results were later confirmed and extended by Reinert and Stein (38) and by Melchior et al. (39), who showed that the gel-to-liquid-crystalline lipid phase transition is a property of living cells and that about 85-90% of the lipid participates in the gel-to-liquid-crystalline phase transition. These studies provided strong, direct experimental evidence for the hypothesis that lipids are organized as a liquid-crystalline bilayer in biological membranes, a basic feature of the currently well-accepted fluid-mosaic model of membrane structure.

Representative high-sensitivity DSC initial heating scans of viable cells, isolated membranes, and total membrane lipid dispersions are shown in Fig. 5; in this instance, cells, membranes, and lipids were made nearly homogenous in elaidic acid (13). The fully reversible gel-to-liquid-crystalline lipid phase transitions observed in cells and membranes essentially have identical phase transition temperatures, enthalpies, and degrees of cooperativity, which suggests that membrane lipid organization in these two samples is very similar or identical. In contrast, the midpoint of the chain-melting transition of the membrane lipid dispersion is shifted to a higher temperature, exhibits a greater enthalpy, and is considerably less cooperative than in cells or membranes, which suggests that native membrane lipid organization has been perturbed during extraction and resuspension of the membrane lipids in water. The thermal denaturation of the proteins in the cells and membranes has absolutely no effect on the peak temperature or cooperativity of the lipid phase transition. However, about 15% of the lipids do not participate in the
membrane lipid transition centered near the rat liver inner mitochondrial membrane reported a narrower membrane transition (44). However, a more recent study of these results, except that the outer membrane transition seemed chondria and of isolated inner and outer membranes confirmed and in extracted lipids (42). A later study of both intact mito-
phase transition centered at 0°C. DSC and other techniques. The earliest work with whole mi-
are low in cholesterol, have been studied by several groups using higher cooperativity than does the membrane lipid dispersion and membranes exhibits a similar temperature maximum and a higher cooperativity than does the membrane lipid dispersion favors the former interpretation.

The presence of high levels of cholesterol in many eukaryotic membranes, particularly plasma membranes, abolishes a discrete cooperative gel-to-liquid–crystalline membrane lipid phase transition in these systems. Thus, no lipid phase transitions could be detected by DSC in the cholesterol-rich erythrocyte (40) or myelin (41) membranes. The thermotropic behavior of rat liver microsomal membranes, which contain moderate levels of cholesterol, has been studied by DSC. An early study using conventional DSC revealed a single reversible, broad phase transition occurring between ~15°C and ~5°C in both intact membranes and isolated lipids (42). A more recent high-sensitivity DSC study confirmed the absence of a reversible phase transition above 0°C (43). Rat liver mitochondrial membranes, which are low in cholesterol, have been studied by several groups using DSC and other techniques. The earliest work with whole mi-
tochondrial revealed a reversible broad gel-to-liquid–crystalline phase transition centered at 0°C in mitochondrial membranes and in extracted lipids (42). A later study of both intact mito-
chondria and of isolated inner and outer membranes confirmed these results, except that the outer membrane transition seemed to occur at a slightly lower temperature than did the inner membrane transition (44). However, a more recent study of the rat liver inner mitochondrial membrane reported a narrower membrane lipid transition centered near ~10°C; by artificially increasing cholesterol content some 10-fold to about 30 mol %, the inner membrane gel-to-liquid–crystalline phase transition could be lowered and broadened, and its AH° transition was reduced to less than one tenth that of the native membrane (45). It has also been reported that in beef heart mitochondrial inner membranes, a broad reversible endothermic phase transition centered at ~10°C occurs.

DSC has been used to study the individual protein com-
ponents of biological membranes of relatively simply protein composition and the interaction of several of these components with lipids and with other proteins. The red blood cell mem-
brane, which has been most intensively studied, exhibits five discrete protein transitions, each of which has been assigned to a specific membrane protein. The response of each of these ther-
mal transitions to variations in temperature and pH as well as to treatment with proteases, phospholipases, specific labelling reagents, and modifiers and inhibitors of selected membrane activities, has provided much useful information on the interactions and functions of these components in the intact erythrocyte membrane (46–49). Similar approaches have been applied to the bovine rod outer segment membrane (50) and to the spinach chloroplast thylakoid membrane (51).

Acknowledgments
Research performed in the authors’ laboratory has been sup-
ported by operating and major equipment grants from the Cana-
dian Institutes of Health Research, operating grants from the Natural Sciences and Engineering Research Council of Canada, and by major equipment and personnel support grants from the Alberta Heritage Foundation for Medical Research.

References


Lipids are hydrophobic or amphiphilic low-molecular-weight substances with a low solubility in water. Glycerophospholipids, sphingolipids, and cholesterol are the building blocks of cellular membranes, and triacylglycerols are the major molecular storage forms of metabolic energy. Various lipids serve as signaling substances either as first or second messengers in signal transduction; as autocrine, paracrine, or endocrine regulators; or are covalently bound to cellular proteins. Several diseases are caused by, or at least associated with, alterations in lipid metabolism, such as inherited disorders of lipid metabolism, atherosclerosis, diabetes, obesity, or Alzheimer’s disease.

Lipidomics is the determination of a lipid profile of a given source under given conditions with the aim to understand lipid function in a biologic system. This understanding might be achieved by correlating changes in this composition with physiologic or pathophysiologic alterations of the system. Current research focuses on the determination of lipid profiles, which is attempted predominantly by the combination of extraction and separation techniques coupled with mass spectrometry. A major challenge is the variety of molecular lipid species found in a given cell. This article provides an overview on current methodologic approaches that find application in lipidomics. Because the function of a lipid can be associated critically with its distinct subcellular or even suborganellar localization, some limitations of this approach are discussed also.

Lipidomics (1) is the approach to determine and to understand the lipid profile of a given biologic source in terms of systems biology (2). This “understanding” requires information about the interacting partners. Systems biology requires a comprehensive set of quantitative data that can be interpreted by bioinformatic approaches to obtain insight into structure and dynamics of a given system. Current attempts to treat the hydrophobic part of the metabolome in this respect face technological problems that must be overcome to determine lipid profiles comprehensively and accurately. With the advent of modern mass spectrometric techniques, metabolic snapshots of parts of the lipidome have become possible. They can be compared with corresponding snapshots of other genotypes and with those obtained under pathologic conditions or after drug treatment. Such differences can be visualized in the form of a network model to correlate compositional differences with functional aspects. However, the information on the levels of all lipid classes in a relevant system, such as an organ, tissue, or cell type, is often incomplete or even entirely missing. This incomplete system is, in part, caused by the large heterogeneity found among cellular lipids, and also by the structural complexity especially of glycolipids, the unavailability of appropriate standard substances, and difficulties with the extraction and ionization of certain lipid classes such as the phosphatidylinositol phosphates. Moreover, decades of lipid research taught us that not only is the concentration of a lipid critical, but also its subcellular distribution can be critical for its function (3). This article provides a brief overview on current techniques used within lipidomics.

Lipids

Lipids are a structurally heterogeneous group of low molecular compounds with common solubility properties. The majority of lipids is only slightly soluble in water, but they can be extracted from biologic sources with organic solvents. Lipids can be defined as hydrophobic or amphiphilic small molecules that originate entirely or in part by carbanion-based condensation of thioesters such as fatty acids, polyketides, and their derivatives, or by carbenium ion-based condensation of isoprene units such as the terpenes, which include the steroids. A recently suggested nomenclature and assignment system for lipids will facilitate the data exchange and processing required in lipidomics (4). It consists of a 12-digit code for each molecular lipid species and classifies lipids into eight categories: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, and polyketides (4). The first two letters of the alphanumeric Lipid-ID are a fixed database designation, the next two letters specify the category, and the numbers specify class, subclass, and the individual lipid. In this way, also more hydrophilic lipid derivatives can be considered within the field of lipidomics. The structures of selected lipids are shown in Fig. 1.
Lipid aggregates

Lipids differ not only in their chemical structures but also in the types of aggregates that they form in the aqueous environment of a living cell. Highly hydrophobic lipids are fats and oils, which are used as storage forms of metabolic energy. In an aqueous environment, these substances aggregate to lipid globules. Within cells, these lipids are deposited as insoluble droplets of up to 200 \( \mu m \) in size especially in the cytoplasm of adipocytes. Closer examination of these aggregates found in vivo reveals a highly metabolically active and organized complex-layered structure of an own organelle, in which triacylglycerols, diacylglycerols, cholesterol, and cholesterol esters are surrounded by a phospholipid/lysophospholipid monolayer, together with accessory proteins (5). In addition, amphiphilic lipids contain structural elements of higher polarity. Dependent on their molecular shape, these lipids form different supramolecular aggregates in an aqueous environment. The most important feature of amphiphilic lipids, with a cylindrical shape like many glycerophospholipids and sphingolipids, is their ability to form lamellar phases spontaneously, the structural basis of biologic membranes. Cone-shaped lipids with small hydrophilic headgroups such as phosphatidylethanolamine (PE), phosphatidic acid (PA),...
by flippases and floppases, but it can also be destroyed actively. The asymmetric distribution is maintained actively. Phosphates are localized predominantly on the cytoplasmic leaflet (PS), the PE, and also the phosphatidylinositols and their (ceramide-1-phosphorylcholine; SM) are found predominantly in the apical and basolateral membranes of a polarized cell. Although glycosphingolipids (with glucosylceramide as an exception) are found exclusively on the extracellular leaflet, phosphatidylcholine (PC) and sphingomyelin (SM) can show the coexistence of different lipid phases at a given temperature. This differential miscibility can lead to lipid microdomains that might also be found in biologic membranes under physiologic conditions. A crucial role for membrane properties is played by the sterols, cholesteryl in animals, stigmasterol in plants, and ergosterol in lower eukaryotes. Within evolution, they became available after the occurrence of higher oxygen concentrations in the ancient atmosphere (see Mourizen (7)) and enabled eukaryotic cells to form the so-called liquid ordered membrane phases. This liquid is essential for membrane stability, impermeability, and for the function of many membrane proteins. Within lipidomics, the analysis of membrane-forming lipids, or "structural" lipids, such as phospholipids and cholesterol, can be distinguished from "mediator-lipidomics" (8). In the latter, highly bioactive lipids are determined, which are present usually in very low concentrations but are found in greater amounts in response to extracellular stimuli. Lipid mediators are substances such as the eicosanoids, platelet activating factor, LPA, sphingosine-1-phosphate, and ceramide-1-phosphate. Because "structural" lipids determine the function of membranes and can be altered under pathophysiologic conditions, both lipid classes have to be determined in a comprehensive approach.

Membranes

Biologic membranes serve several vital functions such as compartmentalization, maintenance of gradients, signal transduction, and many others. Depending on their source, membranes differ in their total lipid composition, but they also show lateral heterogeneity as well as differences between their two leaflets. Although glycosphingolipids (with glucosylceramide as an exception) are found exclusively on the extracellular leaflet, phosphatidylcholine (PC) and sphingomyelin (ceramide-1-phosphorylcholine; SM) are found predominantly in the apical leaflet, and the aminophospholipids phosphatidylethanolamine (PE), the PE, and also the phosphatidylinositol and their phosphates are localized predominantly on the cytoplasmic leaflet (6). This asymmetric distribution is maintained actively by flipases and floppases, but it can also be destroyed actively by scramblases in response to signals. For example, surface exposure of PS is a sign of pathologically altered cells (10). Lateral differences in lipid compositions cause microdomains to develop that account for differences in bilayer thickness and, subsequently, protein transport. Although it is still a matter of debate how these lipids are organized laterally, on which time scale the resulting microdomains ("rafts") exist and with which functional relevance (11), lipid compositions differ not only between such short-living microdomains, but also between the apical and basolateral membranes of a polarized cell.

Although information on the spatial distribution of membrane lipids can be important (3), most mass-spectrometric approaches used in lipidomics focus currently on the determination of total lipid composition. An advanced mass-spectrometric surface analysis technique, however, secondary ion mass spectrometry (SIMS) determines lipid compositions within different areas of freeze-fractured cells (12). This technique and related imaging techniques will become important tools in lipidomics. Different local concentrations of phosphoinositides are one of the many examples of the involvement of lipids in signal transduction processes (13). Lipid mediators can act not only as membrane components, but also as autocrine and paracrine mediators (14); when bound to proteins, lipids fulfill crucial functions. Starting with the pioneering work of Rudolf Schoenheimer on cholesterol, it has been recognized that lipid metabolism and transport are tightly controlled processes. More recently, the molecular mechanisms that underlie lipid homeostasis have been recognized (15). Much of the interest in lipidomics results from new possibilities for the understanding, diagnosis, and treatment of human diseases caused by disturbances of this homeostasis. Therefore, determination and interpretation of lipid profiles is the aim of lipidomics. A key discipline within lipidomics is lipid analysis. Mass-spectrometric methods, however, allow not only the determination of lipid classes that share a common head group, such as PC or SM, but also the distribution of molecular species that develop by combination of different acyl chains, which can be determined with high sensitivity.

Lipid Analysis

Lipid analysis starts with sample preparation, followed by extraction of the lipids from their biologic matrices. The resulting crude lipid extracts can be subjected directly to extensive and global mass spectrometric (MS) analysis, which is an approach sometimes called " shotgun lipidomics." Alternatively, lipids or lipid classes are separated prior to analysis, either by gas chromatography (GC), liquid chromatography (LC), thin layer chromatography (TLC), or—in less frequently—capillary electrophoresis (CE). These methods are coupled online or offline to mass spectrometry or tandem mass spectrometry (M/S/M; e.g., GC/MS, LC/MS, LC/MS/MS, CE/MS). In both cases, structural heterogeneity and complexity of lipids provide a huge amount of data. Accordingly, bioinformatic treatment of the data is used increasingly to facilitate data handling and evaluation. Sample preparation and lipid extraction

In many cases, sample preparation is mandatory prior to lipid extraction. This crucial process might include tissue homogenization, determination of dry weights, cell numbers, protein content, or DNA content for normalization purposes. The addition of internal standards is not only used to control extraction efficiency, but also required for lipid quantification by mass spectrometric analysis, in which an ion current is translated into a lipid concentration. Many standard lipids that can be distinguished from endogenously occurring lipids by using rare
Ionization (APCI), can be less than 30% in classic Folch and Bligh Dyer extraction, recovery of the acidic lipids PA and phosphatidylglycerol (PG) and must be validated to ensure reliable results. For example, the time, and temperature influence the recovery of extracted lipids lipid alteration or decomposition (16).

Because of the heterogeneous nature of lipids, no single extraction method exists that extracts all lipids efficiently. Lipid extraction can be achieved by liquid-liquid extraction, Soxhlet extraction, or solid-phase extraction. In 1957, Folch developed a popular method by which many lipids are extracted using a mixture of chloroform and methanol in a volume ratio of 2:1. The extraction is followed by a wash step with water and is frequently applied for total lipid determination. The occasional formation of emulsions is a disadvantage, and more hydrophilic lipids like gangliosides are separated into the upper phase. To date, most extractions are performed by a method introduced by Bligh and Dyer in 1959 or by modifications of this method. In this case, the extraction is carried out by a monophasic mixture of chloroform, methanol, and water in a volume ratio of 1:2:5. Changing the ratio to 2:2:1.8 (V/V/V) leads to a phase separation, in which most lipids can be recovered from the lower phase. Proteins precipitate mostly between the two phases. It must be considered that quantitative lipid extraction by either of these methods might require repeated re-extraction, and that lipid exposure to air or enzyme sources might lead to lipid alteration or decomposition (16).

Extraction parameters such as solvent type, mixture ratios, metal ion concentration, pH of the aqueous phase, extraction time, and temperature influence the recovery of extracted lipids and must be validated to ensure reliable results. For example, the recovery of the acidic lipids PA and phosphatidylglycerol (PG) can exceed 80% in both Folch and Bligh Dyer extraction, where these lipids can become bound to proteins tightly (17).

Lipids bound to proteins covalently are only released under appropriate conditions, which depend on the type of lipid-protein linkage. For example, ceramides bound to the cornified envelope in the human skin (18) can be extracted after mild alkaline hydrolysis of the ester linkage between lipid and protein.

Mass spectrometry

The dynamic development of mass spectrometry has had a huge impact on lipid analysis. Currently, a variety of suitable mass spectrometers is available. In principal, a mass spectrometer consists of an ion source, a mass analyzer, and an ion detector. The typical features of each instrument (Fig. 2) result mostly from the types of ion source and mass analyzer. To date, the ionization techniques applied to lipid analysis include Electrospray Ionization (ESI) or nano-ESI, Atmospheric Pressure Chemical Ionization (APCI), Matrix-Assisted Laser Desorption/Ionization (MALDI), and, more recently, Atmospheric Pressure Photo Ionization (APPI) and Desorption Electrospray Ionization (DESI).

For the majority of analytical tasks in lipidomics, ESI and nano-ESI are the most common choices. ESI uses a high cone voltage to produce single- or multiple-charged ions from an analyte-containing solution by creating a fine spray of highly charged droplets. In the case of high flow rates, this process can be assisted pneumatically by nitrogen as drying gas. ESI is a soft ionization method that can be run without fragmentation of the analyte molecules. It is suitable for membrane lipids (19). Generally, neutral lipids show poor ionization efficiencies in classic ESI, so that additional measures like adduct formation, derivatization, or application of nano-ESI are required. Typical drawbacks of ESI include signal suppression in complex matrices and varying adduct patterns.

Nano-ESI is a miniaturization of ESI with enhanced sensitivity and reduced signal suppression. It can be carried out either as classic offline approach analysis or as online nano-flow LC/MS, which are both attractive for lipidomics. Performance differences regarding the mass spectrometric response may result from different conductive coatings of the used sample capillaries and the used ionization mode. Many lipid classes, which include glycerophospholipids (20–22); sphingolipids (23, 24); triacylglycerols, cholesterol, and ergosterol (25); and cholesterol esters (26), have been analyzed using ESI and nano-ESI (20).

This approach has already been extended for automated (21) and high-throughput investigations (26).

APCI is complementary to ESI for compounds of limited polarity (27). In APCI, the analyte solution is subjected to a heater of 400–500 °C. The resulting plasma is ionized at a corona discharge needle with the help of nitrogen. The nitrogen molecules transfer the charge onto the analyte molecules in an indirect fashion. Because of the heat, the ionization conditions are less soft than ESI, and sensitive molecules may show fragmentation. The solvent plays a less dominant role for ionization than in ESI. This result can be beneficial in case of normal phase LC/MS approaches with organic solvents such as chloroform or hexane. Adduct formation can be reduced when suitable conditions are used. APCI has been used to analyze ceramides (28), sphingomyelin (29), triacylglycerols, as well as cholesterol and its oxidation products (30).

APPI is derived from APCI, but instead of the corona discharge needle, the ionization takes place after irradiation with a krypton lamp that emits photons of 10.0-10.6 electron volt. Different methods with and without dopant have been recommended. APPI, in combination with LC, has been used to analyze glycosphingolipids (31). In addition, APPI has been reported to be more sensitive and efficient than APCI and ESI for the analysis of fatty acid esters and acylglycerols (32, 33). Because the method has not found widespread use to date, it is too early to estimate its potential.

MALDI is another soft-ionization method, but in contrast to ESI and APCI, it is carried out in solid state. The analyte is placed onto a metallic plate together with an UV-absorbing matrix. Usually, the matrix consists of aromatic acids, which are cocrystallized with the sample. Irradiation of suitable positions on the target with a laser of appropriate wavelength causes analytic desorption and ionization. MALDI is used widely in the...
analysis of proteins and polymers. Recently, it has also been applied to the analysis of lipids and in particular glycolipids (34). MALDI is a qualitative method that normally provides no quantitative results. The matrix used, the heterogeneities within the matrix, analyte concentrations, and preparation techniques influence the results greatly. MALDI is mostly used offline. LC/MALDI coupling can be implemented via fraction collection and automated target preparation. Compared with ESI, MALDI is more tolerant to salts and other disturbing components. DESI is performed by directing electrosprayed-charged droplets onto a surface for analysis under atmospheric conditions. The collision of the charged droplets with the surface leads to the ionization and desorption of the analyte (35). Then, the ions produced in the gas phase are sampled by an atmospheric interfaced mass analyzer. DESI has been used to create two-dimensional images related to the distribution of lipid species in human tissues (36).

Mass analyzers differ in sensitivity, accuracy, and resolution. Among others, they include quadrupole, triple quadrupole (TQ), ion trap, linear ion trap, time of flight, Fourier transform ion cyclotron resonance, and hybrid combinations such as quadrupole-time of flight (Q-TOF) analyzers.

The deliberate generation of fragments is a valuable tool in structure elucidation and quantification. It is mainly carried out as a second mass analysis after determination of the mass per charge (m/z) ratio of the molecular ions. This approach is called tandem mass spectrometry and it can be carried out either as tandem-in-space using a triple quadrupole (or combinations of a quadrupole and another mass analyzer), or as tandem-in-time using an ion trap. The latter also allows multiple-stage mass spectrometry (MS^n). On the other hand, a triple quadrupole allows different MS/MS scan modes, namely product ion scan, precursor or parent ion scan, and neutral loss scan.

TQ mass spectrometers (Fig. 2) have been used frequently for lipid analysis. They consist of three quadrupoles, Q1, Q2, and Q3, which are located in a row, with Q2 as the fragmentation cell filled with an inert collision gas (e.g., argon). A product ion scan starts with selecting a lipid species of a certain m/z-value in Q1, fragmenting it through collisions with the collision gas (collision-induced dissociation; CID) in Q2, and determining the m/z values of the fragments in Q3. Usually, this scan type is used to ascertain characteristic fragments and fragmentation conditions such as the suitable CID energy. Parent (precursor) ion scan (PIS) is the conversion of the product scan. Here, Q3 is fixed to m/z values of desired fragments, whereas the corresponding parent molecular ions are scanned in Q1. PIS is used to identify lipid classes and individual lipid species according to characteristic charged fragments of their functional head groups or their backbones. Neutral loss scan (NL), finally, represents the loss of a neutral fragment from a charged parent molecule. In this case, both Q1 and Q3 are used to scan, but with a constant mass offset. NL has been frequently used to analyze a variety of lipids. For example, many phospholipids have been identified recently according to the NL of their fatty acids in the positive ionization mode. A structural identification
Lipidomics

Figure 3  Tandem electrospray mass spectrum (positive mode) of lactosylceramide, which shows the neutral loss of two hexose residues.

These scan modes or combinations of them have been used not only for identification, but also for quantification of lipids. In this context, multiple parent ion scan (MPIS) of head group and backbone (fatty acids) fragments has been shown recently to be suitable for the determination of individual phospholipid amounts relative to an internal standard. This analysis has been performed directly from crude lipid extracts by using a modified Q-TOF mass spectrometer (21). On the other hand, multiple reaction monitoring (MRM) is applied frequently for lipid quantification. MRM is a technique based on monitoring compound-specific transitions of precursor ions to product ions. In addition to specificity, sensitivity and selectivity are key advantages of MRM. This advantage is particularly valuable for the quantification of small compound amounts and/or by the occurrence of increased background levels. An interesting application of MRM is the quantification of many endogenous sphingolipids carried out by Zheng et al. (37). In a single LC run, multiple combinations of specific precursor-ion/product-ion transitions can be obtained, monitored, and applied to achieve sensitive quantification of individual lipid species.

As mentioned above, two different mass spectrometric approaches to lipid analysis exist: The first one is performed directly from lipid extracts without prior chromatographic separation and is referred to as "shotgun" lipidomics. Here, lipid classes are separated in the ion source according to their intrinsic electrical properties (24). Detailed and unambiguous structural and quantitative analysis of individual species is obtained by means of multiplexed mass spectrometry using NL, PIS, and combinations of them. A remarkable contribution to this field is the analysis of ceramides, cerebrosides (β-glycosyl ceramides), and phospholipids by Han. Refinements of this approach using advanced mass spectrometers have led to additional improvements.

On the other hand, the complexity of the lipid mixtures that compose a variety of isobaric compounds as well as the signal suppression effects during MS ionization frequently require a chromatographic separation. Therefore, LC/MS and LC/MS/MS are valuable tools. Zheng et al. (37) have developed a normal phase LC-ESI-MS/MS method that allows the separation of sphingolipid classes and the subsequent quantification of molecular sphingolipid species, which includes ceramides, glucosylceramides, lactosylceramides, and sphingomyelin in one run. Others have shown that a combination of MALDI-MS and ESI-MS/MS with thin layer chromatography (TLC) enables a successful analysis of gangliosides (38) and of low abundant phosphoinositides (39), respectively.

Bioinformatics

The amount of data obtained by mass spectrometry is enormous, particularly when many lipid classes or several lipid species are investigated simultaneously. As such, manual management of the different data sets is impractical. For this reason, various attempts have been initiated to create software that is capable of evaluating and handling the generated data sets in a qualitative, quantitative, and comparative manner. One example is the software called Lipid Profiler, which has been used to achieve automated identification, deconvolution, and absolute quantification of lipids. This software is capable of handling large data sets and can provide a comprehensive analysis of lipid species.
of glycerophospholipid species analyzed by MPIS (21). A further model is given by a collection of computer algorithms referred to as the computational lipid analysis program, which has been developed to monitor time- or treatment-dependent changes in cellular phospholipids (40–42). This program consists of data handling routines and a package of multistep statistical analysis and is based on normalization of the raw signal intensities observed at a particular m/z value. Such normalization allows the comparison of dynamic changes within different phospholipid profiles. As in other fields of systems biology, the interest on developing management systems for the amount of generated mass spectrometric data is increasing.

The mentioned data management systems are only first steps toward comprehensive lipid databases and global lipid networks. An example for such a data bank is LIPIDMAPS (http://www.lipidmaps.org), which covers structures and annotations of biologically relevant lipids (43). The structures originate from the core laboratories of the LIPIDMAPS consortium and their partners. In this database, users can search the LIPID MAPS proteome database using either text-based or structure-based search options. In addition to LIPIDMAPS, other databases in Europe (http://www.lipidomics.net) and Japan (http://www.lipidbank.jp) have been initiated.

After maturation of the technology used for lipid analysis, bioinformatics will be needed to correlate quantitative data on lipid concentrations with those of the enzymes and other proteins involved in lipid metabolism on the mRNA and proteome level. These data might be generated from experiments that use RT-PCR, DNA-microarrays, cellular fluorescence imaging, or flow cytometry (44). To this end, an important role in lipidomics will be exchange and integration of data from genomics and proteomics. Thus, the entire lipidomics approach must be understood as a part of system biology, including the above-mentioned fields.

Applications and Challenges

Lipid metabolism is a tightly regulated process. Alterations in lipid metabolism and aberrant lipid levels have been associated with frequently occurring diseases, for example, atherosclerosis, pancreatitis, diabetes, cancer, schizophrenia, neurodegenerative and respiratory diseases, and Parkinson or Alzheimer disease. Therefore, lipidomics found entrance into diagnosis and drug development, such as in the discovery and optimization of biomarkers, therapeutic targets, and lead compounds (1). In addition to this field, lipid levels can be altered in rare inherited disorders of lipid metabolism, or in response to drugs such as the large group of cationic amphiphilic drugs that are used, for example, antidepressants, neuroleptics, -Adrenoceptor antagonists, or Antihistaminics. A growing group of inherited disorders is known to affect lipid metabolism. Examples are defects in proteins required for cellular lipid trafficking (44), other defects impair sphingolipid metabolism (45), and even others might be unrecognized causes of even apparently unrelated diseases such as mental disorders or epilepsy (45). Although most of the genetic defects underlying these diseases and the identities of the major storage substances are known for more than a decade, the detailed changes in lipid composition and the pathogenic mechanisms that lead to the different phenotypes are far from clear. An improved understanding of the pathogenesis of these and other diseases can be expected from the application of lipidomics to this field.

Because of the inherent difficulties in lipid analysis, a comprehensive treatment of the lipidome in terms of systems biology has not been achieved to date, but it will be attempted for systems of reduced complexity. For example, alterations of all lipid classes in response to genetic and pharmacologic changes will be analyzed in murine macrophages by the multicentered LIPIDMAPS consortium. Recently, the protein and lipid composition of synaptic vesicles isolated from rat brain have been determined by mass spectrometry (46). With the exception of cholesterol, whose content of about 40% is very high in the vesicles, the lipid analysis confirmed data obtained by classic lipid analysis. But because ESI-MS allows determination of the chain length distribution in phospholipids, these data have been collected also for this (undisturbed) system. From a certain cell type via a certain vesicle type, let us mention a certain lipid. Cardiolipin is a bacterial lipid, and in eukaryotes it is the only phospholipid synthesized in the mitochondria. The chain length distribution of molecular cardiolipin species of rat liver and bovine heart is known for fifteen years. Changes in cardiolipin content, acyl chain composition, and cardiolipin peroxidation affects mitochondrial function and can contribute to human diseases like ischemia, hypothyroidism, aging, heart failure, and cardiakoskeletal myopathy (47).

Usually, lipidomics is understood as a way to determine the global lipid composition of a sample like tissues or cells. On the subcellular level, however, lipid content of certain organelles, the function of the lipid patterns, as well as the identity of the interacting partners is an actual area of research. For example, the lipidome of the nucleus, where a remodeling of phosphatidylcholine to molecular species with saturated acyl chains takes place, is analyzed currently not only with “static” but also with “dynamic” lipid profiling with the aid of stable isotope labeling (48). Indirect methods like antibody staining or photoaffinity labeling were necessary to determine lipid sorting associated with the maturation of intralysosomal membranes: Membranes of lysosomes are prepared for their lyosomal degradation by reduction of their cholesterol content and enrichment of a lipid characteristic for intralysosomal membranes, bismonosyglycerol)-phosphate (45). Novel lipid species have been discovered by mass spectrometric analysis. Knowledge of their structure and occurrence is required for the understanding of their function for the organism in health and disease, which includes a potential use as biomarkers. Only three examples for the growing number of such discoveries are mentioned briefly here. Fatty acid amides such as N- arachidonylthanolamine and N-oleoylthanolamine are signaling substances found in the brain. Within “targeted lipidomics,” the structures of low abundant acylamides like N- arachidonyl- dopamine and N- arachidonyl-glycerine have been determined by a combination of HPLC and ESI-MS (49). These lipids act in part as endogenously occurring cannabinoids and can regulate pain, immune function, reproduction,
and appetite. A second example is the discovery of 3-ketohyd-
droceramides. This metabolite is usually not found because
the precursor lipid, 3-ketosphinganine, is reduced rapidly. But
it occurs in cultured cells that overexpress serum palmito-
yltransferase and that are supplied simultaneously with sufficient
amounts of the metabolic precursors serine and palmitic acid
(17). Under these conditions, the capacity for ketone reduction
seems to be exhausted, and excess ketosphinganine is acylated
by one or more of the cellular acyltransferases. A final example
is the discovery of a protein-bound glycosphingolipid. Knock-
out animals as source offer the advantage of greater amounts
of otherwise not detectable metabolites. In mice deficient in
the degradation of glucosylceramide, a novel posttranslational
modification has been identified. In the animals, glucosyce-
ramides \( \omega \)-hydroxylated in the acyl residue are bound covalently
to proteins of the cornified envelope of the skin (50). Covalently
bound lipids are essential not only for organ development, but
also for the function of the human skin (18).

As mentioned before, knowledge of total lipid composition
is not necessarily sufficient to understand lipid function. For
example, the total cholesterol content of neurons from mice with
a deficiency of the Niemann–Pick disease C1-protein is not
different from that of control animals. Closer examination of
this animal model of the human Niemann–Pick disease, type C1,
however, revealed that the cholesterol level in cell bodies is
increased drastically at the expense of that in distal axons (51).
A determination of members of all lipid classes present in a
tissue is still difficult, and analysis of lipids within different
areas of a cell [e.g., by IMS (see above)] requires sample
preparation by freeze fracture. For the analysis of some prob-
tems, techniques such as metabolic labeling with radioactive
biosynthetic precursors might be appropriate. Results obtained
in experiments with artificial (e.g., fluorescence-labeled) lipids
must be interpreted with care because of entirely different prop-
eties compared with that of their native counterpart. After a
period of development of methods for analysis, data processing,
exchange, and visualisation, a significant impact of this tech-
nology on our understanding of lipid metabolism and function
can be expected.

References
1622-1624.
5. Marm S, Parton RG. Lipid droplets: a unified view of a dynamic
6. Janney PA, Kinnunen PK. Biophysical properties of lipids and
7. Morrison OG. Life-As A Matter of Fat. The Emerging Science
of Lipidomics. 2005; Springer, Berlin, Germany.
8. Sieber CN, Hong S, Luy. Lipid mediator informatics/lipidomics:
novel pathways in mapping resolution. AAPS J. 2006;8:E284-
E297.
10. Zwaal RF, Comribius P, Bava EM. Surface exposure of phos-
971-988.
11. Hancock JF. Lipid rafts: contentous only from simplistic stand-
12. Ostrowski SG, Kurucz ME, Rody TP, Winogard N, Ewing AD.
Secondary ion MS imaging to relatively quantitate cholesterol
in the membranes of individual cells from differentially treated
13. Carlton JR, Cullen PJ. Coincidence detection in phosphafluoride
and Other Lipid Mediat. 2005;77:197-209.
15. Kalyan RTY, Mangelsdorf DJ. LXR AND FXR: the Yin and
2006;68:159-191.
16. Matirakz H, Covic A, Schepers P. Comparison of different
extraction methods for lipid content determination in various
Dekker, New York, p. 254.
18. Holleran WM, Takagi Y, Uchida F. Epidermal sphingolipid
metabolism, function, and roles in skin disorders. FEBS Lett.
2006;580:5456-5466.
Quantitative analysis of biological membrane lipids at the
low picomole level by nano electrospray ionization tandem mass
20. Ekroos K, Chernomorov IC, Simons K, Shevchenko A. Quantita-
tive profiling of phospholipids by multiple precursor ion scanning
on a hybrid quadrupole time-of-flight mass spectrometer. Anal
C, Ekroos K, Shevchenko A. A uniformised ionization and quanti-
tication of glycosphingolipid molecular species by multiple
C, Ekroos K, Shevchenko A. Automated identification and quan-
tification of glycosphospholipids from biological samples by
nano electrospray ionization tandem mass spectrometry. Anal.
23. Han X. Characterization and direct quantitation of ceramide
25. Sandhoff R, Brügger B, Hough F, Lehmann WD. Determination of
cholesterol in biological membrane lipids at the low picomole level
by nano-electrospray ionization tandem mass spectrometry. J.
26. Liischt G, Binder M, Schifferer R, Langmann T, Schulz B,
Schmitz G. High throughput quantification of cholesterol and
cholesterol ester by electrospray ionization tandem mass spectrom-
27. Byrdwell WC. Atmospheric pressure chemical ionization mass
separation of stearin corn oil ceramides with detection by
covalent light scattering and APCI mass spectrometry. Anal.

See Also

Lipid Domains, Chemistry of
Lipids from Whole Cells and Tissues, Extraction of
Lipid Bilayers, Properties of

Further Reading

Lipid structure and dynamics are keys factors for cell life. Phospholipids are essential in signalling pathways as they are recognition sites at the membrane surface. Other lipids such as cholesterol and sphingomyelin are crucial molecules in maintaining membrane integrity and are involved in membrane domains (rafts). NMR both with its liquid and solid-state facets is a very powerful tool for deciphering both the structure and the dynamics of various classes of lipids. Lipids extracted from membranes and dissolved in organic solvents can reveal their molecular structure when observed with multinuclear one-dimensional or two-dimensional solution NMR. Multinuclear, wide-line, solid-state NMR brings information on the nature of the membrane phase (lamellar, hexagonal, isotropic, etc.), its dynamics (diffusion, fluid/gel, or liquid-ordered with cholesterol) and the molecular structure of embedded lipids when using the Magic Angle Spinning technique.
NMR of Lipids
Basics for liquid-state and solid-state
NMR of Lipids
Structure and Dynamics of Lipids
by NMR
NMR Interactions in Liquids
One-dimensional (1-D) NMR spectra in liquids are made of
very sharp lines whose linewidth is less than 1 Hz, because
the internal magnetic interactions that modulate the applied
external magnetic field are averaged to isotropic values (iso) by fast
(frequencies of 10^{-2}-10^{-15} Hz) isotropic tumbling of molecules
in solution. Chemical shielding (\delta) and indirect spin-spin (J)
interactions remain, which lead to one-dimensional (Hz) spectra
(N resonant nuclei):
\[
\delta(i) = \sum_{i} \nu_{iso}^{i} + \sum_{i<j} J_{ij}
\]

The sharp line positions, \nu_{iso}^{i}, are often converted into a mag-
netic field-independent variable, the well-known chemical shift,
\delta, which is expressed in parts per million (ppm) relative to
the magnetic field. Three types of information can be obtained
from a high-resolution 1D-NMR spectrum: 1) the NMR line
position, the chemical shift, reports on the isotropic local elec-
tric environment of the nucleus (atom) in the molecule. This
in principle allows identification of small molecular species us-
ing known table of chemical shifts. 2) Some lines exhibit fine
structure; they are split into multiplets (doublets, triplets, etc.),
which lead to information on the number of nearest-neighbor-
ning nuclei having an indirect spin-spin interaction (J). This
process is also used for molecular structure identification. 3) The
line intensity (area) is proportional to the number of magneti-

cal nuclei which means that NMR is a quantitative
method that can be used for finding the proportion of different
species in mixtures. For more complex molecular structures or
for lipid mixtures, 1-D-NMR is limited because spectra contain
thousands of lines. Two-dimensional (2-D) or three-dimensional
(3-D) NMR is then used. The principle of these experiments
is to apply a pulse sequence that will separate magnetic in-

cations or make correlations between them. For instance, a

correlation spectroscopy-type sequence (COSY, TOCSY, etc.)
will allow identification of pieces of molecular structure by cor-

relation of chemical shifts of different nuclei through J-coupling.
Through-space information can be obtained using pulse se-
quences (NOESY or ROESY) that afford measurement of direct
spin-spin interaction, which is inversely proportional to the
sixth of the distance separating nuclei. This information affords
reconstruction of the three-dimensional molecular structure at
atomic resolution.

Anisotropy of NMR Interactions in Membranes
or in Solid Aggregates
In liquid-crystalline media or in aggregates, molecular motions
are much slower and are anisotropic. Two consequences result:

Individual NMR lines may span several thousand hertz and
depend on the orientation of molecules with respect to the mag-
netic field (1-4):
\[
\nu_{A}(\beta, \alpha) = \nu_{iso} + A \left( \frac{3\cos^2\beta - 1}{2} + \frac{\nu_{iso} \cos^2\alpha \sin^2\beta}{2} \right)
\]

In cases where the asymmetry parameter \nu_{iso} is zero, when axial
symmetry exists for structural or dynamical reasons as in lipid
membranes, the equation simplifies to
\[
\nu_{A}(\beta, \alpha) = \nu_{iso} + A \left( \frac{3\cos^2\beta - 1}{2} \right) S
\]

\nu_{A}(\beta, \alpha) is the frequency of an individual NMR line, \alpha the azi-
muthal angle between the main axis of symmetry (lipid rotation axis for instance, \alpha in Fig. 2a) and the mag-
netic field direction B_{0}, and S the order parameter (0 to 1)
linked to dynamics. \nu_{iso} stands for the isotropic frequency as
detected in purely isotropic liquids. The constant A depends on
the internal magnetic interaction of interest: direct spin-spin or
dipolar interaction (D), chemical shielding interaction (s), or
electric quadrupolar interaction (Q). The D interaction is found
when two nuclei are close in space (few Angstrom) and when

An example of a line broadening due to the dipolar interaction
is shown in Fig. 1a.
it dominates $^1$H spectra; $\sigma$ mainly accounts for $^{31}$P, $^{19}$F, and $^{15}$N spectra; and $Q$ is only found for nuclei with a nuclear spin greater than one half that possess an electric quadrupolar moment ($^{2}$H, $^{14}$N, $^{17}$O, etc.). The indirect spin-spin interaction ($J$) is so weak compared with the other interactions that it is barely detected in solid-state spectra. As observed in Fig. 1a–c, the NMR lines shift during change in orientation of the membrane plane (angle $\beta$) with respect to the magnetic field. A doublet appears for $D$ and $Q$ interactions (Fig. 1a); it is considered here that only one pair of spin-1/2 nuclei interacting ($^1$H-$^1$H, $^1$H-$^13$C, etc.) or one spin-1 nucleus ($^2$H, $^{14}$N). This doublet collapses when the axis of motional averaging is oriented at the magic angle $\beta_m = 54.7^\circ$ with respect to $B_0$. Liposomes or cells usually do not orient in magnetic fields. Because these large ($\mu$m) entities slowly reorient, compared with the NMR frequencies, all membrane orientations ($\beta$ varying from 0 to 90$^\circ$) lead to a so-called “powder spectrum.” Traces of Fig. 1a–g are powder spectra for $\sigma$, $D$, and $Q$ interactions under conditions of axial symmetry as described by Equation 3, whereas traces in Fig. 1h stand for $\sigma$ and $Q$ interactions where no axial symmetry exists (Eq. 2) as in the pure solid state. As observed in Fig. 1, solid-state NMR lineshapes span several hundred kilohertz (“wide-line”); they are very different depending on the NMR interaction and depend on whether the sample is oriented. In some cases, interactions may be of the same strength, and for many nuclei (e.g., protons in lipids), spectra may become very complex.

Spectra and the effect of motional averages

As observed in Figs. 1 and 2, considerable differences exist between spectra for samples that are in the pure solid state and for molecules in solution, because of the considerable differences in motional regimes. Motions that may be detected by NMR range from picoseconds to seconds and can be intramolecular, involving bond rotations, and molecular such as rotation-diffusion of the entire molecule or collective (i.e., because of concerted diffusion of molecules in a bilayer [hydrodynamics]). Both spectra and relaxation times ($T_1, T_2$, etc.) may bring dynamical information. One of the powers of wide-line, solid-state NMR resides in the fact that motional processes with correlation times $\tau_c$
Figure 2. Isotropic NMR spectra. (a) Lipid–water dispersion in nonrotating mode (lack of high resolution). (b) Schematics of a lipid-containing rotor spinning at the magic angle. (c) High-resolution $^{13}$C spectrum of cholesterol (insert), proton-decoupled, and edited such as CH$_3$/CH show positive peaks and CH$_2$/quaternary, negative. (d) 2-D map correlating $^{31}$P and $^1$H isotropic chemical shifts in a mixture of dipalmitoylphosphatidylcholine and dioleoylphosphatidylglycerol.

(time for reorientation in a given phase) shorter than the reciprocal of the spectrum width will modify the observed line shape (4, 5). Figure 1 shows the effect of consecutive fast axial motions on NMR powder spectra; it has to be looked at from bottom to top, considering a hierarchy in time of motions (i.e., going from the fastest to the slowest). Intramolecular motions (Fig. 1g) typically occur at $\sim 10^{-12}$ s and average the static magnetic interactions ($\sigma$, $D$, $Q$) leading to axially symmetric lineshapes of smaller width. The molecular and collective motions ($\sim 10^{-9}$–$10^{-10}$ s) produce additional averaging that reduces the linewidth more (Fig. 1f). As a general statement, all anisotropic motions of $\tau_c$ shorter than 1 µs will reduce spectra width, which is simply translated in terms of less and less ordering (i.e., a reduction of the order parameter $S$ of Eq. 3). A special case is also shown in Fig. 1e: a fast motion that is present in hexagonal phases. Here the lipids are distributed along long tubes around which a fast rotation occurs. Because the new rotation axis makes an angle of 90$^\circ$ with respect to the former bilayer normal, this fast rotation leads to a geometry-induced reduction by a factor of two of the spectrum width, compared with a bilayer-type spectrum. In the case of the $\sigma$ interaction, this fast rotation leads to an inversion in the symmetry of the lineshape. Finally, the effect of an isotropic motion is shown in Figure 1d; the consequence of such motion is a complete canceling of the angular-dependent term of Equation 2 (i.e., all solid-state interactions vanish). The spectrum is then that of a sample in the liquid state where only the isotropic $\sigma$ and $J$ interactions are present. It must be mentioned here that isotropic motions may occur in membrane phases (i.e., the local structure may be organized (micelles, small vesicles, cubic phases, etc.), but the tumbling of the small (1 to 30 nm) object or the fast lateral diffusion on highly curved structures may lead to spectra of liquid-state NMR). Interestingly, for intermediate sizes (30 nm to 500 nm), the NMR lineshape is in between that of liquids and solids and may be used for determining vesicle size (6).

A special case of motional averaging: magic angle sample spinning

As fast spinning around a molecular axis leads to efficient averaging of the solid-state interactions, placing the sample inside a tube that spins at several kilohertz around an axis that is oriented at $\beta_m = 54.7^\circ$ (Fig. 2a) should in principle cancel the angular-dependent term of Equation 2. This result is what is obtained for lipid aggregates or membranes, the spectrum being left with the isotropic terms of $\sigma$ and $J$ interactions, as in pure liquids. How far lines can be narrowed depends on several...
Because the motions are activated thermally, their activation 31P and 1H chemical shifts \( \text{inverse HMQC-TOCSY} \) cases of a lipid mixtures pseudo-3-D experiment allowing correlated resonances. In this table are also reported 1H chemical shifts of lipids except for double-bond resonances or carbonyl resonances are falling in the same range and do not allow sorting by their head group and backbone resonances. The chain in chloroform. These two classes of lipids may be differentiated by their head group and backbone resonances. For phosphorus-containing lipids, 31P NMR can be applied in both liquid and membranous states and allows sorting various phospholipid classes.

**Analysis of hydrated lipid aggregates using solid-state NMR**

In the following discussion, some examples of wide-line NMR used to probe structure and dynamics of lipids in hydrated systems are presented.

**Lipid phases, transition temperatures**

As demonstrated in Fig. 1, wide-line NMR spectra are good tools to determine in which hydrated phase a lipid may be. This process has been used widely to determine phase diagrams (e.g., Dipalmitoylphosphatidylcholine(DPPC)-cholesterol using 2H-NMR (8) and Phosphatidylethanolamine(PE)-cholesterol (9, 10) with 1H and 31P NMR). Fig. 3a reports a temperature-composition diagram of short-chain and long-chain PC mixtures (11, 12). This system may form Lamellar phases (L), small disks (D), and isotropically tumbling very small objects (I) depending on composition and temperature. The small (40-80-nm) disks are called bicelles and are oriented by magnetic fields, with their normal parallel or perpendicular to the field \( B_0 \). Fig. 3a (inserts), which leads to unambiguous spectral assignment using the naturally abundant 2H and 14N nuclei as reporters (Fig. 3b).

An isotropic line is obtained for isotropically tumbling objects (800 ppm, bottom). Two (2H) or four (14N) sharp lines corresponding to small-chain and long-chain lipids are detected for disk normal oriented perpendicular or parallel to the field (middle), and an isotropic line superimposed on a powder pattern is observed for a biphase region where lamellar and isotropic phases coexist (top). Because phase diagrams are tiresome to build, partial phase changes can nonetheless be obtained by following spectral changes with temperature or hydration. PE or monoglycosyldiglycerides favor the appearance of hexagonal phases (13), lysolipids, diglycosyldiglycerides, and diacylglycerols of isotropic phases (14). Raising or lowering gel-to-fluid phase transition temperatures \( T_m \) can be followed by wide-line NMR (1H, 14N, 31P): Calcium-neutralization of negative head group changes of dialyglycerophosphatic acids increases \( T_m \) (15), whereas the addition of unsaturated lipids such as sea urchin Phosphatidylglycerol decreases \( T_m \) (14, 16).

**Bilayer internal dynamics: order parameters, membrane thickness, sterols**

Besides monitoring morphological changes on hydrated systems, wide-line NMR can also report on internal membrane dynamics (i.e., determine the membrane fluidity when in the lamellar fluid state). This is where the concept of order parameter takes place. Order parameters \( S \) are time and space averages of the quantity \( \langle \cos^2 \theta - 1/2 \rangle \). They can be expressed at the intramolecular level describing the fluctuation.
### Table 1: 1H and 13C chemical shifts of lipids in C2HCl3, reference TMS (δ ppm).

<table>
<thead>
<tr>
<th></th>
<th>Sphingomyelin</th>
<th>Liver PC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1H, CDC13</td>
<td>1H, H2O(DL)</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Head Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α</td>
<td>3.72</td>
<td>63.37</td>
</tr>
<tr>
<td>β</td>
<td>3.92</td>
<td>nd</td>
</tr>
<tr>
<td>γ</td>
<td>3.34</td>
<td>3.66</td>
</tr>
<tr>
<td>Back Bone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>3.92</td>
<td>nd</td>
</tr>
<tr>
<td>G2</td>
<td>5.21</td>
<td>3.13</td>
</tr>
<tr>
<td>G3</td>
<td>4.14</td>
<td>4.00</td>
</tr>
<tr>
<td>1H, H2O(DL)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.36</td>
<td>5.71</td>
</tr>
<tr>
<td>1'</td>
<td>173.2/173.58</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.36</td>
<td>5.71</td>
</tr>
<tr>
<td>2'</td>
<td>2.27/2.29</td>
<td>2.22/2.31</td>
</tr>
<tr>
<td>3</td>
<td>2.00</td>
<td>2.02</td>
</tr>
<tr>
<td>3'</td>
<td>1.57</td>
<td>1.60</td>
</tr>
<tr>
<td>4</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>4'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>5'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>6</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>6'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>7</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>7'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>8</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>8'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>Fatty Acyl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>9'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>10</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>10'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>11</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>11'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>12</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>12'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>13</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>13'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>14</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>14'</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>15</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>15'</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>16</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>16'</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

nd: not determined. Range numbers (e.g., 127.52–132.51) indicate that there are more than 2 lines in the range that cannot be assigned with confidence. Numbers separated by '/' (e.g., 127.83/129.89) indicate that two lines of indicated chemical shifts exist, which cannot be assigned with confidence. Primes stand for positions in sn-2 chain.
of bonds, at the molecular level, picturing the angular reorientation of molecules, or at the mesoscopic scale reflecting membrane deformations (4, 5). When expressed with respect to the local axis of symmetry, S values range from 1 to 0 and are simple tools to say whether a system is rigid (S ≈ 1) or liquid-like (S ≈ 0). As indicated in Equation 3, S is measured directly from the spectrum width. The best magnetic interaction to deal with when measuring order parameters is the quadrupolar interaction (C–2H bonds); however, the dipolar or the chemical shielding interaction can also be used (17). In Figure 4a (18) is shown a deuterium wide-line bicelle spectrum of fully 2H-labeled DMPC together with the molecule showing labeled positions. Assignment was made based on selective labeling (3, 19), and except for hindered positions, the smaller the doublet the smaller S. Chain ends and head group are very mobile, and positions 2–8 display the same high ordering called the “plateau” of order parameters. Relaxation times T1 can also report on membrane dynamics. Figure 4c shows how 1/T1 varies with labeled position. It is observed that both the acyl chains and the head group are dynamic, particularly toward the ends.

Orientation of molecules in membranes

Equation 2 indicates that frequency lines depend on the orientation of molecules with respect to the magnetic field. The spectrum is used to determine the average orientation of molecules in the membrane knowing the orientation of the membrane plane with respect to B0 and atomic coordinates of the rigid molecule of interest. Whereas bicelle membranes naturally orient in the field, classic membranes must be sandwiched in between glass plates that are oriented macroscopically by means of a goniometric device. The orientation of 2H-labeled cholesterol in membranes has thus been determined by combining neutron coordinates and wide-line deuterium NMR. Cholesterol is vertical in the membrane, which offers the minimum molecular area to optimize lipid packing, whereas α-cholesterol, the non-naturally occurring isomer is tilted (25). Labeling of lipid head groups also afforded measuring changes in orientation at the membrane surface during variation of the electric field (pH) (26).
Membrane dynamics from picoseconds to millisecond

From measurement of relaxation times, a minute analysis of intramolecular, molecular, and collective motions that occur in a membrane during temperature changes may be accomplished [Fig. 5 (5)]. Correlation times are plotted in a log scale as a function of the reciprocal of temperature in order to have a direct reading of activation energies from the slopes of the lines: the steeper the slope, the higher ${E_a}$. In the gel phase ($L_{β'}$, $P_{β'}$), only intramolecular ($τ_{1,2,3}$) and molecular motions ($τ_{//}$, $τ_{⊥}$) occur whose $τ_c$ ranges from the millisecond to the nanosecond. On going to the fluid $L_α$ phase, above $T_m$, an increase occurs by 3 orders of magnitude of the motion speeds. Intramolecular motions go into the picosecond range that is barely detected by NMR. Collective modes of motions (membrane large-scale deformation) come into play only in the fluid phase. The addition of cholesterol to this system levels off the large temperature-induced change in $τ_c$ at $T_m$ (27). Lateral diffusion of molecules in the membrane plane can also be measured using a combination of oriented samples and pulsed gradient sequences (28). Indication of restricted diffusion as for domain formation was found in “raft” mixtures of sterols and sphingolipids.

Chemical Tools and Techniques

Chemicals

Natural lipids are extracted following known procedures (29). Synthetic lipids are also prepared according to reported procedures (30). Many lipids are available commercially. Membranous lipids can also be prepared directly from cells using several detergent-centrifugation steps.

Sample preparation

Lipids for solution NMR are dissolved in deuterated organic solvents (CDCl$_3$, CD$_3$OH, DMSO, etc.) at concentrations varying from µM to mM. Solutions are transferred into
NMR of Lipids

The classic battery of solution NMR experiments is used for lipids in solution or spinning at the magic angle (31). For samples that are under the form of aggregates, cross-polarization experiments are often used to enhance signals of low concentrated and/or diluted nuclei (13C, 31P) by taking advantage of the proton reservoir from which the magnetization may be transferred (31). The most used 1-D sequences for static experiments are spin-echoes that refocus the chemical-shielding, dipolar or quadrupolar interactions and cope with the quickly damping signals often observed in solids (2). These sequences are implemented into 2-D variants capable of correlating dipolar/quadrupolar interactions versus chemical shielding (32). Measurement of relaxation times and diffusion constants are made using classic sequences (inversion recovery, echoes, gradients) that are merged with spin-echo sequences.

Conclusion and Perspective

NMR is the only nondestructive and quantitative method capable of bringing the atomic 3-D structure and dynamics of lipids both in solution and in aggregated form. The only drawback of the method is its relatively modest sensitivity: It requires micrograms to milligrams of material. However, NMR is an ongoing science that still develops both by inventing new sequences and by developing new technologies (high field and cryoprobes). It is expected that detection limits will be lowered and that analyses will be made in living materials.

References


WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.


See Also

Analytical Chemistry in Biology
Labeling Techniques: Lipids
Lipid Bilayers, Properties of
Membranes, Fluidity of
Physical Chemistry in Biology
Tools for Detection of DNA Polymorphisms

Sascha Sauer, Max-Planck-Institute for Molecular Genetics, Department of Vertebrate Genomics, Berlin, Germany
doi: 10.1002/9780470048672.wecb464

DNA polymorphisms, such as single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs), are used as genetic markers in linkage and association studies to dissect monogenic and polygenic traits. Structural variation and DNA methylation are additional sources of genetic variation and are becoming increasingly important, particularly in the analysis of cancer. In this article, biologic background, basic chemistries, and currently used technologies for the detection of these polymorphisms are presented with a focus on SNP analysis.

Biologic Background

Variations in DNA sequence between individuals are termed "DNA polymorphisms." In biology, a polymorphism—which derives from ancient Greek "many" and morph "form"—generally is defined as the occurrence in the same habitat of two or more forms of a trait in a population in such frequencies that the rarer form cannot be maintained only by recurrent mutation. Evolutionary forces act on genetic polymorphisms when organisms with different polymorphism alleles reproduce with varying degrees of success. With the availability of the reference sequence of the human genome along with a growing number of additionally sequenced genomes, genetic analysis, which includes genotyping of DNA polymorphisms, enters a new stage of research (1).

Many geneticists aim to study multifactorial traits in humans caused, among others, by genetic constitution and environmental factors. Genotyping, the determination of the specific allelic composition of a genome, a set of genes or a single gene, is a key element in these studies (2). Currently, statistical genetics relies heavily on linkage and association approaches. In general, linkage analysis uses DNA polymorphisms such as short tandem repeats (STRs) (mainly microsatellites) and SNPs as markers to assign chromosome map position to genes by genetic crosses prescribed in databases. Presently, over 10,000 STR sequences are described in databases. Presently, over 10,000 STR sequences are published for the human genome. A selected, informative subset of polymorphic microsatellites is applied readily in forensics. They are used to identify and trace potential criminals because of their individual genetic "barcode" that consists of a combination of varying microsatellite alleles at a few genomic loci. The Federal Bureau of Investigation in the United States has selected 13 STR loci to generate an individual human genetic profile. These profiles are stored in DNA databanks such as CODIS (http://www.fbi.gov/lab/codis/index.html). The ratio of STR genotyping and fingerprinting is based on previous approaches using time-consuming restriction fragment length polymorphism (RFLP) analysis (4). RFLP analysis is performed by using a restriction enzyme to cut the DNA specifically at

Outlook
Single nucleotide polymorphisms (SNPs)

SNPs are single-base substitutions, with the minor allele that occurs at least for 1% in a population (5). Otherwise, they would be termed point mutations. Often, single-base deletions and insertions are included in this definition. The majority of SNPs were discovered in DNA sequencing projects during the last decade. Presently, around 10 million human SNPs are stored in public databases such as dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Not all are validated and confirmed in all major populations. As STRs, SNPs can be used for several genetic applications, such as linkage and association analysis, diagnostics, and forensics.

Moreover, the haplotype of SNPs—the combination of SNP alleles that is found on a single chromosome of a given individual—is important in association analysis (6). Unfortunately, molecular haplotyping procedures are presently very expensive and not amenable to automation. Therefore, many researchers use statistical methods to infer haplotype pairs from multilocus genotypes (7). Resulting haplotypes are prone to misclassification, particularly if a large number of heterozygous SNPs are analyzed. The international HapMap project (http://www.hapmap.org/index.html.en) started in 2002 to compare human genome sequences of different reference populations to identify chromosomal regions where genetic variants are shared (i.e., haplotypes of SNPs). Nowadays, tag SNPs that comprise several redundant SNPs are used for whole-genome association scanning (8). In these genotyping experiments, several hundred thousand SNPs are analyzed in a limited number of about 100 selected individuals. The restriction in DNA samples is mainly because of the high cost of these experiments. However, first indication of chromosomal regions that contribute to the traits under investigation can be gained by such studies. A limited number of SNPs (∼1–1000) is used subsequently to investigate selected genomic regions in upsampled numbers of DNA samples. Resulting genotypes are explored either for direct association of causative SNPs or for indirect association of SNPs in an allelic association (often termed linkage disequilibrium) with causative loci.

Structural variations

Another class of polymorphisms that becomes increasingly important is structural variations (9). These variations are defined broadly as genomic alterations such as reciprocal translocations, deletions, duplications, insertions, and inversions. Structural variations involve segments of DNA that are larger than 1 kb, and they can be either microscopic or, in many cases, submicroscopic. Submicroscopic structural variants range from ∼1 kb to 3 Mbp in size and their analysis with current molecular biologic procedures is still much more complicated than the analysis of simple polymorphisms such as SNPs. The most important structural variations in the human genome are copy number variations (CNVs), which include large insertions, deletions, and duplications. CNVs are defined as a segment of genomic DNA, which is 1 kb or larger in size and occurs at a variable copy number compared with a reference genome. As for SNPs, a CNV is termed a copy number polymorphism if a variation occurs for at least 1% in a population. CNVs are known to be involved in several diseases of gene dosage, such as Angelman syndrome. CNVs are also involved in complex (multifactorial) diseases such as cancer; therefore, many researchers suggest a CNV analysis step should be added to all disease association studies, which are based currently on SNPs. Similar to SNPs, databases of structural variation presently are built up for the human genome. The human genome contains ∼3 x 10^6 base pairs, and it is estimated that the nucleotide content of CNVs per genome is about 10–30 million base pairs.

Methylation variation position (MVPs)

'S'-methylation of cytosines at CpG dinucleotides in genomic DNA is another source of DNA variation, which, in contrast to stable SNPs, can vary depending on the developmental stages and the tissues of an organism. These digital variations are called methylation variable position (MVP) (10). "CpG islands" are a large number of CpG dinucleotides adjacent to each other that occur in regulatory genomic regions. As SNPs, MVPs or patterns thereof can predispose to or indicate human diseases such as colon cancer. Moreover, the combination of SNP alleles with MVP alleles, called "haplotype," can lead to differential phenotypes. For example, by modifying regulatory genomic regions such as promoters, MVPs could cause differential gene expression as a result of varying binding efficiencies of RNA polymerase complexes.

Chemistry

In general, the underlying principle of almost all biochemical reactions discussed in this article is the well-known Watson-Crick DNA base-pairing rule; the purine adenine (A) forms a base pair with the pyrimidine thymine (T), and the purine guanine (G) forms a base pair with the pyrimidine cytosine (C) in DNA duplexes. Weak hydrogen bonding is the chemical mechanism that specifies the base-pairing rules. Moreover, electrostatic repulsion and stereoechemical criteria inhibit other potential base-pairing constellations.

DNA modification chemistry

Only some examples of DNA modification chemistry that provide the basis for many different biologic applications can be mentioned here. One development of importance was beyond doubt the introduction of phosphoramidite chemistry for nucleotides and automated synthesis of oligonucleotides in the 1970s (11). Double-strand specific intercalating dyes, such as ethidium bromide and more recent dyes such as SYBR Green I (a cyanine dye), are useful particularly in the analysis of DNA products. The development of sensitive reporter molecules such as fluorescent dyes has also been crucial for the analysis of nucleic acids. These dyes are attached chemically to nucleic acids and nucleotides (12). DNA modification chemistry was also crucial for the development of nonnatural nucleotides
that are substrates of natural or engineered enzymes, such as 2',3'-dideoxynucleotide triphosphates (ddNTPs), used commonly as polymerization terminators in Sanger sequencing and primer extension (13). These dideoxynucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides during DNA strand elongation. Recently, the company Solexa applied chemically modified reversible terminator nucleotides tagged with removable fluorescence dyes between two nucleotides during DNA strand elongation. Recently, the company Solexa applied chemically modified reversible terminators to target DNA sequences that contain an SNP. The difference in thermal stability between a perfectly matched and mismatched hybridization probe and its DNA target is used to distinguish between the SNP alleles. The thermal stability of such a hybrid depends on the nucleotide sequence that flanks the SNP, and additionally on the stringency of the reaction conditions (i.e., temperature and ionic strength). This problem complicates multiplex allele-specific hybridization for genotyping several SNPs simultaneously, but it can be mitigated by careful selection of SNPs on the basis of their performance in hybridization assays. In general, to interrogate a single SNP, several allele-specific oligonucleotide probes are used (16). For example, DNA microarrays are designed to carry four sets of ten allele-specific oligonucleotides that correspond to both strands of the SNP alleles with matched and mismatched probes for each SNP and four additional nucleotide positions flank the polymorphic site. For highly parallelized SNP genotyping, the requirement of such oligonucleotide panels for a SNP results in large arrays. SNP genotypes are assigned according to their joint fluorescence patterns obtained by 40 hybridization reactions for each SNP. The analysis of these patterns is performed by classification-based or model-based algorithms, which are tailored to this application. Two microarrays with a probe feature size of 8 μm that carry 2.5 million different probes per array were applied to analyze over 100,000 SNP genotypes in an individual DNA sample (16). Manufacturing DNA microarrays with oligonucleotides at such high density is performed routinely by using photolithographic techniques and combinatorial synthesis of the oligonucleotide probes on the arrays (17).

Dynamic allele-specific hybridization (DASH) (18) is another hybridization-based approach. DASH, the hybridization state of a duplex formed between PCR products that contain SNPs and allele-specific oligonucleotides is observed dynamically by fluorescence detection of double-strand specific SYBR Green I incorporation. The denaturation profile caused by increasing the temperature steadily during hybridization can be monitored by decreasing fluorescence signals. SNP genotypes are detected according to different melting temperatures of allele-specific oligonucleotides. DASH can be performed in different formats using, for example, membranes or microarray surfaces.

Enzyme-assisted SNP genotyping procedures based on DNA polymerase reactions and/or ligase reactions provide highly specific detection of SNP alleles (see Fig. 1). Primer extension and ligations are significantly less dependent on the DNA sequence context compared with hybridization; therefore, they enable genotyping of most SNPs using similar reaction conditions. An obvious advantage of enzyme-based SNP typing compared with conventional allele-specific hybridization is that it reduces the number of oligonucleotides required for multiplex SNP analysis.

DNA ligation is highly specific enzymes that join two adjacent and perfectly matched DNA strands to a template such as genomic DNA or a PCR product (19). Ligated products can be amplified and detected. This principle has been exploited in the oligo ligation assay and related procedures.

Primer extension reaction has become one of the most widely used methods for SNP genotyping because of its robustness and flexibility (15). In this reaction, an extension primer hybridizing to both strands of the SNP under investigation, is extended with a thermostable DNA polymerase. The simplest reaction principle is allele-specific hybridization of oligonucleotides to target DNA sequences that contain an SNP. The difference in thermal stability between a perfectly matched and mismatched hybridization probe and its DNA target is used to distinguish between the SNP alleles. The thermal stability of such a hybrid depends on the nucleotide sequence that flanks an SNP, and additionally on the stringency of the reaction conditions (i.e., temperature and ionic strength). This problem complicates multiplex allele-specific hybridization for genotyping several SNPs simultaneously, but it can be mitigated by careful selection of SNPs on the basis of their performance in hybridization assays. In general, to interrogate a single SNP, several allele-specific oligonucleotide probes are used (16). For example, DNA microarrays are designed to carry four sets of ten allele-specific oligonucleotides that correspond to both strands of the SNP alleles with matched and mismatched probes for each SNP and four additional nucleotide positions flank the polymorphic site. For highly parallelized SNP genotyping, the requirement of such oligonucleotide panels for a SNP results in large arrays. SNP genotypes are assigned according to their joint fluorescence patterns obtained by 40 hybridization reactions for each SNP. The analysis of these patterns is performed by classification-based or model-based algorithms, which are tailored to this application. Two microarrays with a probe feature size of 8 μm that carry 2.5 million different probes per array were applied to analyze over 100,000 SNP genotypes in an individual DNA sample (16). Manufacturing DNA microarrays with oligonucleotides at such high density is performed routinely by using photolithographic techniques and combinatorial synthesis of the oligonucleotide probes on the arrays (17).

Dynamic allele-specific hybridization (DASH) (18) is another hybridization-based approach. DASH, the hybridization state of a duplex formed between PCR products that contain SNPs and allele-specific oligonucleotides is observed dynamically by fluorescence detection of double-strand specific SYBR Green I incorporation. The denaturation profile caused by increasing the temperature steadily during hybridization can be monitored by decreasing fluorescence signals. SNP genotypes are detected according to different melting temperatures of allele-specific oligonucleotides. DASH can be performed in different formats using, for example, membranes or microarray surfaces.

Enzyme-assisted SNP genotyping procedures based on DNA polymerase reactions and/or ligase reactions provide highly specific detection of SNP alleles (see Fig. 1). Primer extension and ligation are significantly less dependent on the DNA sequence context compared with hybridization; therefore, they enable genotyping of most SNPs using similar reaction conditions. An obvious advantage of enzyme-based SNP typing compared with conventional allele-specific hybridization is that it reduces the number of oligonucleotides required for multiplex SNP analysis.

DNA ligation is highly specific enzymes that join two adjacent and perfectly matched DNA strands to a template such as genomic DNA or a PCR product (19). Ligated products can be amplified and detected. This principle has been exploited in the oligo ligation assay and related procedures.

Primer extension reaction has become one of the most widely used methods for SNP genotyping because of its robustness and flexibility (15). In this reaction, an extension primer hybridizing upstream of the SNP under investigation, is extended with ddNTPs and/or ddNTPs by a thermostable DNA polymerase. The DNA polymerase elongates the 3'-end of the primer in a temperature-cycled reaction by incorporating nucleotides that are complementary to the DNA template, which leads to a linear amplification of products of SNPs. The reaction terminates at the first nucleobase in the template where a nucleotide occurs that is complementary to one of the ddNTPs in the template. Depending on the reaction conditions, the extension products can be small.

DNA sequencing reactions

Using a modified protocol of the primer extension reaction, the so-called chain termination protocol, several different products (sequencing ladders with sizes up to about 1000 bases) are obtained. Essentially, this principle is applied in the Sanger sequencing method developed by Frederick Sanger et al. in the 1970s (13). This procedure requires a DNA template such as a PCR product, a primer, a DNA polymerase, and a mix of deoxynucleotides (dATP, dGTP, dCTP, and dTTP) and dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) that terminate DNA strand elongation. The dideoxynucleotides are added at significantly lower concentration than the standard deoxynucleotides to the reaction mix. Stochastic incorporation of a dideoxynucleotide into the nascent DNA strands terminates extension, which results in various DNA fragments of different lengths. Detection can be performed on gel-based or capillary-based systems using radioactively or nowadays fluorescently labeled nucleotides or primers. Although this method was applied very successfully, for example, to sequence the human genome, some nucleotide repeat-rich genomic regions such as telomeres, centromeres, and AT-rich intergenic repeats have often proved difficult to analyze with polymerase-based sequencing protocols (1).
Also in the 1970s, Maxam and Gilbert developed a "chemical sequencing" method, which is still useful for the study of DNA–protein interactions, nucleic acid structure, and epigenetic modifications (20). Similar chemistry was later used for SNP detection methods by chemical mismatch cleavage, a variant of enzymatic mismatch cleavage. The Maxam–Gilbert procedure requires radioactive labeling at one end and purification of the DNA molecule to be sequenced. Chemical treatment that uses four specific reactions (G, A+C, C, and C+T) produces breaks at a small proportion of one or two of the four nucleotide bases in the DNA template. Thus, a variety of labeled DNA fragments is generated, from the labeled terminus to the first fragmentation site in each molecule. The fragments are size-separated by polyacrylamide gel electrophoresis with the four reactions arranged side by side. To visualize the products, the gel is exposed to a film for autoradiography, which yields an image of DNA fragment ladders in four columns from which the sequence can be interpreted.

A new DNA sequencing approach developed by Solexa (www.illumina.com) uses massively parallel sequencing of millions of fragments by reversible terminator-based sequencing chemistry (14). The method relies on the attachment of randomly fragmented genomic DNA to a planar, optically transparent surface and solid-phase DNA amplification to generate a high-density flow cell with more than 10 million clusters, each of which contains about 1000 copies of template per cm². The DNA templates are sequenced in parallel on the surface by using a four-color sequencing-by-synthesis procedure, which applies reversible terminators with removable fluorescence. This procedure generates sequences with high accuracy and minimizes artifacts with homopolymeric DNA base repeats. Short sequence reads of about 25 bases can be aligned against a reference genome and genetic differences can be detected by applying specific software.

Pyrosequencing is another novel procedure for DNA sequencing that—as the Solexa sequencing method—also relies on the sequencing-by-synthesis principle (21). This method is based on a chemiluminescent enzymatic reaction, which is triggered when DNA polymerization occurs. Each time a nucleotide is incorporated specifically into a growing DNA chain by a polymerase accompanied by release of pyrophosphate in an equimolar quantity to the amount of incorporated nucleotide, a cascade of enzymatic reactions is activated. First, ATP sulfurylase converts pyrophosphate quantitatively to ATP in the presence of adenosine 5′-phosphosulfate. The newly synthesized ATP induces a luciferase-mediated conversion of luciferin to oxyluciferin, which produces visible light in amounts that are proportional to the number of nucleotides incorporated. It can be detected by a charge-coupled device camera and observed as a signal in a so-called pyrogram. Apyrase is used to degrade unincorporated dNTPs and excess ATP continuously. When this enzyme completes degradation, another dNTP is added (one at a time). Alpha-thio triphosphate-dATP is used as DNA polymerase substrate because it is not recognized by the luciferase. As the process continues, the complementary DNA strand is synthesized and the nucleotide sequence is determined by reading the signals in the pyrogram. Currently, sequences of up to 150 DNA bases can be reliably read by this technique (www.454.com).

Sequencing by ligation can be accomplished through a series of iterative cycles of ligation reactions, in which a sequencing primer is added followed by a pool of fluorescently tagged...
B-mer oligonucleotide probes (22). The probes are designed in sets of four, in which the fifth base serves as the interrogating base. After ligation, the array is washed, imaged, and the probes are cleaved away. Once the dye is removed, the native DNA strands remain so that the ligation, detection, and cleavage cycles can be repeated. This process allows a signal to be generated at the 10th base in the second cycle, at the 20th base in the third cycle, at the 20th base in the fourth cycle, and so on. The intervening bases are detected using different nested sequencing primers. The whole process is repeated until all intervening positions in the sequence are imaged. The ligase approach could be less vulnerable to repeat-rich regions compared with the polymerase-based approaches. For more information, the reader is referred to the Applied Biosystems website (www.appliedbiosystems.com).

The last three methods described belong to the second-generation sequencing approaches. For DNA polymorphism analysis, a current bottleneck in the widespread application of these methods consists in the lack of an efficient template preparation procedure to focus on selective sequencing of genetically or medically relevant genomic regions.

DNA amplification

Efficient DNA amplification methods were important for many genetic applications in the last two decades. The polymerase chain reaction (PCR), developed in the 1980s, is a crucial molecular biology method to replicate short stretches of DNA enzymatically without using a living organism (in vitro) (23). Kjell Kleppe et al. (24) already conceived the principles of PCR in 1971 (24). Kary Mullis et al. (23) established first PCR protocols in the 1980s at the biotechnology company Cetus. They also used thermostable DNA polymerases from thermophilic bacteria such as Thermus aquaticus, which were isolated and first studied functionally by Kaledin et al. (25).

PCR has become a standard method to reduce the complexity of genomes that consist of millions or billions of bases, and to increase specificity of the copy number of short stretches of DNA templates to levels required for analysis. PCR has been refined and extended to several applications, such as genotyping of DNA polymorphisms and quantitative mRNA expression analysis.

In general, during PCR, template DNA is denatured at a high temperature (~95 °C), which separates the double DNA strand, allows subsequent annealing of oligonucleotides called primers at lower temperatures (~55 °C), and extends nucleotide sequences along opposite strands at ~72 °C. This thermocycling process is repeated about 30 times, and it results in exponential accumulation of short DNA products (with sizes ~100–300 base pairs). Usually, PCR products are electrophoresed on agarose or polyacrylamide gels, which are made visible by DNA double-strand intercalating dyes and are validated by comparison with size standard markers.

Genetic markers such as SNPs or STRs can be amplified for genotyping by PCR using primers that anneal to unique flanking regions of these polymorphisms. STR genotypes can be analyzed easily according to their product sizes. PCR primers that flank microsatellite loci are simple and easy to use, but the development of correctly functioning primers can be a tedious and expensive process. In general, STR analysis, as well as amplification of repeat regions of a genome, suffers from slippage of the DNA polymerase, which leads to truncated products that are observed as "stutter bands" in electrophoresis.

A n strategy for DNA amplification represents the rolling-circle amplification—a mode of DNA replication used by bacteria and viruses to amplify circular plasmid DNAs—which generates molecules that look similar to lariats. A single primer is required to anneal to the target DNA and to synthesize a concatenated DNA strand that consists of thousands of copies using T7 polymerase, which also has a 3’–5’ exonuclease activity (26). In contrast to PCR, rolling-circle amplification is an isothermal method that is performed usually at 37 °C, which leads to linear and clonal amplification and takes place at the target DNA.

Parallelization of DNA sample preparation

A high degree of parallelization of sample preparation is of paramount importance to reduce costs in DNA genotyping and sequencing. The ability to perform highly parallel genomic assays to detect hundreds of thousands of SNPs simultaneously depends on two fundamental prerequisites: a highly parallel read-out and an intrinsically scalable, multiplexable sample preparation. The difficulty in optimizing multiplex PCRs consists in sequence-dependent differences in PCR efficiency. This difficulty increases exponentially with the number of genomic target regions to be amplified because of potential interactions between the PCR primers in the reaction mixture. This interaction leads to preferential amplification of undesired short "primer-dimer" artifacts. The multiplexing level that can be achieved by PCR is low (up to about 100 simultaneous reactions) and does not reach the level of high-parallel technologies based on miniaturized DNA microarrays or sequence read-out devices. However, current detection sensitivity of fluorescence or luminescence light scanners still requires an amplification step to generate a reasonable amount of DNA product.

A main bottleneck in DNA sequencing consists of the preparation of DNA samples from the bacterial colonies that contain a DNA library. In the Human Genome Project, each clone was picked individually and grown-up, and target DNA was extracted or amplified out of the clone (1). Recently, "polonies" (contraction of "polynucleotide colony") were used to prepare DNA templates for high-parallel DNA sequence analysis (27). Polonies represent discrete clonal PCR amplifications of a single DNA molecule, which is embedded in a polymer matrix by diffusion. Local amplification of a single DNA is performed by using a primer immobilized to the gel matrix. Several variations of this principle, such as bridge-PCR (28), can be employed, for example, by using planar surfaces and two immobilized primers (www.illumina.com). BEAMing (beads, emulsions, amplification, and magnetic)-based cloning on beads is a popular method to parallelize PCR amplification of many genomic loci in which a library of clones is grown on beads through the use of compartmentalized water-in-oil emulsion PCR (29). DNA and beads are diluted as far as, on average, a single micrometer-sized bead and a single target molecule co-occupy a single compartment (30, 31). PCR amplifies a clonal population of DNAs on the
A popular approach to reduce genomic complexity before SNP typing on high-density oligonucleotide microarrays consists of cleavage of genomic DNA by applying restriction enzymes and ligation of common adaptor sequences to the genomic fragments produced. Then, the adaptor sequences are used as binding sites for amplification by PCR using a single universal primer (15). The complexity of the human genome can be decreased by a factor of 50 using a single restriction enzyme followed by PCR with a universal primer to amplify fragments 250–1000 bp in size (see Fig. 2). This method allows for genotyping of 10,000 selected SNPs, located in the amplified fragments using allele-specific hybridization to oligonucleotides immobilized on high-density microarrays. The multiplexing level of this approach has been increased recently to 500,000 SNPs by applying a combination of two restriction enzymes for the complexity-reduction and amplifying larger PCR fragments up to 2000 bp in separate reactions. Recent SNP arrays allow for parallel genotyping of about 1,000,000 SNPs.

Carrying out the allele-specific reaction directly on genomic DNA, and subsequently applying PCR only to increase the number of detectable products, is a reverse, flexible approach to address the problem of amplifying a large number of genomic loci specifically (see Fig. 3) (32). One approach uses hundred or thousands of “circularizable” oligonucleotides called padlock or molecular inversion probes. These probes have two sequences that anneal complementary to adjacent regions in genomic DNA that contain SNPs, followed by an enzymatic step to close the circle allele-specifically. In the case of the padlock probes, the enzymatic step is an allele-specific ligation, whereas an additional gap-filling reaction for the molecular inversion probes is accomplished by a DNA polymerase followed by circularization using a ligase. The closed circles can be amplified exponentially by PCR or by rolling circle amplification. Circularizable probes allow for flexible in vitro genotyping in the range of 1000–100,000 SNPs in a single reaction (33). The application of generic microarrays that carry “molecular barcode” (tag) sequences to capture the amplified ligated products allow more flexible genotyping than the approach described in the previous paragraph.

Alternatively, oligonucleotide pairs are used to recognize the target DNA flanking the SNPs before enzyme-assisted genotyping (see Fig. 4). The allele-specific reactions produce templates, which were amplified subsequently by PCR using a primer for which universal binding sites have been inserted in the oligonucleotide pairs. Furthermore, in the probe oligonucleotides, molecular barcode sequences were incorporated, which were used to capture specifically the amplified products on beads that carry immobilized complementary DNA sequences. The beads were then assayed on optic fiber devices. This “Golden Gate” assay can be multiplexed to detect panels of about 2000 SNPs in parallel (34). Moreover, bead arrays were used for genome-wide genotyping of hundreds of thousands of SNPs (35). Therefore, a large number (about $10^6-10^7$) of fragmented copies of the genome were generated by whole-genome amplification using short random primers, and a

Figure 2: Genomic complexity reduction for microarray hybridization. Genomic DNA is digested with a restriction enzyme, and adaptors are ligated to the ends of resulting fragments. These fragments are amplified by using one of the strands of the adaptor as a primer. Fragments in the size range 250–1000 bp are amplified by PCR. The narrow size range of the PCR products represents ~400 base pairs which corresponds to a 50-fold reduction in genome complexity. To optimize efficient hybridization, the PCR products are fragmented by DNase I. The products are biotinylated and are subsequently hybridized to oligonucleotide microarrays. After a series of washing and signal generation steps, the fluorescence signals on the microarrays are scanned and genotypes are detected based on hybridization signal intensities.

Signal amplification method was applied to achieve sufficient fluorescence-detection sensitivity after genotyping. This procedure for whole-genome genotyping provides the basis for a new system to genotype a panel of about 300,000–550,000 human SNPs and is presently being extended to about 1,000,000 SNPs.

Specific analytical chemistry for structural variation detection

To define a structural variant, the analysis of structural variants relies on the comparison with a reference genome (9). Reference sequences such as the human genome sequence represent a hybrid assembly, which reflects the hierarchical mapping and sequencing approach that was applied (3). So far, no human DNA source has yet been defined as a standard reference control. The lack of a single reference DNA complicates both the identification and interpretation of the data, which is required for determining the hierarchy of mapping and sequencing methods used to detect variations on different studies and on standardization of databases, which contain information on structural variations.

Large structural variations are detected by microscopy-applying chromosome-banding methods using, for example, Giemsa

© 2008, John Wiley & Sons, Inc.
Affymetrix, Molecular Inversion Probes

Hybridization

PCR Amplification + Labeling

Figure 3 Reaction steps of the molecular inversion probe assay from Affymetrix illustrated for a heterozygous SNP. Oligonucleotides that contain target-specific regions (red), binding sites for PCR primers (gray), and molecular barcode sequences for capture on a solid support (blue) are allowed to anneal directly to genomic DNA. The 3′-end of the upstream target-specific oligonucleotide probe is elongated by a DNA polymerase, followed by ligation to the downstream oligonucleotide. The resulting circular molecule is cleaved at a site between the PCR primers before and after PCR amplification, which is followed by fluorescent labeling. The labeled molecules are captured finally on GeneChip glass microarray that carries complementary molecular barcode sequences for fluorescence detection. With permission from Ref. (15).

staining of elongated prometaphase chromosomes to reveal discrete structural variations (9). Recent developments in fluorescence in situ hybridization (FISH) analysis (36) on interphase cells and metaphase chromosomes enable a more refined characterization of the extent of these variants. In high-resolution fiber FISH, oligonucleotide probes fluorescence labeled in different colors are hybridized to stretch chromosome fibers mechanically, which allows copy numbers to be counted directly.

Recently, whole-genome scanning array technologies and comparative DNA-sequence analyses such as clone-based comparative genome hybridization (array-CGH) revealed DNA variation, which involves genomic segments that are smaller than those that can be observed by microscopy (37). Copy number variants that are greater than 50 kb could be detected efficiently by CGH. Array-CGH can even be used to discover CNVs in formalin-fixed, paraffin-embedded clinical samples that imply varying levels of degradation. CGH procedures apply fluorescent dyes such as Cy5 to label fragments from a genome under investigation. The genomic fragments are hybridized competitively with a second differentially labeled reference genome using Cy3 tags to surfaces that are printed with cloned DNA fragments, which reveals eventually copy number differences between the two genomes under investigation. Genomic clones such as bacterial artificial chromosomes (BACs) and
Tools for Detection of DNA Polymorphisms

Illumina Golden Gate Assay

Hybridization

Primer-Extension

Ligation

Capture on Bead Array

Figure 4 Reaction steps of the GoldenGate assay from Illumina. The fragmented genomic DNA is biotinylated and immobilized on avidin-coated microparticles. For polymerase extension and ligation reaction, a pair of allele-specific primers is used with similar functional blocks as the approach in Fig. 3. The allele-specific DNA molecules are amplified by PCR using universal primers that are tagged with fluorescence labels. The labeled molecules are finally captured on beads carrying complementary molecular barcode sequences for fluorescence detection. Finally, the beads are arrayed on optic fiber devices. With permission from Ref. (15).

Specific analytical chemistry for MVP detection

The analysis of MVPs requires some additional experimental steps compared with that for SNPs (39). One approach uses cleavage of genomic DNA by methylation-sensitive restriction enzymes. However, methylation-sensitive enzymes do not interrogate all CpG sites. About one-third of all CpGs in the human genome can be assayed using a combination of enzymes. Alternative procedures apply immunoprecipitation with methyl-binding protein or antibodies against methylated cytosines. The two approaches provided here are used in combination with highly parallel microarray detection.

The standard approach to analyze MVPs requires the conversion of the methylation difference into a sequence difference. This conversion is accomplished by bisulfite treatment of DNA, in which the 5’-methylcytosine remains stable while cytosines are deamminated (40). Using PCR amplification, these deaminated bases are converted to thymine bases. The amplicons that result can be detected by modified DNA sequencing methods, by array-based hybridization, or by application of SNP genotyping assays. Detection of cytosine alleles in these methods indicates a methylation position. For all bisulfite-based procedures, the challenges exist in dealing with the bad DNA quality because of harsh chemical conditions as well as the fact that the DNA strands are not complementary after bisulfite reaction. Furthermore, PCR efficiencies suffer from increased sequence redundancy.

Chemical Tools and Techniques

Nucleic acids research widely relies on optical detection using fluorescence or luminescence detection. Moreover, mass spectrometry and electro-chemical methods are useful depending on the specific application of these detection methods (41, 42). The reaction principles described in detail above and appropriate detection devices can be efficiently combined to tailor analytical methods for the detection of DNA polymorphisms with particular strengths depending on the application, which is illustrated for important SNPs genotyping methods in Fig. 5.
DNA sequencer tools

Capillary sequencers that use sensitive fluorescence detection in a 96-well or a 384-well format provide accurate detection of base sequences in a genome and are currently applied widely for DNA analysis problems. Semi-automated analysis of sequence traces can be accomplished by using software tailored to these applications. The DNA-sequencer can also be applied for SNP genotyping, for example, a 48-plex SNP genotyping method called SNPlex assay (www.appliedbiosystems.com). This technology is based on the allele-specific ligase reaction and the generation of 96 fluorescence-labeled products of 48 SNPs that are separated and detected on a capillary sequencer.

Currently, very promising “second-generation” sequencing technologies are of main interest to detect SNPs, MVPs, or structural variation. Key players in this area are the companies 454 (Roche), Solexa (Illumina), and Aurrencp Personal Genomics (Applied Biosystems). In general, these new sequencing technologies rely on the reduction of genome complexity based on fragmentation and the use of universal linkers and subsequent clonal amplification. The FLX sequencer of 454 uses BEAMing for sample preparation and pyrosequencing, which includes highly parallel luminescence read-out, by charge-coupled cameras devices. The 1G genome analyzer system of Solexa uses clonal amplification, primer extension with reversible terminators that carry removable fluorescence dyes, and highly parallel fluorescence read-out. Applied Biosystems’ SOLiD (Supported Oligo Ligation Detection) sequencing procedure relies on step-wise ligation of short oligonucleotides on arrayed DNA fragments and parallel fluorescence detection. For more details, the reader is referred to the websites of these companies. In these techniques, single molecule amplification is performed on haploid clones. Obtained sequences represent molecular haplotypes that could be assembled for large genomic regions. Current problems of all new sequencing techniques consist in dealing with the short sequence reads and in the large amounts of image data produced per run, which is in the range of several terabytes.

Microarray tools

For genetic studies that involve hundreds of thousands of genetic markers such as SNPs, the highly parallelized methods developed by Affymetrix and Illumina, which were discussed above intensely have become gold standards. However, optimized second-generation sequencing methods might overcome, or at least complement, these microarray techniques in the future. For small-scale applications, the combination of multiplex PCR with primer extension and parallel fluorescence detection on microarrays is useful. This approach is applied in the SNPStream assay from Beckman Coulter that allows the simultaneous generation and detection of about 40 SNPs.

To analyze structural variation efficiently, companies such as NimbleGen and Agilent Technologies offer oligonucleotide microarrays for direct (nonrepresentational) CGH. The resolution of these arrays is currently in the 30-50 kb range, which will probably increase as higher-resolution arrays become available. Dense SNP microarrays that provide genotype information have the additional advantage of being useful in copy number variation analysis that uses hybridization signal intensities. Signal
Another approach for SNP genotyping relies on competitive allele-specific hybridization of primers for PCR amplification, which consists of placing of the 3′-end of one of two primers allele-specifically and directly over the SNP position (see Fig. 3). Matching one of the alternative nucleotides ensures the specificity and the amplification of the PCR, which is observed by accumulating fluorescence of double-strand DNA-binding dyes like SYBR Green I (47). In general, mismatch amplification would be delayed by more than 10 cycles. The commercially available AmpliFluor procedure uses universal oligonucleotides to detect allele-specific amplification (www.chemicon.com). In general, all mentioned PCR-based methods represent simple SNP genotyping approaches in which all reagents are dispensed into a reaction well that is sealed with adhesive film and subjected to thermal cycling.

**Outlook**

In the last decade, several powerful sequencing and genotyping technologies have been developed. Additional refinements will be achieved certainly in the next years. Currently, DNA sequencing is in the focus of technology development. More sensitive and precise detection tools, as well as efficient templates and sequencing reaction principles, are being developed.
Acknowledgment

I would like to thank Tabea Binger for preparing the figures.

References


Further Reading


http://www.home.agilent.com/agilent/home.jsp?cc=US&lc=eng-


http://www.home.agilent.com/agilent/home.jsp?cc=US&lc=eng-


The power of fluorescence spectroscopy in the study of nucleic acids relies on the availability of hundreds of different fluorophores that span the visible spectrum of light and the possibility of using them to label oligonucleotides (short fragments of DNA or RNA obtained by chemical synthesis) specifically. Any sequence can be obtained with one or more fluorophores and fluorophore quenchers attached at preselected positions by linkages with suitable parameters to ensure the optimal properties for the fluorescent oligonucleotides. Fluorescent labeling of long DNA or RNA fragments is also possible by enzymatic methods. Fluorescence-based studies of nucleic acids are widely used in different areas such as basic research, molecular and clinical diagnostics, disease monitoring, therapeutic development, food technology, environmental sciences, and biotechnology. After a few words on biological and fluorescence backgrounds, the main labeling strategies will be developed and examples of biological questions that can be addressed with fluorescent oligonucleotides and the major types of fluorescence spectroscopy techniques will be reported.

Fluorescence-based studies of nucleic acids (NAs) rely on two key contributions of synthetic chemistry: the automatization of oligonucleotide (ON) syntheses (1) and the development of efficient methods for site-specific fluorescent labeling of these ONs (2–7). Many applications are based on the complementary hybridization between short fluorescent synthetic ONs (15–25 nucleotides in length) and the NA to be analyzed. ON sequences can also be chosen for assembling into structures that mimic those found in living biological systems, which provides tools for structural, dynamic, and interaction studies. Because the intrinsic fluorescence emission of the major nucleic bases is too weak to be used for detection, fluorescence-based studies have to rely on synthetically modified nucleotides with more desirable emissive characteristics or exogenous fluorescent labels (Fs). It is a challenge for the chemist to develop ON analogs that do not perturb the samples to be analyzed while providing the desired spectroscopic properties. Fluorescence-based techniques can be used both in solution and in solid-phase applications. In the last few years, the development of fluorescent ONs (FONs) that show a modified emission in the presence of the target has been the focus of intense research (2–7). These new FON probes simplify the analyses in vitro and provide the possibility of applications in living organisms. New strategies for specific enzymatic labeling of large NAs are also in development. Many reviews have been published over the last 3 years on the use of fluorescence to study NAs (2–7). After a brief account on biological and fluorescence backgrounds, this article will focus on the main strategies used to label ONs with Fs and examples will be given of biological questions that can be addressed with FON using the major fluorescence parameters.

Background

Biological background

The sequencing of the human genome provides the full genetic map of the NAs of a human being but lacks information as to how they are involved in the processes within cells. A analysis of the interindividual sequence variations in healthy and diseased people can help to identify and understand inherited or acquired pathologies as well as drug side effects (8). The ability to detect, localize, and quantify the different RNAs (pre-mRNA, mRNA, micro RNA) in living cells and tissues, in real time, will offer tremendous opportunities for biological and disease studies, and it will have a significant impact on medical diagnostics and drug discovery (8, 9). To understand RNA functions in the cell, it is also necessary to monitor their conformational changes directly during pre-mRNA splicing, ribosome assembly, and RNA processing. These different biological questions can be addressed...
Fluorescence Techniques for Nucleic Acids

### Table 1

<table>
<thead>
<tr>
<th>Fluorescent labels</th>
<th>( \lambda_{\text{exc}} ) (nm)</th>
<th>( \lambda_{\text{em}} ) (nm)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xanthene dyes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAM</td>
<td>494</td>
<td>517</td>
<td>3, 4, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25</td>
</tr>
<tr>
<td>TAMRA</td>
<td>555</td>
<td>577</td>
<td>15, 17, 18, 26, 30, 23, 27, 24, 25, 33</td>
</tr>
<tr>
<td>Cyanine dyes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cy3.5</td>
<td>582 (550)</td>
<td>593 (570)</td>
<td>15</td>
</tr>
<tr>
<td>Cy5</td>
<td>655</td>
<td>670</td>
<td>13, 19, 21, 22, 27, 28</td>
</tr>
<tr>
<td>Thiazole orange</td>
<td>480</td>
<td>530</td>
<td>29, 30, 31, 32, 33</td>
</tr>
<tr>
<td>Alexa Fluor 546</td>
<td>556</td>
<td>570</td>
<td>34</td>
</tr>
<tr>
<td>Polycarbocyclic dyes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyrene</td>
<td>240, 345</td>
<td>375, 395, 480</td>
<td>2, 4, 5, 7, 16, 35, 36</td>
</tr>
<tr>
<td>Perylene</td>
<td>444</td>
<td>461, 487</td>
<td>4, 5, 16</td>
</tr>
<tr>
<td>Triphenyl methane dye</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malachite green</td>
<td>610 nm</td>
<td>648 nm</td>
<td>37</td>
</tr>
<tr>
<td>Quenchers</td>
<td>( \lambda_{\text{abs}} ) (nm)</td>
<td></td>
<td>References</td>
</tr>
<tr>
<td>TAMRA</td>
<td>540</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>Dabcyl</td>
<td>474</td>
<td></td>
<td>15, 36, 18, 19, 20, 22, 25</td>
</tr>
<tr>
<td>BHQ-1</td>
<td>534</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>Metal</td>
<td></td>
<td></td>
<td>26</td>
</tr>
</tbody>
</table>

with different ON probe structures conveniently labeled with selected Fs that exploit the main characteristics of fluorescence.

### Fluorescence background

Emission of light, which occurs when an F excited to the singlet state relaxes back to the ground state, can be characterized by parameters such as emission spectrum, fluorescence intensity (enhancement or quenching), fluorescence lifetimes, and anisotropy (10). The fluorescence exhibited by a given F depends on its intrinsic properties, its exposure to solvents, as well as its interaction with other environmental elements. In our applications, the F can interact with the labeled ON. Interactions of the ON with its target sequence or with other nucleic acid binding molecules induce additional contacts of the F leading to quenching or enhancement of the fluorescence. The fluorescence signal can also be modified by the use of more than one F (identical or different) or by a suitable fluorescent/fluorescent quencher (F/Q) pairs linked at convenient positions on the ONs. When two different Fs are in close proximity and the emission spectrum of the donor overlaps the excitation spectrum of the acceptor, energy is transferred from donor F to acceptor F in a distance-dependent manner that follows a nonradiative process (Förster). Another strategy to induce changes of the fluorescent signal consists in the use of modified fluorescent ONs that can adopt specific structures prone to changes during interaction with their targets. When two or more aromatic hydrocarbon molecules such as pyrene are in close proximity, an excited-state pyrene monomer and ground-state monomer can form an excimer state that fluoresces at a longer wavelength than the monomer emission (Birks). Fluorescence resonance energy transfer (FRET) and excimer formation coupled with quenching/unquenching strategies are the most frequently used fluorescence parameters because of the ease of their usage. Fluorescence lifetime and anisotropy measurements require more expertise and sophisticated equipment. Fluorescence detection and measurements can be performed with various instruments (spectrofluorometers, microplate readers, scanners, microscopes, and flow cytometers) that provide different pieces of information. The possibility of detection at different wavelengths associated with the suitable fluorescent labeling of ONs provides the possibility of multiplexing. Finally, progress made in the field of instrumentation associated with the development of very sensitive Fs allows detection at the single-molecule level (11).

### Nucleic Acids Labeling Strategies

Fluorescence labeling of NAs relies on chemical or enzymatic methods. However, intrinsic fluorescence emissions of the five most frequent natural nucleobases is too weak to be used for detection, and fluorescent labeling relies on synthetically modified nucleosides or exogenous Fs incorporated at convenient positions inside the ONs or NA sequences to provide the required spectroscopic properties.

**Different classes of Fs and F quenchers (Qs)**

Fs used to detect and study NAs involve nucleoside analogs, organic dyes, metal complexes, nanoparticles, and quantum dots (2-7). These Fs cover much of the visible spectrum and are available from many suppliers (12-14) (Table 1). Several Qs are also commercially available. A very useful study on F/Q pair efficiencies has been reported (15). Researchers are also continuously developing Fs and Qs with spectroscopically tuned properties to fit the requirements of new applications into different formats and new detection instruments (2-5).
Fluorescence Techniques for Nucleic Acids

Different chemical labeling strategies

Non-specific labeling

Many organic dyes bind to NAs by electrostatic and hydrophobic (including intercalation) interactions that are not specific to the NA sequence. They provide information about the amount of NA available in a sample. These applications mainly include staining of the NA in polyacrylamide and agarose gels as well as DNA quantification during real-time PCR or cell labeling. These compounds, which are available from many suppliers, are used in the presence of the targets is the focus of intense research because their use can simplify the analyses.

Specific labeling

The selective chemical covalent labeling of ONs (DNA and RNA) can be achieved by either the incorporation of modified fluorescent nucleosides or the covalent attachment of Fs (and Qs) at preselected positions on the ON (2-5, 16). Fluorescent nucleotide analogs include either modified bases (isomeric base analogs, purine analogs, extended base analogs, natural bases conjugated to Fs, etc.), or base replacement (polyaromatic Fs...). These analogs can be incorporated via their phosphoramidate or H-phosphonate derivatives during the ON synthesis on solid phase. The phosphoramidite derivatives of the most common organic Fs (fluorescein, rhodamine, a few cyanines), Q (dabcyl...), and nucleosides with pyrene and perylene pendant groups as well as functionalized supports that allow the incorporation of the Fs and Qs at the 3’-end of the ONs can be obtained from different suppliers (12-14). However, the direct incorporation of the Fs during the ON synthesis requires that the Fs or nucleoside analogs can withstand the chemical conditions needed for the deprotection step and that they are soluble in organic solvents. Another strategy consists in the postsynthetic attachment of Fs to various positions of ONs by specific reactions between convenient functional groups incorporated at preselected positions of the ONs and in the Fs. Many Fs with groups that can react with the amino or thiol functionalized linker incorporated into ONs are commercially available (12, 13). The preparation of FONs with new specific properties can be achieved by the refinement of the F structures together with the development of new linkage parameters. Commonly used positions are the 5’- or 3’-termini (3, 4, 15, 17, 18, 20-22, 25, 31, 32, 34-36, 39-41), internucleotidic position (29, 30, 32) and the 2’-position of the sugar residue (3 and 4, positioning the F in the minor groove) (Fig. 2). It is also possible to attach Fs to nucleobases typically at the 5-position of T(U) or C (positioning the F in the major groove), the 7-C carbon of 7-deazaadenine or 7-deazaguanine, and the 8-position of adenine or guanine (2-5, 16, 42 and 43). In addition to the amide, thiourea, and thioether bonds, other linkages such as amine, phosphoramidate, oxime, hydrazide, phosphothioamide, triazole, squarate, or disulfide can be used. The coupling reactions can take place in different solvents, aqueous buffers, or aqueous buffers/organic solvent mixtures depending on the F solubility (17, 18, 29-32, 35, 39). It is also possible to perform the coupling steps directly in methanol in the presence of crown-ether to solubilize the deprotected ON (31).

Multilabeling

The chemical strategy allows the positioning of Fs at many selected sites along the ON. When several Fs are involved, they must be linked to the ON probes in positions that enable the most efficient properties required (FRET (17, 39), excimer (35)). Tunable intensities of the excimer dependent on the pyrene number can also be observed (36).

Hybridization formats

The design of hybridization formats that enable important changes (fluorescence intensity increase or wavelength shifts) in the presence of the targets is the focus of intense research because their use can simplify the analyses in vitro and provide the possibility of applications in living organisms. The most frequently used involve binary probes (35), competitive hybridization probes (18), linear probes with only one F (2, 5, 16, 29, 31), and molecular beacons (MBs) (30-41).
Fluorescence Techniques for Nucleic Acids

Enzymatic labeling

The enzymatic incorporation of modified fluorescent nucleosides or analogs via their triphosphate derivatives is also possible using DNA polymerases (3). One of the most important examples of applying this labeling strategy concerns the methods for NA sequencing (42). More recently, a DNA sequencing method by synthesis on a solid surface has been proposed (43). This method is based on the four nucleotide 5′-triphosphates, which each contain a unique F with a distinct fluorescence emission at the base, modified at their 3′ position to be reversible terminators. This method has yielded the Solexa sequencing technology (44). However, the sequence-specific internal labeling of large DNA requires the development of other strategies as for example the use of DNA methyltransferase associated with a cofactor labeled with a F (45). A new strategy for the site-specific labeling of DNA and RNA is based on the expansion of the genetic alphabet by unnatural base pairs (46).

Selected Fluorescence Techniques and Applications

Different biological questions can be addressed using different ON probe structures conveniently labeled with a selected F that exploits the main characteristics of fluorescence, which is the most sensitive spectroscopic technique (10). Their definitions are widely reported in companion articles and will not be discussed here. The applications can be divided into different series. One application concerns synthetic FONs that mimic the different structures found in living systems and aims at structural, dynamics, and interaction studies in vitro. A closer

In vitro applications

Single-nucleotide polymorphism (SNP) analysis by fluorescence sensing and FRET

Most SNP genotyping assays detectable by fluorescence can be separated into four groups based on molecular mechanisms: primer extension, oligonucleotide ligation, invasive cleavage, and allele-specific hybridization (3, 47). Examples of specific hybridization applications are reported. Methods based on variations of the emitted light of nucleoside analogs inserted in the position opposite the mutation site on the DNA sequence to be analyzed have attracted a great deal of attention in the past few years. Most of these are based on the use of size-expanded base-discriminating fluorescent nucleosides or pyrene-labeled nucleosides. These FONs are attractive because they are easy to prepare and are cheap (2). A genotyping method without amplification of genomic DNA has been reported (34). It is based on the different conformations of a cationic polythiophene F in the presence of single- and double-stranded DNA that results in a fluorescence signal increase in the presence of the latter. Fluorescence can be enhanced via FRET by convenient labeling of

Figure 2 "The most frequently used positions for the covalent attachment of Fs onto the oligonucleotides."
Fluorescence Techniques for Nucleic Acids

Detection of terminal mismatches on DNA duplexes by fluorescence sensing
FON probes labeled at their 5'-ends with thiazole orange can discern perfect duplexes from those that involve a terminal mismatch, as well as those that involve mismatches at the penultimate or last two positions under nonstringent conditions (31).

Increasing the specificity of detection with binary probes
Duplexes of 15-25 base pairs are often too stable to be sensitive to the presence of a mismatched base pair. The division of the ON probes into two parts that lead to good selectivity of the NA recognition has recently been reported. These examples are autoligating probes detectable by FRET (19) or aptamers (37) assembled in the presence of the NA target. The formation of the complex resulted in a great increase in the fluorescence emission of the reporter molecule. A malachite green aptamer can discriminate 41 out of 42 possible single-nucleotide substitutions in a 14-mer DNA target (37).

Detection of DNA abasic site with light-up probes
The detection of an abasic site located at the central position of a 13-base pair duplex by a modified nucleobase inserted at the opposite position was signaled by a 7-fold enhancement of the fluorescence emission when compared with that of the perfectly matched duplex (38).

Detection of mRNA splicing by FRET and single-molecule spectroscopy
Two probes fluorescently labeled to undergo FRET can hybridize to areas flanking the splice sites of an RNA molecule and demonstrate a difference in FRET efficiency between spliced and unspliced mRNA as shown by both bulk solution and single-molecule spectroscopy measurements (33).

Multiplex pathogen detection
The development of a highly accessible and easily adaptable multiplex system for the detection of pathogens remains the ultimate goal in molecular diagnostic laboratories (48). By attaching the MB to bar-coded nanowires, it is possible to detect multiple target sequences with only one F (26), whereas multiplex detection by MBs in a homogeneous solution requires different F/Q pairs (20). Recently, MBs labeled with 6-FAM/BHQ-1 or Cy5/BHQ-2 quencher pairs were linked to microspheres of different sizes by a biotin-streptavidine linkage and used for the multiplex detection by flow cytometry. The average limit of detection for these beads specific for SARS coronavirus, HCV, PIV-3, and RSV was found to be 37 fmol (21).

FISH to size the telomeric sequences
The length of the telomere repeats at individual chromosome ends influences biological functions that range from aging to carcinogenesis (9). Measurements of the telomere length can be

---

Figure 3: Schematic representation of the main hybridization formats. (a) Binary probes. (b) Competitive hybridization probes. (c) Linear probe with one F. (d) Molecular Beacon. (e) Aptamer.
Table 2: A selection of applications of the fluorescent oligonucleotides reported in this article

<table>
<thead>
<tr>
<th>Applications</th>
<th>Hybridization formats</th>
<th>F(s), Q(s), positions</th>
<th>Fluorescent techniques</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP analysis in homogeneous solutions</td>
<td>M onolabeled uniprobes</td>
<td>Fluorescent nucleosides in opposite position to the interrogation position (internal position of duplexes)</td>
<td>Fluorescent sensing</td>
<td>2</td>
</tr>
<tr>
<td>SNP analysis with genomic DNA</td>
<td>M onolabeled uniprobes</td>
<td>F at the 5'-end &amp; free minor groove binder F</td>
<td>FRET</td>
<td>34</td>
</tr>
<tr>
<td>Improving specificity of sequence detection</td>
<td>Binary aptamer</td>
<td>Noncovalent labeling</td>
<td>Fluorescent sensing</td>
<td>37</td>
</tr>
<tr>
<td>Detection of terminal mismatches</td>
<td>M onolabeled uniprobes</td>
<td>F at the 5'-end</td>
<td>Fluorescent sensing</td>
<td>31</td>
</tr>
<tr>
<td>Detection of abasic site</td>
<td>M onolabeled uniprobes</td>
<td>Modified nucleoside in opposite position to the interrogation position (internal position of duplexes)</td>
<td>Fluorescent sensing</td>
<td>38</td>
</tr>
<tr>
<td>Sequencing by synthesis</td>
<td>Templated enzymatic incorporation</td>
<td>F attached to bases of nucleoside triphosphates</td>
<td>Fluorescence sensing, multicolour detection</td>
<td>43</td>
</tr>
<tr>
<td>Pathogen detection (multiplex)</td>
<td>MB on beads</td>
<td>F/Q pairs at both ends</td>
<td>Fluorescence sensing, multicolour detection</td>
<td>20, 21</td>
</tr>
<tr>
<td>Virus localisation in living cells</td>
<td>MB</td>
<td>F/Q pair at both ends</td>
<td>Fluorescence sensing</td>
<td>25</td>
</tr>
<tr>
<td>Detection of RNAs in living cells</td>
<td>Autoligating binary probes, Unprobe</td>
<td>Two F and one Q (at selected positions) M B or MB pair F at internal position</td>
<td>FRET, flow cytometry and confocal microscopy</td>
<td>19, 29, 40-42,</td>
</tr>
<tr>
<td>Visualization of mRNA in cellular extracts</td>
<td>Binary probes</td>
<td>Two Pyrene. One at the 5'-end of one ON probe and the second at the 3'-end of the second ON probe</td>
<td>Excimer detection, Time-resolved microscopy</td>
<td>35</td>
</tr>
</tbody>
</table>

achieved by fluorescence in situ hybridization (FISH). Recently, a flow FISH method has been developed (22).

**Probing nucleic acid structures, dynamics, and interactions**

Studies of the structural dynamics of NA in solution and their complexes are very important for the understanding of their functions in living organisms (4, 49). This research can be easily performed by using fluorescent base analogs that resemble natural ones with respect to their dimensions and hydrogen binding patterns (5). Among them, 2-aminopurine and pteridines have been used in numerous applications (5, 16). Alternatively, the covalent attachment of Fs at convenient positions of the NA to be analyzed can be performed to allow FRET to take
Detection of viral sequences by fluorescence sensing

The direct visualization of poliovirus NA in living host cells was achieved with MB, involving TAMRA as F and Dabcyl as Q, targeting the viral plus-strand RNA (25). The poliovirus plus strand was observed to display different distribution patterns at different post-infection time points in living vero cells.

The visualization of RNAs in living cells using FRET, excimer and light-up probes

Competitive hybridization probes (18), binary hybridization probes (35), MBs (40), light-up probes (29), and quenched auto-ligation probes (QUAL) (19) can be used. Among them, the quenched probe strategy offers the advantage of a decreased background signal in the absence of the RNA target. By using QUAL probes, mRNAs as well as 28S ribosomal RNA were detected in HL-60 cells by FC and were visible by confocal microscopy (19). A cell membrane-permeant couple of MBs designed to produce FRET in the presence of the target sequence was successfully used to detect human GAPDH and surviving mRNAs in living cells (41). Oligo-α-thymidylates that involve thiazole orange at the internucleotidic position were used to visualize the mRNAs intracellular distribution in HOS cells (29). The accurate detection of mRNA in living cells with F-ON probes requires homogenous distribution of the F-ONs throughout the cells. To make visible the cytoplasmic mRNAs in living cells, MBs were linked to bRNA (53).

Summary

In this article, we have attempted to show, through a few examples, the power of fluorescence strategies based on fluorescent oligonucleotides to study nucleic acids. This process has been made possible through chemistry that makes the synthesis of fluorescent oligonucleotides possible with the required properties suitable for experiments in varied formats including living cell experiments. We apologize to authors whose work has inadvertently not been cited. It is our hope that this article will stimulate further reading and help the design of new strategies.

References


Profiling microRNA expression in biological samples

Because of the importance of micro-RNAs in gene regulation in both plants and animals, there is a need for profiling their expression in biological samples (51). Multiplexed detection methods in solution and in solid phase have been developed (52).
Fluorescence Techniques for Nucleic Acids


Fluorescence Techniques for Nucleic Acids


Further Reading


See Also

 Fluorescence in Living Systems: Overview of Applications in Chemical Biology
 Fluorescent Spectroscopy: Overview of Applications in Chemical Biology
 Tools to Detect DNA Polymorphisms (SNPs focus)
 Peptide Nucleic Acids (PNAs)
 Solid-Phase Synthesis of Biomolecules
 Single Molecule Studies of Biomolecules
Regulation of gene expression plays a central role in controlling and shaping the functions of a cell. Tools for quantification of the expression of individual genes have been available for years; but over the past decade, development of the microarray technology and accompanying bioinformatics tools has made it possible to generate comprehensive overviews of the transcriptional events in both diseased and normal cells and tissues. This review covers various approaches for transcript profiling from single genes to more global analyzes approaches. In addition, a detailed description of different microarray-based technology platforms will be provided, which all enable an essentially genome-wide characterization of the transcript levels. We also discuss the current challenges and future trends within transcriptional profiling, and we introduce briefly the next generation DNA sequencing technology that will enable a more detailed description of the entire transcriptome, which includes various small RNA species (e.g., microRNA) and other noncoding transcripts.

Transcription is the essential cellular and biochemical process that links the genetic information encoded in the genome (DNA) to the functionally active macromolecules, proteins, which carry out most tasks in a cell. Transcription generates an RNA transcript, and the process is regulated at multiple levels, which includes both synthesis and degradation of the transcript. To understand the complexity of this mechanism and to connect the process to the phenotype of an organism, we need to measure the transcript levels accurately under various situations and samples. In many cases, it is advantageous to do this in a genome-wide and unbiased manner.

Biologic Background

The central dogma of molecular biology states that genetic information flows from genes, via RNA, to proteins. In this flow of information, messenger RNA (mRNA) is generated in a process called transcription and is subsequently processed to yield a mature transcript. The transcriptome is the combined set of all transcripts present in a cell at a certain time, but it should be noted that mRNA is only a minor component of the entire RNA population of a cell. The cell contains highly abundant ribosomal RNA, transfer RNA, microRNA, small nuclear RNA, small nuclear RNA, and additional rare types of RNA, but the focus of this review is on the protein-coding mRNA transcripts. The maturation of these protein-coding transcripts consists of several distinct steps, all of which are regulated specifically (Fig. 1)

For years, it was assumed that the rate of RNA synthesis was the rate-limiting step that controls indirectly the amount of protein synthesized. However, this simplistic view has been replaced by results clearly showing that mRNA and protein levels do not always correlate fully, and that extensive regulation of mRNA transcript processing and availability for translation occurs.

A typical human cell contains approximately 300,000-500,000 transcripts, and most genes are transcribed at low to moderate level, whereas only a small number of genes encode for a large number of transcripts. The underlying assumption of most transcript profiling studies is that the pattern of mRNA transcripts in a cell at a certain time can be used to explain the phenotype of the cell and the activities within the cell. For
example, by comparing a cancer cell with a normal active pathway, signaling mechanisms in the cancer cells can be identified; in the next step, we can attempt to alter these pathways in treatment of the cancer.

Interestingly, another level of complexity has been added by research during the last few years that has shown that only a fraction of the transcribed loci generate protein-coding transcripts and that almost the entire genome is transcribed (1, 2). Even though these transcripts do not code for a protein, they may be functional in other ways; therefore, they constitute interesting target transcripts that may be analyzed using the same approaches as the protein-coding transcripts. This review will focus on analysis of the protein-coding transcripts, but it should be kept in mind that the same tools can be used to analyze of many other types of transcripts and RNA molecules as well.

This review is sectioned in the following way; a brief overview of the various approaches for transcript profiling is given in the section on Tools and techniques. In the section on Microarray-based transcriptional profiling, we provide a detailed description of the microarray technology, which enables essentially a genome-wide characterization of the transcript levels. The last chapter discusses the current challenges and future trends within transcriptional profiling.

Tools and Techniques

Essentially every cell in an organism is, at any given time, transcribing thousands of its genes in various quantities. As described in the previous section, the amount of an mRNA transcript is regulated, and an interest exists, from both basic science and clinical perspectives, in quantifying the transcripts levels accurately. The available techniques can be divided broadly into either gene-by-gene methods (see sections on Northern blot and Quantitative real-time RT-PCR) or global methods (see sections

![Diagram of gene expression processes](image-url)
on Sequencing-based transcriptional profiling and onward). The gene-by-gene methods aim at high-accuracy quantification of a low number of transcripts, whereas the global methods aim at a highly parallelized quantification of many transcripts, often even genome-wide. The focus of this review is on the microarray technology, which is the most widely used technology for genome-wide transcriptional profiling (see the section on Microarray transcriptional profiling). The other techniques are reviewed briefly below.

Northern blot

The northern blot technique allows quantification and size determination of a transcript in a complex mixture (e.g., the entire transcriptome) first by separating the transcripts by denaturing agarose gel electrophoresis, followed by a transfer to a membrane strip and hybridization with a labeled probe. Historically, Northern blotting has been used widely, but during the recent years, a shift toward more sensitive methods has taken place. These alternative methods are often less sensitive to RNA degradation, and they have a wider dynamic range.

Quantitative real-time RT-PCR

Quantitative real-time reverse-transcription PCR (qRT-PCR), which is the current “gold standard” for high-accuracy transcript profiling, provides superior sensitivity for analysis of transcript levels compared with other methods. A complex mixture of total RNA is converted to cDNA using reverse transcriptase with either random or gene-specific priming. Next, a 100–200 bp fragment is PCR-amplified using gene-specific primers that often target two different exons of the transcript, and the accumulation of the amplicons is monitored in real-time using a fluorophore that either specifically targets the amplicon and the accumulation of the amplicons is monitored in real-time or any double-stranded DNA. Theoretically, during an exponential phase of the amplification, each PCR cycle doubles the amount of product, which in log2 scale corresponds with a linear increase. Extrapolation of this linear increase back to the base-line level provides an estimate of the initial starting amount of mRNA.

Using qRT-PCR has many advantages, which makes it the method of choice for high accuracy—but low throughput—gene expression analysis for the following reasons: 1) it offers, at its best, a dynamic range of 7–8 log orders of magnitude; 2) it can achieve single-copy detection; 3) it can be carried out in one step; 4) it has low coefficients of variation facilitating detection of small differences between samples; and 5) design of specific amplicons allows for discrimination between similar transcripts, such as transcript isoforms or different gene family members. The drawback of the method is that genes must be targeted individually. Therefore, a large-scale approach is not feasible.

Sequencing-based transcriptional profiling

Recently, several platforms for high-throughput sequencing have been developed (see section on Next generation of sequence technologies), and these platforms offer an impressive improvement of the number of bases sequenced compared with previous technologies. However, use of standard capillary sequencing has also provided means for large-scale sequencing-based transcript profiling, albeit at lower throughput levels. These different sequencing-based methods are outlined briefly below and are discussed in more detail elsewhere.

Expressed sequence tag analysis

Expressed sequence tag (EST) sequencing generates random, 200–900 bp single-pass sequences of cDNA clones. The initial purpose of these sequences was to facilitate gene detection, but the technique has been used subsequently to estimate gene expression levels. The underlying assumption is that the EST sequences are generated randomly, and hence the EST counts correspond with the transcript’s abundance in the original sample. The main drawback is the low throughput (number of counts) caused by the high data generation costs (library generation and sequencing). Given that gene expression levels follow a distribution with many genes expressed at low levels and a few genes expressed at high levels, transcript profiling that uses EST technology enables reliable detection only of a small number of genes expressed at moderate or high levels.

Serial analysis of gene expression

Serial analysis of gene expression (SAGE) (10) was the first approach to provide large-scale absolute estimates of transcript frequencies, and it relies on a biotinylated primer, streptavidin-coated beads, and type IIIs restriction endonucleases that cleave outside the recognition site to generate short tags of each transcript. The tags are concatenated and sequenced using standard sequencing technology to derive a digital representation of the transcript frequencies (counts). In the original approach, SAGE was used to isolate approximately 14-bp 3′ tags, but the method has been later modified to allow for isolation of 5′ tags and of longer 26-bp tags.

Cap analysis of gene expression

Cap analysis of gene expression (CAGE) (11) uses 5′ trapping methods to isolate full-length cDNA s selectively and generates 20-bp tags from these. After isolation, the tags are ligated to yield ~700 bp of concatenated sequence, which is cloned into a vector and sequenced. The first-strand synthesis can be primed with random primers, which allows for analysis of polyA-negative transcripts. Recently, CAGE has been used for large-scale transcription start site mapping (12).

Massive parallel signature sequencing

In massive parallel signature sequencing (13), 3′ sequences of each transcript are isolated using biotinylated primer in the cDNA synthesis and are cleaved with a restriction enzyme that generates a cohesive end. Next, the 3′ signature sequences are...
ligated into specifically designed plasmid vectors that contain 32-nt oligonucleotide tags (in total 16.8 × 10^9 different tags), and amplified using PCR. Use of a large number of tags provides a unique tag for each 3' signature sequence, which is subsequently coupled to a 5-µm microbead. Each bead contains one type of capture tag complementary to one of the 32-nt oligonucleotide tags. Next, the captured signature sequences are sequenced on beads to yield 16–20 nt signature tags, which are counted to derive a global estimate of transcript levels.

Next generation of sequencing technologies

The technology development driven by the race toward low-cost sequencing of the entire human genome has provided the research community with new ultra-high-throughput DNA sequencers, which in the near future may open up for sequencing-based analysis to generate a global overview of the transcriptome. The new sequencing approaches include bead-based Genome Sequencer pyrosequencing instrument (454 Life Sciences, Branford, CT, USA) that produces more than 100 million bases of sequence per run, Solexa Clonal Single Molecule Array (Illumina, San Diego, CA, USA) producing up to one billion bases of sequence per run, and also the SOLiD sequencing chemistry (Applied Biosystems, Foster City, CA, USA) producing read lengths of almost 40 nt.

Common for all these techniques is that they are based on random fragmentation of the sample to be analyzed (e.g., a genome), ligation of adapter molecules to both ends of the fragment, and an amplification step (e.g., emulsion PCR for Genome Sequencer and the SOLiD technologies) on a solid-phase surface. This step is followed by the actual sequence reading step that is based on detection of fluorescence (Solexa and SOLiD technologies) or emitted light (Genome Sequencer) using a CCD camera. The large amount of data generated is a consequence of millions to tens of billions of sequence reads (10–300,000 for Genome Sequencer, tens of millions for SOLiD and Solexa). The read lengths are however significantly longer for the Genome Sequencer (200–300 nt) than for the other two technologies (25–35 nt). All platforms are expected to be improved in the near future.

In addition to de novo and whole-genome resequencing, all these technologies open up also for large-scale transcriptome analysis, possibly in combination with the approaches described above. For example, the Genome Sequencer platform has been used in several transcriptome analysis studies (http://www.454.com/news-events/publications.asp?cat=4).

Microarray technology

Since the first publications in 1990s (14–16), the use of the microarray technology for transcriptional profiling has become widespread with more than 25,000 publications in the NCBI’s PubMed database. In addition, the number of microarray hybridizations in two public data repositories (see section on Data sharing) is rapidly increasing, already approaching 250,000 hybridizations. Finally, today an entire industry exists that provides resources (e.g., arrays and reagents) and analysis support for the microarray community, which brings the technology within reach of essentially every researcher.

The term “microarray” refers to a solid-phase support on which multiple capture probes have been immobilized in an ordered fashion, and which participate in a capture reaction of a specific target molecule. Microarrays are used commonly to measure levels of mRNA transcripts, microRNAs, and proteins, but also to analyze characteristics of genomes (e.g., SNPs, gene copy number changes, and larger chromosomal gains and deletions). Since their first use in the mid-1990s, the different microarray platforms have been modified and improved extensively. The fundamental underlying advantage of the technique is that a simultaneous, highly parallelized measurement of thousands of different targets is possible; in some cases allowing for analysis of all known protein-coding transcripts.

A detailed description of the technology is provided in the section on Microarray-based transcriptional profiling.

In situ hybridizations

In situ hybridizations are based on labeled probes that base pair and identify target transcripts in fixed samples. This technology is the only approach to provide a snapshot of transcripts and their cellular localization. The probe is labeled either using a radioactively isotope, a fluorophore, or an antigen. After wash, the probes that hybridized to their target transcripts can be detected using autoradiography, fluorescence microscopy, or immunohistochemistry, respectively. Use of fluorescence or antigens allows for use of probes for multiple transcripts simultaneously, which allows identification of genes with overlapping expression patterns.

The in situ hybridization technique is used in several large-scale efforts to provide a comprehensive picture of the gene expression pattern in for example mouse embryos and brain (17).

Microarray-Based Transcriptional Profiling

A typical transcript analysis starts with isolation of total RNA, followed by cDNA synthesis and labeling. Next, the purified and labeled cDNA is applied onto a microarray that contains thousands of immobilized probes, hybridized, washed and scanned, and the signal for every probe estimated and analyzed (Fig. 2).

The aim of a typical microarray-based transcriptional profiling experiment is to identify target genes for downstream validation experiments, for example to identify genes that are expressed differentially after treatment with a certain compound. Likewise, the array technology can be used to verify a hypothesis, for example to verify that a compound does induce a certain expected effect. Furthermore, microarrays are used widely in various classification studies in which an initial set of samples with known “labels” (e.g., type of leukemia) are profiled, followed by a profiling of a different set of samples with unknown labels (e.g., patient biopsies) and assigning these into the previously identified classes based on their expression profile.
Typical workflow of a microarray experiment

This section describes the use of the microarray technology for transcriptional profiling. The section includes a summary of the experimental procedures (section on Experimental design), the most widely used array platforms (section on Platforms for Gene Expression Analysis), and briefly summarizes the data analysis steps, the software that can be used, and the public microarray data repositories (sections on Analysis of microarray data and Data sharing).

Experimental design

Microarray experiments should be designed to be maximally informative given a certain amount of resources, and they need to answer the primary questions of the experiment. Extensive reviews on experimental design are available elsewhere (18–20). The consequence of a nonoptimal design ranges from loss of statistical power and an increased number of false negatives to inability to answer the primary scientific question of the experiment. The number of available arrays is often determined by financial resources; therefore, one of the most important questions is to determine how to allocate the different samples to a given set of arrays (hybridization scheme) and what to replicate (e.g., biologic samples or hybridizations). In addition, selection of the array platform and the target preparation approach must be considered.

For Affymetrix (Santa Clara, CA) (see below) and for other single-channel experiments, the hybridization scheme is straightforward, but for two-channel platforms, the allocation of samples to arrays is important to prioritize the primary scientific question of the study. Furthermore, balanced designs should be used so that treatments are not confounded with technical issues such as dye assignments, batch of slides, or day of hybridization. Replication is carried out to control the three levels of variation in an experiment: biologic variation (e.g., differences between animals), technical variation (e.g., differences caused by the RNA amplification), and measurement error (e.g., uneven hybridizations). Statistical testing can be carried out on any of these levels, but interpretations of the results differ; is the purpose to analyze the difference between two mice (inference at the level of technical replicates), or is the purpose to generalize the results and to draw conclusions at the level of a population (inference at the level of biologic replicates)? It can be safely assumed that the purpose of most, if not all, experiments is to analyze differences at the population level; hence, biologic replication is essential.

Sample preparation

Direct analysis of complex samples such as unfractionated tissue is often of little value because of cellular heterogeneity. Consider bulk brain for example, which is a mixture of hundreds of different cell types. Unless a specific cell type is selected prior to mRNA extraction, the obtained gene expression profile will be a weighted average of the total gene expression of all different cell types. To enrich for a certain cell type or to obtain homogeneous samples several different approaches have been used. First, experiments can be designed to include sampling shortly after perturbation, which allows for monitoring of early events before secondary changes accumulate. Second, synchronized cell cultures can be used, which allows for analysis of cell cycle phase-specific gene expression patterns (21). Third, fluorescence-activated cell sorting that uses one or multiple fluorophore-conjugated antibodies to identify cells that express a combination of different cell-surface molecules provides a rapid and sensitive cell fractionation assay. Finally, laser-capture microdissection based on a microscopic evaluation of the sample and use of a computer-controlled laser to excise and to isolate specific cells into a collection vessel can be used.
Transcript Profiling: Tools for Transcript Profiling

Target preparation

Depending on the array platform, a labeled sample that originates from 1–20 μg of total RNA is required for each hybridization. This corresponds to 0.1–2 million cells (assuming 10 pg of total RNA per cell), which is obtainable in cell culturing studies. However, use of various sample preparation methods (see previous section) compromises the yield; hence, a target amplification method is often required. Linear T7-based in vitro transcription (IVT) typically yields 300–1000-fold amplification, and a higher amplification can be obtained by performing up to three consecutive rounds of amplification. A double-stranded DNA template that contains a T7 RNA polymerase binding site in the 5’ end is synthesized using the mRNA as template, and subsequently transcribed in a 3–12 hour isothermal IVT reaction during which the amplified accumulates RNA linearly. It should be noted that all samples used on Affymetrix arrays are subjected to IVT amplification. PCR-based exponential amplification methods are diverse and typically are based on ligation of linker sequences to both ends of double-stranded cDNA, followed by a limited number of PCR cycles to yield double-stranded DNA. Generally, these methods are assumed to introduce bias to the data because of transcript-length dependent or base composition differences in amplification efficiencies. To circumvent this problem, approaches have been developed that restrict the length of the template and make it more uniform (24, 25). The advantage of PCR-based methods over linear IVT methods is that a much faster and a higher amplification is achievable (26). Comprehensive literature reviews of target amplification approaches are available (27, 28).

Labeling using fluorophores is typically carried out in a prehybridization fashion for the spotted array platforms, and in a posthybridization manner for the Affymetrix platform. For Affymetrix arrays, biotin-modified nucleotides are incorporated into the aRNA, during the IVT step, and the dye coupling is carried out after hybridization using phycocerythrin-streptavidin and biotinylated anti-streptavidin-antibody conjugates. For spotted arrays, target labeling and incorporation of the fluorophore can be carried out either directly (the fluorophore is attached to the nitrogenuous base of one of the nucleotides) or indirectly (the fluorophore is attached to modified nucleotides after cDNA synthesis using a chemical coupling). Direct labeling is often affected by incorporation difficulties and differences in efficiency between the dyes. Indirect labeling avoids these problems by using only one type of modified nucleotide in the cDNA synthesis. Alternatively, an emerging approach, based on labeled platinum conjugates, can be used to label the RNA or DNA chemically (28). The possibility to omit all enzymatic steps makes this approach interesting and promising.

Hybridization

In the hybridization of the labeled target to the immobilized probe and subsequent washing, two opposing forces need to be balanced—too stringent conditions develop low signals, whereas too unspecific hybridizations yield compressed ratios with little differential expression. An extensive analysis of the conditions is beyond the scope of this review, but the main parameters that must be considered are probe length, hybridizations buffer composition, hybridization temperature, duration of hybridization, mixing, and wash stringency.

Scanning and image analysis

Typically, spotted microarrays are scanned at 5- or 10-μm resolution one channel at a time, which generates two 25-100 Mb 16-bit images. To facilitate the image analysis and visualization, the two images are overlaid to generate one 24-bit RGB pseudo-color image with the red, green, and yellow spots associated commonly with microarray data. Affymetrix arrays are scanned using only one wavelength. After scanning, and irrespective of array platform, the purpose of the image analysis step is to separate foreground and background pixels, to derive an estimate of the gene expression level for each feature, and to calculate various intensity and quality-control parameters (30). The background intensity is considered to represent the contribution of nonspecific hybridization to the slide surface and to the immobilized DNA. Various correction approaches have been described to account for this binding, but it should be noted that the subtraction issue is controversial, and the effect may vary depending on the dataset being analyzed (31).

Platforms for gene expression analysis

Multiple platforms are available for high-throughput, microarray-based, genome-wide transcriptional profiling. They mainly differ in the type of probe attached to the surface, the number of target samples that can be hybridized on each array, and the principal expression measurement (ratio for two-channel arrays and absolute level estimate for single-channel experiments). In addition, the target labeling and hybridization, image analysis, and initial low-level data analysis aspects often differ. At the high-level data analysis stage (where biological inference is sought), the data analyzes for the different array platforms converge, and the approaches and the interpretation of results generated are similar.

Several different microarray platforms have been compared in the Microarray quality control project using commercial RNA samples. This comprehensive study of reproducibility and variability, both within and between different microarray platforms, stated clearly that the data obtained using the microarray platform is generally of high quality. The study also included a large-scale analysis of the expression levels by quantitative RT-PCR, and these results allowed first time the microarray technology to be benchmarked against the current, “golden standard” technique (32).

The next section describes briefly the platforms that are used most widely for transcript profiling, starting with the two-channel platforms.

Spotted cDNA arrays

The relatively low cost of cDNA array production and the access to thousands of EST clones in the freezers in many laboratories and the commercial distribution of EST clone collections propelled the early development and popularity of the cDNA arrays in the late 1990s (15). The arrays are generated through PCK-amplification of cloned 200–4000 bp insert sequences using vector-specific primers. The double-stranded amplicons are


then purified (ethanol precipitation or filter plates), printed, and immobilized on coated glass slides. To avoid plate-handling errors, rigorous quality control steps, which include complete, partial, or random resequencing, and aarsene gel electrophoresis analysis of the purified clones is advantageous, but it is labor-intensive and costly.

The general advantages of the spotted cDNA arrays include the following: 1) the low cost of arrays that allow for design of experiments with extensive replication, 2) the possibility to use two-color detection and thereby reducing the number of arrays, 3) large-scale clone collections are widely available from multiple sources, 4) they are compatible with most target amplification protocols, and 5) they include established laboratory protocols. The drawbacks (many of which are shared with oligonucleotides and Affymetrix arrays, see below) include: 1) unspecific target-probe interaction because of long probes, 2) false negatives caused by drop-outs during probe preparation or array printing, 3) batch-to-batch variability in array production, 4) incomplete transcriptome coverage, 5) uncertainty over which region or isoform of a transcript is targeted with a given probe (the complete probe sequence is rarely known), 6) difficulties in maintaining high-quality probe collections (avoidance of evaporation, well-to-well contamination, plate rotation, etc.), and 7) confounded measurement of sense and antisense transcripts.

Spotted long-oligonucleotide arrays

Spotted arrays with (50–90 nt) oligonucleotides have been available for several years, and they offer a higher specificity than is achievable using the cDNA arrays. Using publicly available genome sequences, oligonucleotides are designed in silico for each gene in a genome. The melting temperatures are also taken into account to achieve uniform hybridization conditions. Typically, the oligonucleotides are bought presynthesized and are dissolved in appropriate printing buffer and printed. Use of presynthesized oligonucleotides offers several advantages. First, probes can be generated for any organism given that its genome sequence and gene predictions are available. Second, the probes are targeted to specific regions of genes, which allows for some differentiation of splice variants. Third, clone handling is reduced, minimizing the risk for plate or clone handling errors. Fourth, replacement plates are easy to obtain. Last, the probes are designed to have the same sense as the mRNA; hence, they are complementary to the labeled cDNA generated from the mRNA, and a confounded measurement between sense and antisense strands is avoided. In addition to many of the drawbacks listed for cDNA arrays (see points 2, 3, 4 and 6 in cDNA array section), the initial purchase investment for oligonucleotide collections is substantial.

Affymetrix genechip arrays

Affymetrix arrays are one of the most widely used platforms for transcriptional profiling. The probes are designed in silico and are synthesized directly on the array using photolithography (33, 34). Each gene transcript is interrogated by a “probe set,” which constitutes 11–20 different 25-mer perfect match (PM) probes and their corresponding mismatch (MM) probes. The MM probes differ from their PM probes by one mismatched base in the central position that functions to destabilize the probe-to-target complementarity. Depending on the data processing approach, the intensities from the MM probes can be used to correct the signal from the PM probes, but other approaches are available (35).

Several advantages exist with Affymetrix arrays. First, arrays and operating procedures are standardized, which allows for direct comparison of data between projects and laboratories. Second, the direct synthesis of the probes on the array avoids problems with plate handling and ensures that the batch-to-batch variability is minimized. Three, small feature sizes yield dense arrays, which allows for genome-wide transcriptional profiling with multiple probes per gene. Fourth, probes are single-stranded, and nonconfounded measurements between overlapping transcripts are obtained. The drawbacks include the inflexibility in probe content caused by the initially high production costs, and sample preparation always includes linear amplification (i.e., extra enzymatic steps that may introduce bias into the results).

Recently, Affymetrix also launched their exon arrays (i.e., arrays that use multiple probes to target essentially every exon of each transcript variant). This platform is the most detailed for analysis of gene expression levels, and it allows for identification of alternative splicing in addition to the standard genome-wide transcriptional analysis.

Affymetrix provides arrays that not only analyze expression levels of genes, but also interrogate the entire genome for transcriptional activity. These arrays are termed “tiling arrays,” and they contain probes that are more or less evenly spaced (approximately 35 bp) along the genome. The arrays have been used in several studies to identify extensive transcriptional activity from regions that were not considered to encode for genes. Typically, these arrays are used to provide a genome-wide transcriptional mapping or are used in chromatin immunoprecipitation studies.

Illumina’s beadarray platform

The BeadArray technology (Illumina, San Diego, CA, USA) allows for genome-wide transcript profiling using a bead-based, high-density microarray platform. The 3–5 μm beads are coated with hundreds of thousands of copies of a specific capture probe, and these beads are assembled into an array of beads on either a fiber-optic bundle substrate or planar silica slides with etched microwells. For transcript profiling arrays, the oligonucleotides are 50 nt long. Multiple beads per target transcript are used to generate a large redundancy in the data, which increases the precision of the final measurement. In addition, quality control steps in the oligonucleotide synthesis and bead attachment process ensures that the frequency of dropouts is kept low (36). In addition to transcript profiling, the BeadArray platform can also be used to analyze DNA, for example, in a comparative genomic hybridization experiment.

NimbleGen gene expression platform

NimbleGen arrays (NimbleGen Systems, Inc., Madison, WI) are similar to the Affymetrix arrays in that they have been produced using a light-directed probe synthesis approach, but they use computer-controlled micromirrors instead of masks to
The generation of the raw data is followed by an extensive analysis of the data. A detailed description of this analysis is beyond the scope of this review, but it is available in a recent review (37). Briefly, in most cases the analysis is divided into two sections: pre-processing of the data (low-level data processing) and subsequent data mining (high-level data processing). The purpose of the pre-processing is to identify and to correct for systematic and nonsystematic technical artifacts and other nonbiologic bias in the data. Steps that are typically included are correction of the hybridization background levels, exclusion of nonreliable data (e.g., dust particles on hybridization surface), log2-transformation of the data, and normalization to account for technical intra- and inter-slide differences. The high-level data mining include steps in which biologic inference are sought for; example, identification of differentially expressed genes using a moderated t-test (38, 39), identification of enriched (overrepresented) biologic themes from the Gene Ontology database (40), clustering analyzes to identify coregulated genes and patterns in the data, and various other dimension reduction and classification tools (41, 42).

A typical microarray experiment generates large quantities of data, and to analyze these data efficiently, both commercial and open-source software solutions have been developed. The open-source software have gained extensively in popularity, mainly because the availability of R packages that provide tools for analysis steps described in publications, a large user community that improves existing functions, the possibility to modify and to automate analysis steps, and the fact that the software is available at no cost. R is a programming language and an environment for statistical computing and graphics (43), and its functionality can be extended by packages such as the Bioconductor project (44), which provides a comprehensive collection of tools for all steps of microarray data analysis. TM 4 is another open-source software suite that provides an easy-to-use, java-based graphical interface (45).

Data sharing

To facilitate comparisons between experiments and especially meta-analyses, two raw data storage and exchange repositories are available: ArrayExpress, which is run by European Bioinformatics Institute (Cambridge, UK) (46), and Gene Expression Omnibus (GEO), which is run by the National Center for Biotechnology Information (Bethesda, MD) (47). Both accept submissions that fulfill the Minimum Information About a Microarray Experiment (MIAME) standards (48). The purpose of the MIAME standards is to ensure that all essential information regarding the experiment underlying a publication is available, and that the interpretation of the results can be carried out properly. An increasing number of journals are also requiring the data to be available publicly in the repositories to publish the results. Hence, it is not surprising that these repositories are widely used; ArrayExpress contains more than 2,500 experiments and over 80,000 hybridizations, whereas GEO has passed 6,000 experiments and 160,000 hybridizations (August 2007).

Practical Aspects and Future Trends

Since the early years of microarray technology, the field has matured rapidly and the technology development has resulted in numerous different platforms for genome-wide transcript profiling. With the availability of the genome sequence and gene predictions of an organism, probes can be designed easily, and arrays can be generated. In parallel with the development of the experimental platforms, tools and the statistical framework for the analysis have been developed strongly. Today, multiple excellent tools are available as both commercial and open-source software packages. The sheer magnitude of the data has also required development of data management systems that can efficiently store, back-up, and process large data sets in an efficient way. A together, during the last 15 years, the array technology has been shown multiple times to yield an accurate description of the transcriptional status of a large number of genes in a cell, but it has to be remembered that our understanding of the transcription process, the transcriptome and the central dogma itself is rapidly changing and that the transcript profiling technologies irrespective of platform need to adjust to this changing picture where new details are emerging continuously.

The new information gained on transcription of nonprotein-coding transcripts has changed the view of transcription in many ways during the recent years, and it has changed the requirements for transcript profiling. Future technology development will have to deal with improved separation of transcript isoforms, for example, to separate coding and noncoding (e.g., partial) versions of a transcript, and transcript strand assignment (i.e., sense or antisense transcription). In addition, the identification of transcription start site may be useful to understand in more detail how transcription is regulated. Furthermore, the sensitivity of the methods needs to be improved to measure transcripts of low abundance accurately.

The recently published genome of an individual human (49), together with development of the next generation of sequencing technologies, prepares ground for an exciting alternative to microarray-based transcription profiling. These new technologies have a capacity that is unmatched by the older sequencing approaches, and with extended read lengths additional information of the transcripts can be obtained, such as detailed
Transcript Profiling, Tools for

transcript isoform information. In combination with the information obtained on individual genomes, exciting possibilities to combine gene copy number variation and gene expression levels may emerge in the near future. Hence, the sequencing-based methods for transcript profiling not only will challenge the microarray-based transcript profiling methods, but also will complement them with additional detailed information of the transcripts and the genome from which they originate.

References

19. Glonem GF, Solomion PJ. Factorial and time course designs for cDNA microarray experiments. Biostatistics 2004;5:89-111.
Transcript Profiling: Tools for


See Also

Array-Based Techniques for Proteins
Array-Based Tools for Nucleic Acids
Nucleic Acids, Design and Engineering of Oligonucleotide Arrays to Monitor Polymorphisms
Electron Paramagnetic Resonance in Enzymology

Aimin Liu, Department of Chemistry, Georgia State University, Atlanta, Georgia

doi: 10.1002/9780470048672.wecb668

Electron paramagnetic resonance (EPR) is a versatile tool for chemical biology research. A remarkable wealth of information about the mechanistic enzymology has been obtained by EPR experiments. As an analytical method, EPR can obtain these two useful pieces of information in a study of enzymatic reaction mechanisms: 1) detection and complete description of free radicals and 2) characterization of the electronic structures of paramagnetic metal ions and their response to changes of the protein environment or substrate binding. The first feature has spurred the development of the new field known as radical enzymology; the second has enhanced the understanding of the mechanisms of metalloenzyme action. The potential of EPR spectroscopy to serve as an important biophysical tool for the future development of enzymology is highlighted by several selected examples.

Introduction

This article does not intend to be a survey of electron paramagnetic resonance (EPR) spectroscopy in chemical biology, but rather the focus is to provide an entry-level introduction to chemists who are interested in the research opportunities that EPR may provide for enzymology studies. The principles of this technique will be briefly described, as well as the breadth of the technique’s applicability to free radicals and metallo_centers in enzymology. Potential applications in the future will also be outlined.

History

EPR was first applied to biological materials in 1954 (1), 10 years after the discovery of an experimental approach to detecting electron spin resonance by the Russian physicist Zavoisky in 1944 (2, 3). This technique became useful for studying enzymes when a spectrometer that was capable of detecting about $10^{-10}$ moles of unpaired (nonbonding or “free”) electron spins in samples that contained about 0.15 milliliter of liquid water was developed by Townsend and colleagues in 1957 (4). The first successful application of this technique to elucidating the structure of substrate radicals in enzymatic reactions was in 1958 (5–7). In the ensuing 50 years, numerous seminal and classical contributions have been made to the development of enzymology by scientists employing EPR spectroscopy analytically.

Application

It is an advantage to analyze EPR samples of aqueous solution directly because the buffered solution is the most physiologically relevant condition. However, EPR samples are not restricted to samples in the solution state, gaseous and solid samples and single crystals can also be analyzed directly by EPR spectroscopy. The only absolute requirement for preparing an EPR sample is that it must contain unpaired electron spins or that such status can be achieved before EPR measurements by oxidation/reduction reactions. EPR spectroscopy can determine unambiguously free radicals that are present in the sample or radicals produced by oxidation or irradiation with light, X-rays, and γ-rays. Because paramagnetic transition metals contain unpaired electrons, EPR spectroscopy can determine the detailed geometric and electronic environment of the metal centers. This technique can also indicate the degree of molecular motion in a sample with unpaired electrons and analyze the populations in a heterogeneous sample that has two or more paramagnetic species. Its ability to focus on the paramagnetic active sites without interference from the rest of the diamagnetic species makes this technique an ideal method for mechanistic studies. Therefore, it is commonly employed in studies of the structure of enzyme active site, interactions between enzyme and substrate, conformational dynamics, electron transfer, and reaction kinetics.
Electron Paramagnetic Resonance in Enzymology

**EPR Basics**

As an analytical spectroscopic technique, EPR is similar in concept to the more widely used nuclear magnetic resonance (NMR) spectroscopy [see NMR: Overview of Applications in Chemical Biology]. In fact, EPR and NMR are complementary to each other. Both techniques detect magnetic moments, but NMR determines the chemical structures in solution, whereas EPR describes more precisely the electronic and chemical structures of a particular region of the biological system, such as electron transfer centers, metal ions, and an intermediate state of the enzyme or substrate. It is not possible to present a full description of the theory of EPR in an article with this scope. Therefore, only sufficient information is provided here to enable the readers to understand the practical aspects of this analytical tool in enzymology.

**Resonance and the Zeeman effect**

Electrons have “spin,” and when they are unpaired (or “free”), this “spin” gives them a measurable magnetic moment. In the absence of an external magnetic field, the electron’s magnetic moment will orient randomly. The electron with a spin quantum number $S = \frac{1}{2}$ can have two orientations ($m_s = \pm \frac{1}{2}$) in a magnetic field ($B_0$). Therefore, the electrons will align either parallel or antiparallel to the external magnetic field (Fig. 1). The orientation of unpaired electrons causes discrete energy levels (Fig. 2). The orientation antiparallel to $B_0$ is energetically higher than the parallel. The displacement of discrete energy levels is known as the Zeeman effect.

The two spin states ($M_s = \pm \frac{1}{2}$) of an $S = \frac{1}{2}$ electron have the same energy when $B_0$ is 0. In the presence of an external magnetic field, two discrete energy levels occur, and the electronic-Zeeman energy diverges linearly as the strength of the magnetic field increases. In the event that a constant electromagnetic radiation frequency is present, peak absorption of the energy will occur when the magnetic field strength is varied and tuned to the two spin states so that the electronic Zeeman energy matches the energy of the allowed spin transitions. At this point, resonances are observed. The detection of the first...
derivative of energy absorption of the radiation for transitions results in an EPR spectrum (Fig. 2).

For a spin \( S > \frac{1}{2} \) system, such as a system with 5 unpaired electrons each occupying an orbital, and all 5 pointing upward (i.e., in high symmetry), application of an external magnetic field yields uniformly spaced multiplet splitting of energy levels at \( \Delta E = \mu_B H \) for an unpaired electron with spin \( S = \frac{1}{2} \). This energy difference is predominantly because of the interaction of the sample’s unpaired electrons with the external magnetic field, and each \( M_s \) converges to a common origin at zero field. If the electrons are in an asymmetry environment, for example, some point upward and some downward, then the energy levels may not be uniformly spaced. A measurable parameter, zero-field splitting, has been introduced to describe the energy-level splittings of an electron spin (\( S > \frac{1}{2} \)) in the absence of an applied magnetic field (see the materials in the Further Reading section).

**g-factor**

The difference between the energies of the unpaired electron in the initial and final spin states of an allowed transition is \( \Delta E \), which can be described by the fundamental equation of EPR spectroscopy:

\[
\Delta E = h \nu = g \mu_B B_0
\]

where \( h \) is Planck’s constant \((6.6260695 \times 10^{-34} \text{ J s})\), \( \nu \) is the electromagnetic wave of frequency, \( \mu_B \) is the Bohr magneton \((9.2740093 \times 10^{-24} \text{ JT}^{-1})\), and \( g \) is the electron Zeeman factor. Equation 1 shows that \( \Delta E \) is proportional to \( B_0 \), a constant in a sample during an EPR measurement. The interaction between the electron and the nearby nuclei is called hyperfine interaction or hyperfine coupling. Such an interaction would cause additional splitting of the energy levels of unpaired electrons and, in turn, would produce multiplicity in EPR spectrum that requires spin Hamiltonian terms to describe the coupling of the spin system (see the Further Reading section). However, the number of the spectral lines in the EPR spectrum can be predicted by the following simple formula:

\[
g \propto \frac{n}{I+1} + 1
\]

where \( n \) is the number of the equivalent nuclei, that being coupled to and \( I \) is the nuclear spin number. Table 1 summarizes the nuclear spin of common elements that may be used in analyzing EPR data.

At 9.5 GHz, which is in the X-band frequency range, the magnetic field where the EPR signal should occur for a free electron can be calculated from Eq. 3, for example, \( B_0 = 714.4775 \times 2.002 3193 \times 9.5 \approx 3390 \text{ Gauss} \).

\[
B_0 (\text{Gauss}) = \frac{714.4775 (\text{Gauss}/\text{GHz})}{\nu (\text{GHz})}
\]

The value of \( g \)-factor is a measure of the coupling between the spin of an unpaired electron and an external magnetic field. It is not only dependent on the spin species but also on its environment. A single numerical value of \( g \) is applicable only to systems that behave isotropically. With anisotropic systems, a modified term that accommodates the variability of \( g \) with orientations relative to the external field is introduced as \( g \)-tensor. Three values, \( g_x \), \( g_y \), and \( g_z \), which represent principal \( g \)-values, \( B_{xx} \), \( B_{yy} \), and \( B_{zz} \) values of the \( g \)-matrix, are important EPR parameters.

**Hyperfine coupling**

The actual field at each spin species is not necessarily only the externally imposed magnetic field. In addition to this field, local fields from nearby nuclei may add to the external field. Nuclei can also have spins, and they possess an intrinsic spin angular momentum. It is common practice to represent the total angular momentum of a nucleus by the symbol \( I \) and to call it “nuclear spin.” Nonzero nucleus spins affect the local magnetic field where the unpaired electrons are located in a sample during an EPR measurement. The interaction between the electron and the nearby nuclei is called hyperfine interaction or hyperfine coupling. Such an interaction would cause additional splitting of the energy levels of unpaired electrons and, in turn, would produce multiplicity in EPR spectrum that requires spin Hamiltonian terms to describe the coupling of the spin system (see the Further Reading section).

However, the number of the spectral lines in the EPR spectrum can be predicted by the following simple formula:

\[
\text{Hyperfine splitting lines} = 2nI + 1
\]

where \( n \) is the number of the equivalent nuclei, that are being coupled to and \( I \) is the nuclear spin number. Table 1 summarizes the nuclear spin of common elements that may be used in analyzing EPR data.
Table 1  Nuclear spin data of common elements and isotopes

<table>
<thead>
<tr>
<th>Isotope</th>
<th>Spin (I)</th>
<th>Natural abundance (%)</th>
<th>Isotope</th>
<th>Spin (I)</th>
<th>Natural abundance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>¹H</td>
<td>¹/₂</td>
<td>99.99</td>
<td>⁶⁰Cu</td>
<td>³/₂</td>
<td>69.20</td>
</tr>
<tr>
<td>²H</td>
<td>1</td>
<td>0.01</td>
<td>⁶³Cu</td>
<td>³/₂</td>
<td>30.80</td>
</tr>
<tr>
<td>¹²C</td>
<td>0</td>
<td>98.89</td>
<td>⁶⁵Co</td>
<td>³/₂</td>
<td>100</td>
</tr>
<tr>
<td>¹³C</td>
<td>¹/₂</td>
<td>1.11</td>
<td>⁶⁶Co⁺</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>³⁵Cl</td>
<td>³/₂</td>
<td>75.77</td>
<td>⁵³Cr</td>
<td>³/₂</td>
<td>9.50</td>
</tr>
<tr>
<td>³⁷Cl</td>
<td>³/₂</td>
<td>24.23</td>
<td>⁵⁴Fe</td>
<td>³/₂</td>
<td>5.85</td>
</tr>
<tr>
<td>¹⁴N</td>
<td>¹/₂</td>
<td>99.63</td>
<td>⁵⁵Fe⁺</td>
<td>³/₂</td>
<td>0</td>
</tr>
<tr>
<td>¹⁵N</td>
<td>¹/₂</td>
<td>0.37</td>
<td>⁵⁶Fe</td>
<td>³/₂</td>
<td>0</td>
</tr>
<tr>
<td>¹⁶O</td>
<td>0</td>
<td>99.95</td>
<td>⁵⁷Fe</td>
<td>³/₂</td>
<td>0.28</td>
</tr>
<tr>
<td>¹⁷O</td>
<td>³/₂</td>
<td>0.04</td>
<td>⁵⁸Fe</td>
<td>³/₂</td>
<td>91.72</td>
</tr>
<tr>
<td>³³S</td>
<td>³/₂</td>
<td>0.75</td>
<td>⁵⁹Co</td>
<td>³/₂</td>
<td>1.13</td>
</tr>
<tr>
<td>³¹P</td>
<td>¹/₂</td>
<td>100</td>
<td>⁶⁰Co</td>
<td>³/₂</td>
<td>15.90</td>
</tr>
<tr>
<td>¹⁸F</td>
<td>¹/₂</td>
<td>100</td>
<td>⁶¹Ni</td>
<td>³/₂</td>
<td>9.60</td>
</tr>
</tbody>
</table>

*Radioactive.
**Half-life: 2.7 years.

Figure 3  Energy-level diagram and corresponding EPR signals at 9.6 GHz (X-band) with maximized resolution for an unpaired electron in an imposed magnetic field that interacts with a) a proton (nuclear spin I = ¹/₂), b) a ¹⁴N nucleus (I = ¹), and c) a ¹⁴N nucleus in which the energy levels are additionally split by a nearby proton. The spectra were produced by the EPR simulation program SimFonia for illustration of the hyperfine splitting purpose.
the hyperfine splitting lines = (2 × 1 + 1)2 × 1/2(2 + 1) = 2 × 3 = 6. These spectral lines are not always fully resolved. They could partially overlap, which would result in an EPR signal with a hyperfine structure.

An important EPR parameter, hyperfine coupling constant, A, is required to describe the hyperfine interaction. The A values can be obtained directly from EPR spectra whose equally spaced spectral lines are exhibited because of the hyperfine interactions (Fig. 3). In many cases, a matrix, known as A tensor, is required to describe the A values for a more complex anisotropic system.

Because of the hyperfine interaction, EPR is very sensitive to the local surroundings. From a chemical standpoint, this property provides a wealth of information such as the identity and number of atoms that makes up a molecule or complex and their distances from the unpaired electron. This information can often be translated, in favorable cases, into the molecular structure of the sample.

Much of the information EPR provides about the composition, structure, and bonding of a paramagnetic metal center is obtained through the analysis of the hyperfine coupling constants that represent interactions between the spin of the unpaired electron(s) and the spins of nuclei associated with the metal center. These coupling constants are calculated from splittings seen in the EPR spectrum. In many occasions, the splittings are not resolvable in conventional EPR spectroscopy. Advanced EPR spectroscopy, such as electron nuclear double resonance (ENDOR) and electron spin echo envelope modulation (ESEEM) spectroscopy, are necessary to retrieve the missing information (10). ENDOR and ESEEM provide an NMR spectrum of those nuclei that interact with the electron spin of the paramagnetic center, and such spectra display orders-of-magnitude-better resolution than the EPR spectrum of the center. Resolved features in these NMR spectra are characterized by frequencies, and these frequencies directly give the electron-nuclear coupling constants (see the materials in the Further Reading section).

Detection and characterization of enzyme-based free radicals

In the past three decades, evidence for enzyme-based protein radicals as the catalytic driving force or the key transient intermediate in enzymatic reactions has strengthened (11). Spectroscopic data from EPR suggest that these highly reactive species can be stored with remarkable stability inside an enzyme either as a transient intermediate or as a paramagnetic center and catalyze a wide range of biological reactions in electron transfer and oxidation/reduction.

It was not known that a protein could harbor a stable free radical derived from an amino acid residue until the discovery of the first tyrosyl radical from the smaller R2 subunit of the *Escherichia coli* ribonucleotide reductase (RNR) in the 1970s. RNR catalyzes the first committed step in DNA synthesis by reducing the hydroxyl group of all four types of ribonucleotide to produce the corresponding deoxyribonucleotides (12). RNR, thus, is an essential enzyme for all living organisms. The observation of an EPR signal with an asymmetric doublet centered at $g = 2.0047$ in the EPR spectrum and a sharp absorption peak at 410 nm ($\epsilon = 3250 \text{ M}^{-1}\text{cm}^{-1}$) of the optical spectrum led to the identification of a protein-based radical (13). This free radical has a half-life time of a matter of days at room temperature and is an integral part of the R2 subunit. It is essential for the enzyme activity. RNR, thus, has been introduced as “a radical enzyme” (14). Later, it was discovered that a transient cysteine-based thyl radical is formed at the expense of the tyrosyl radical during catalysis and that the transient thyl radical is responsible for the catalytic reduction of the ribonucleotide substrates (15, 16).

It would be expected that all tyrosyl radicals have similar spectroscopic properties; however, the characterized tyrosyl radicals exhibit distinct spectral features. From a spectroscopic point of view, at least three different EPR signals for tyrosyl radicals are found (Fig. 4). The EPR signal shown in Fig. 4a (trace a) is observed in the active form of E. coli RNR R2 protein, which exhibits a 1.8 mT hyperfine splitting (13). EPR signals similar to trace b have been found in *Mycobacterium tuberculosis* and *Salmonella typhimurium* RNRs, photosystem II (PSII), and many other enzyme-based tyrosyl radical intermediates (16, 19, 20). The most complex spectrum, trace c, exhibits the largest (ca. 2.1 mT) hyperfine splitting and is observed from mouse, yeast, and higher plant RNR R2 proteins. These radicals are the representative neutral tyrosyl radicals. The tyrosyl radicals in R2 proteins are stable because the free electron is delocalized to the phenyl ring and the radical center is shielded from reaction by a hydrophobic pocket.

Most known protein-based tyrosyl radicals, either stable or transient, exhibit a line shape that resembles one of the above line shapes unless they are strongly spin-coupled with another paramagnetic center, such as that found in galactose oxidase (17, 18). Although the observed EPR spectra are distinct, the differences in EPR spectral characteristics are mainly because of the dihedral angles $\theta_{xy}$ defined by the locations of the $\beta$-methylene protons, $\beta$-methylene carbon, ring carbon C1, and its 2p axis relative to the phenyl ring of the side chain (19). These dihedral angles (Table 2) can be determined from the...
Figure 4  a) X-band EPR spectra of tyrosyl free radical in (i) *E. coli*, (ii) *Mycobacterium tuberculosis*, and (iii) mouse ribonucleotide reductase R2 proteins (17). All spectra were obtained under nonsaturation conditions at 20 K. b) Spin density distribution of the unpaired electron obtained from isotope-labeling EPR studies. c) The distances between the phenolic oxygen of tyrosyl radical and the nearest Fe ion deduced from the relaxation properties of the tyrosyl radicals.

It should be noted that no structure analysis is available for the radical-containing form in the above enzymes with the exception of one radical-containing form obtained by oxidizing the existing crystals of *E. coli* R2 by H$_2$O$_2$ (22). It is rare to obtain the crystal structure of a biological radical within its catalytic site, although, indeed, some successes have been accomplished. One such case is the intermediate radical form of hydroxyethylidene-thiamine pyrophosphate (HE-TPP), which is visualized in a structure of the enzyme pyruvate ferredoxin oxidoreductase (PFOR) crystallized with HE-TPP (23). The radical is present in the structure because the thiazole ring is apparently puckered. Free radical intermediates are generally not sufficiently stable to survive crystallization and X-ray diffraction, thus EPR spectroscopy is often the most pertinent tool for comprehending the chemical properties of the radicals. Thus far, the electronic structure and orientation of the tyrosyl radicals in enzymes have all been determined by EPR-based techniques in conjunction with isotope-labeling of the radicals.

**The interaction between radical and other paramagnetic species**

EPR can often detect the interaction between radical and other paramagnetic species. The tyrosyl radical in RNR is generated by a nonheme di-iron cluster. It is visualized from the crystal structure of P700$^+$ of *H. pluvialis* at 2.8 Å resolution (24). The tyrosyl radical in RNR is in close proximity to the second Fe ion of the di-iron cluster.

**Table 2** Dihedral angles of β-methylene protons of tyrosyl radicals in PSII and R2 and the hyperfine splitting in the EPR spectra

<table>
<thead>
<tr>
<th>Radical</th>
<th>Hyperfine (mT)</th>
<th>$\theta_1$ (deg)$^1$</th>
<th>$\theta_2$ (deg)$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>M. tb R2</td>
<td>1.0</td>
<td>52(± 6)</td>
<td>68(± 6)</td>
</tr>
<tr>
<td>E. coli R2</td>
<td>1.8</td>
<td>30(± 3)</td>
<td>90(± 3)</td>
</tr>
<tr>
<td>PSII</td>
<td>1.0</td>
<td>52(± 4)</td>
<td>68(± 4)</td>
</tr>
</tbody>
</table>

$^1\theta_1$ and $\theta_2$ represent the dihedral angles between the 2pz orbital at C$_\beta$ and the planes containing the C$_1$-C$_\beta$-H$_\beta$$_1$ bonds and the C$_1$-C$_\beta$-H$_\beta$$_2$ bonds, respectively.

$^2$Y$_D$• is one of the two tyrosyl radicals in PSII.
structures of the enzyme that the phenolic oxygen of the tyrosine is 5.3 Å away from the nearest iron ion in *E. coli* R2 and 6.8–6.9 Å in *Salmonella typhimurium* and *Mycobacterium tuberculosis* R2 proteins (24). This distance is unknown in the mouse R2 because the structure contains a partially occupied metal center (25). It is also uncertain whether the differences of the distance still exist in the tyrosyl radical-containing forms because the structures were determined in the absence of the radicals. EPR spectroscopy can be used to study these unresolved issues. Although the di-ferric center in R2 proteins is diamagnetic because of the antiferromagnetic coupling nature of the two high-spin ferric ions, the interaction of the di-iron center with the adjacent tyrosyl radical can be detected indirectly by EPR spectroscopy. When the tyrosyl radical was found in *E. coli* RNR R2 protein (α-type), it exhibited some unusual relaxation properties relative to an organic free radical that is not part of a protein. Later biochemical and biophysical studies attributed these unusual relaxation properties to the magnetic dipolar and to exchange interactions between the R2 free radical and the iron center. The interaction between the iron-centre and tyrosyl radical in mouse R2 (α-type) is strongest, whereas in *Salmonella typhimurium* and *Mycobacterium tuberculosis* R2 proteins (β-type), it is almost negligible. Therefore, the iron-radical distance must be shorter in mouse and yeast R2 proteins.

The EPR relaxation behavior can be analyzed quantitatively by measuring microwave power saturation profiles of the EPR signal of the radical at various temperatures. The EPR signal intensity increases in proportion to the square root of the microwave power until the onset of saturation of the spin system (26). The power saturation occurs when the rate of absorption of microwave exceeds the rate at which the system returns to equilibrium. A spectral parameter, *P*<sub>1/2</sub>, is used to describe quantitatively the microwave power saturation profile. In the RNR tyrosyl radical case, the *P*<sub>1/2</sub> values at four representative temperatures are given in Table 3. The most straightforward interpretation for the easily saturated radical spectra with very small *P*<sub>1/2</sub> values, as seen in *M. tuberculosis* R2, is that the tyrosyl radical is minimally influenced in its relaxation by the di-ferric cluster. This finding is reverse in mouse and yeast R2 proteins. To obtain the precise distance information in a biological system, advanced techniques such as ESSEM would be more pertinent than the continuous-wave EPR spectroscopy.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>10</th>
<th>77</th>
<th>100</th>
<th>273</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P</em>&lt;sub&gt;1/2&lt;/sub&gt; (mW) of Tyr&lt;sup&gt;•&lt;/sup&gt; of <em>M. tuberculosis</em> R2</td>
<td>0.99</td>
<td>0.72</td>
<td>1.28</td>
<td>144&lt;sup&gt;•&lt;/sup&gt;</td>
</tr>
<tr>
<td><em>P</em>&lt;sub&gt;1/2&lt;/sub&gt; (mW) of Gly&lt;sup&gt;•&lt;/sup&gt; of <em>L. lactis</em> RNR</td>
<td>&lt;0.01</td>
<td>0.10</td>
<td>0.14</td>
<td>7</td>
</tr>
</tbody>
</table>

*The estimated value for the half saturation because the full saturation did not occur with the applied microwave power.*
Electron Paramagnetic Resonance in Enzymology

EPR Characterization of Metallocofactors

For the metalloenzymes (see Metallo-Enzymes and Metallo-Proteins, Chemistry of) with a paramagnetic metal center, NMR spectroscopy may have limitations because the lines of the nuclei near the metal are often broad (see NMR for Proteins). X-ray crystallography reveals the location of atoms when the diffraction quality of crystals is obtainable; however, it does not illustrate the electronic structure, the oxidation state of the metal ion, and the charge distribution, which are the real determinants of the activity of active site in metalloenzymes. Therefore, even with the crystal structure, information about transient or intermediary steps still is needed to understand electron or atom transfer mechanisms and enzymatic reactions. For this purpose, spectroscopy is indispensable, and for obtaining information about paramagnetic centers, EPR-centered methods, including pulse techniques and the technique of electron-nuclear double resonance at high- and low-field, are uniquely suited. The application of EPR spectroscopy for the characterization of metalloenzymes has recently been reviewed in several elegant articles (see the Further Reading section). In this section, three examples will be described to highlight the potential of EPR spectroscopy for characterization of metalloenzymes.

Electronic structure of paramagnetic centers in enzymes and the response to chemical, redox, and environmental changes

In a recent combined biochemical and spectroscopic study, the RNR R2 protein from Chlamydia trachomatis (Ct), a human pathogen, is shown to employ a novel enzyme-bound MnIVFeIII cofactor to initiate catalysis in the R1 protein (33). This heterodinuclear cofactor exhibits an S = 1 ground spin state, which results from antiferromagnetically coupling of the MnIV (d3, S = 3/2) and FeIV (d5, S = 5/2) ions (34). It is reduced to an EPR-active MnIIIFeII form when the Ct R2 protein is treated with the substrate analog 2'-diphosphate (N3-ADP) in the presence of R1 and adenosine triphosphate (ATP) (33). This observation suggests that the MnIVFeIII cofactor can generate protein and substrate radicals in the R1 site during the catalysis. The MnIVFeIII cluster is proposed, therefore, to be a functional equivalent of the MnIVFeIV cluster capable directly to the iron-sulfur center. Using advanced EPR spectroscopic techniques (see the Further Reading section), the EPR spectrum of the MnIIIFeIII form is centered at g ≈ 2 region (1.86, 1.92, 2.00). A similar observation of substrate/cofactor-induced EPR-silent activase (37). The fully reduced activase is proposed to contain an all-ferrous [4Fe-4S]0 cluster because it is a diamagnetic and EPR-inactive. When the activase is reduced by dithioate or a photochemical reductant, a typical low spin spectrum changes, for example, increased rhombicities and distinct line patterns may be observed (see the Further Reading section). In the presence of SAM, the photochemically reduced adenosyl radical is also proposed to be a key intermediate in pyruvate formate lyase (PFL) and anaerobic RNR enzymes in which it produces a stable g-acytyl radical for initiating catalysis. In PFL and RNR, the iron-sulfur cluster that cleaves SAM is located in an activase protein. The resting state of the activase contains a [4Fe-4S]3+/2 cluster, which is diamagnetic and EPR-silent. The activase is proposed to be a functional equivalent of the MnIVFeIII cluster by direct air exposure (Fig. 6). The fully reduced activase is proposed to contain an all-ferrous [4Fe-4S]3+ cluster, which is diamagnetic and EPR-silent. The activase is proposed to contain an all-ferrous [4Fe-4S]3+ cluster, which is diamagnetic and EPR-silent.
Concluding Remarks

EPR data on enzymes and mechanisms have been accumulating rapidly, and the knowledge learned from EPR-centered techniques has significantly contributed to the development of modern enzymology. In this article, the background and technique of EPR spectroscopy is introduced, and a few selected topics are used to illustrate the potential of its application in two areas of enzymology.

EPR spectroscopy has been a pivotal tool in the identification, quantification, and characterization of both cofactor-derived and protein-based radicals. With the assistance of EPR spectroscopy, several important enzymes are now known to be radical enzymes, including the well-studied cytochrome c peroxidase, galactose oxidase, photosystem II, prostaglandin H synthase, pyruvate formate lyase, ribonucleotide reductase, quinoox-containing enzymes, and coenzyme B12, or SAM-dependent superfamily of enzymes. Although as a terminology “radical enzymology” was only formally introduced in 2003 (11), it was recognized long before as a major area of enzymology. EPR-centered techniques will continue to be the primary tools in the future development of radical enzymology.

The oxidation state and electronic structure of protein-based paramagnetic metal ions can also be described in great detail by EPR spectroscopy, and the structural transitions in response to substrate-binding and the redox and environmental changes can be followed by EPR spectroscopy. Therefore, this technique provides an experimental handle in the studies of metalloenzymes.

To solve more complex chemical and biological problems, EPR spectroscopy has grown into a family of techniques. In addition to the conventional X-band continuous-wave EPR spectroscopy, ENDOR and ESEEM have become popular analytic tools (see the Further Reading section). PELDOR is an emerging technique, and its value in enzymology has received considerable attention (40, 41). EPR at a very high field (HF-EPR) is expected to characterize integer spin systems with large zero-field interactions, some of which are EPR silent at lower fields. Examples of where such an approach would be useful are Mn(III) in manganese enzymes, Fe(II) in hemoglobin, Fe(IV), Cof(I) in vitamin B12-binding enzymes, Ni(II), Mo(IV) in oxidases, and W(V) in dehydrogenases (42). The combination of pulse EPR and multifrequency EPR techniques seems to be the most promising use for enzymologists who are studying the dynamics and long-distance interactions in radical enzymes and paramagnetic metalloproteins. The development and successful application of these advanced EPR techniques will continue to play an important role in the future development of enzymology.

Acknowledgment

The NIH subaward of GM069618, an institutional American Cancer Society award, and an ORAU Faculty Enhancement Award in Life Sciences are acknowledged.

Figure 6 EPR spectra of the iron-sulfur cluster in the activase protein of anaerobic BNR from L. lactis. a) (3Fe-4S) cluster in oxidized activase, b) (4Fe-4S) cluster in photochemically reduced activase, and c) in the presence of SAM. Spectra were taken at 10 K (35).

Electron Paramagnetic Resonance in Enzymology

Electron transfer

EPR spectroscopy can serve as an efficient tool for studying the delivery of electrons between biological systems. Nitrogenase, for example, consists of an iron protein and an iron-molybdenum protein. N₂-binding and reduction take place at the iron-molybdenum protein, which contains a 3Fe-9S-Mo-X-homocitrate cofactor (FeMo-co). During catalysis, FeMo-co acquires electrons one at a time from the Fe-protein. Using rapid freeze-quench EPR, Hoffman and Seefeldt et al. have trapped and characterized several reactive intermediates of the enzyme reaction (39). The relaxation properties of the intermediates have been measured by an innovative step-annealing approach, and the information obtained has helped to assign the structure of the observed intermediates with proposed structure. The direct outcome of this work is a significant step forward in understanding the nitrogenase catalytic mechanism. The step-annealing approach employed in the nitrogenase study includes rapidly warming a sample held at 77 K to 253 K by placement for a fixed time in a methanol bath held at that temperature, quench-cooling it back to 77 K, and then collecting a 2 K EPR spectrum. These procedures are based on Davydov and colleagues’ studies of the iron-oxygen intermediates that are directly generated in an EPR sample at 77 K by radiolytic cryoreduction (39).

The oxidation state and electronic structure of protein-based paramagnetic metal ions can also be described in great detail by EPR spectroscopy, and the structural transitions in response to substrate-binding and the redox and environmental changes can be followed by EPR spectroscopy. Therefore, this technique provides an experimental handle in the studies of metalloenzymes.

To solve more complex chemical and biological problems, EPR spectroscopy has grown into a family of techniques. In addition to the conventional X-band continuous-wave EPR spectroscopy, ENDOR and ESEEM have become popular analytic tools (see the Further Reading section). PELDOR is an emerging technique, and its value in enzymology has received considerable attention (40, 41). EPR at a very high field (HF-EPR) is expected to characterize integer spin systems with large zero-field interactions, some of which are EPR silent at lower fields. Examples of where such an approach would be useful are Mn(III) in manganese enzymes, Fe(II) in hemoglobin, Fe(IV), Cof(I) in vitamin B12-binding enzymes, Ni(II), Mo(IV) in oxidases, and W(V) in dehydrogenases (42). The combination of pulse EPR and multifrequency EPR techniques seems to be the most promising use for enzymologists who are studying the dynamics and long-distance interactions in radical enzymes and paramagnetic metalloproteins. The development and successful application of these advanced EPR techniques will continue to play an important role in the future development of enzymology.

Acknowledgment

The NIH subaward of GM069618, an institutional American Cancer Society award, and an ORAU Faculty Enhancement Award in Life Sciences are acknowledged.
References


Further Reading

See Also
Chemistry of B12-Dependent Enzyme Reactions
Metallo-Enzymes and Metallo-Proteins, Chemistry of
Nitric Oxide, Biological Targets of
NMR: Overview of Applications in Chemical Biology
NMR for Proteins
Biologists and chemists are increasingly turning to fluorescence to investigate proteins. Not only is fluorescence detection highly sensitive, but it can also report specific information about a range of properties of proteins. For example, fluorescence can be used to probe the environment of fluorescent molecules, interactions with other proteins, and the motions of proteins over time scales from picoseconds to seconds. This article surveys the properties of biologic fluorophores that can be used to study proteins and introduces the fluorescence techniques commonly applied to proteins, with an emphasis on the information generated. Techniques discussed include steady-state emission, time-resolved detection, fluorescence anisotropy or depolarization, energy transfer, fluorescence correlation spectroscopy, single-molecule spectroscopy, and fluorescence imaging.

Fluorescence is the basis for elegant and sensitive tools for studying proteins. Its applications range from detection of proteins at concentrations as low as single molecules to characterization of protein structure and dynamics. Since Gregorio Weber pioneered protein fluorescence studies in the 1950s (1), biologists and chemists have increasingly turned to fluorescence to probe the properties of proteins. Two fundamental reasons exist to use fluorescence. First, fluorescence is one of the most sensitive detection methods known. Second, fluorescence signals report a wealth of information on protein structure, dynamics, and interactions. This chapter contains a survey of fluorescence techniques with an emphasis on the information generated relevant to proteins. Additional surveys can be found listed in the Further Reading list. More focused review articles are referenced in the context of specific methods or techniques.

**Biologic Fluorophores**

Fluorescence measurements on proteins require both an appropriate fluorescence technique and the presence of a suitable fluorophore. The techniques used for the application of fluorescence to proteins are described later in this article. In this section, we briefly consider three classes of fluorophores that are used widely to study proteins: native fluorophores including fluorescent amino acids, extrinsic fluorescent labels, and autofluorescent proteins. Each has advantages for probing proteins and has distinct drawbacks: No perfect fluorophore exists for studying proteins.

**Native fluorescent amino acids**

The use of fluorophores intrinsic to the protein allows researchers to probe protein structure and dynamics without incorporating non-native fluorophores that could perturb the native structure of the protein. Proteins usually contain one or more fluorescent amino acids, which makes dynamic studies on the native protein feasible. Often, fluorescent amino acid residues can be introduced by site-directed mutation without altering protein structure significantly.

The three aromatic amino acids, tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe), are the only native amino acids with useful fluorescence properties. Figure 1A shows their absorption and fluorescence spectra. Their fluorescence properties are summarized in Table 1. Note that the relative absorption coefficients increase in the order Phe < Tyr < Trp. The fluorescence quantum yields increase in the same order. The product of absorption coefficient and fluorescence quantum yield can be taken as a measure of the brightness of the fluorophore. By the standards of fluorescent dyes (see below), the brightness of all three amino acids is poor. Trp is the brightest of the three, and for proteins with a small number of Trp residues it may be possible to assign fluorescence decays to specific Trp residues. As a result, of the three fluorescent amino acids, Trp is by far the most widely exploited for its fluorescent properties. Fluorescence from Tyr is also detectable but may be masked by Trp fluorescence. Proteins often contain many Tyr residues, so it is often not possible to isolate the fluorescence from individual Tyr residues. Fluorescence from Phe is weak and not often used in fluorescence studies.
Fluorescence Techniques for Proteins

Figure 1  (a) Absorption and fluorescence spectra of fluorescent amino acids blocked with peptide bonds: N-acetyltryptophanamide (black), N-acetyltyrosinamide (green), and N-acetylphenylalaninamide (red) in aqueous solution. The absorption spectra (left) show the relative absorption strengths of the three amino acids. The fluorescence spectra (right) are normalized to the same peak intensity (see Table 1 for relative quantum yields). (b) Fluorescence emission spectra of the maleimide derivatives of four common fluorescent dyes: Alexa Fluor 488 (green), tetramethylrhodamine (yellow), Texas red (orange), and Cy 5 (red).

Table 1. Fluorescence properties of amino acids and other intrinsic fluorophores

<table>
<thead>
<tr>
<th>Fluorophore</th>
<th>Absorption λ&lt;sub&gt;max&lt;/sub&gt; (nm)</th>
<th>Absorption coefficient (M&lt;sup&gt;-1&lt;/sup&gt;cm&lt;sup&gt;-1&lt;/sup&gt;)</th>
<th>Fluorescence λ&lt;sub&gt;max&lt;/sub&gt; (nm)</th>
<th>Lifetime (ns)</th>
<th>Quantum yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tryptophan</td>
<td>280</td>
<td>5,600</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate in water</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyrosine (in water)</td>
<td>274</td>
<td>1,400</td>
<td>304</td>
<td>3.2</td>
<td>0.07</td>
</tr>
<tr>
<td>NADH (in water)</td>
<td>340</td>
<td>6,200</td>
<td>470</td>
<td>0.52</td>
<td>0.02</td>
</tr>
<tr>
<td>Phenylalanine (in water)</td>
<td>257</td>
<td>200</td>
<td>260</td>
<td>6.8</td>
<td>0.02</td>
</tr>
<tr>
<td>FMN</td>
<td>450</td>
<td>12,000</td>
<td>520</td>
<td>4.7</td>
<td>0.2</td>
</tr>
<tr>
<td>FAD</td>
<td></td>
<td></td>
<td></td>
<td>0.5</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Although Trp fluorescence is the strongest of the three fluorescent amino acids, its fluorescence is also the most complex. The Trp fluorescence spectrum is highly sensitive to its environment. Trp residues buried in the protein tend to fluoresce at shorter wavelengths (λ<sub>max</sub> as low as 308 nm), whereas solvent-exposed Trp residues emit at longer wavelengths (λ<sub>max</sub> = 353 nm in water). The environmental sensitivity of λ<sub>max</sub> results from the large increase of the Trp dipole moment in the fluorescing excited state (denoted the 1<sub>La</sub> state). The environment surrounding Trp, including H-bonds and charged groups, can therefore tune the emission wavelength over a large spectral range (see reference 2). The fluorescence quantum yield of Trp is also highly sensitive to the environment, and depends on the distance and the orientation of the indole side chain of Trp relative to groups that quench Trp fluorescence, especially carbonyl groups of the peptide backbone. The fluorescence quantum yield of Trp is modulated further by the presence of two closely lying excited states, 1<sub>La</sub> and 1<sub>Lα</sub>. Several fine reviews of the fluorescence properties of Trp describe these properties in greater detail (see Further Reading). Although the fluorescence from Tyr residues is often less intense than the fluorescence from Trp residues, Tyr fluorescence is not complicated by the presence of two close-lying excited states, and Tyr emission spectra are much less sensitive to the
fluorescence in its reduced, NADH state) and flavin adenine dinucleotide (FAD). NADH is weakly fluorescent in water, but its fluorescence yield increases markedly on binding to a protein-binding site with an emission peak around 470 nm (3). FAD and flavin mononucleotide (FMN) are also fluorescent with an emission maximum around 520 nm, but fluorescence is quenched on binding to many flavoproteins (4).

Fluorescence labels
As described above, the intrinsic fluorophores that nature provides in proteins generally have rather low absorption coefficients and quantum yields. For many applications, brighter fluorescence probes are needed, and emission in the visible region is desirable. One way to meet these needs is by labeling with a fluorescent dye. A wide range of fluorescence probes is now available for this purpose (see Fig. 1b). A summary of the spectroscopic and photophysical properties of many fluorescence probes is available in References 5 and 6.

Extrinsic fluorophores have emerged as remarkable tools for studies of proteins in part because reactive derivatives are available to target labeling to specific functional groups in proteins. Derivatives (which include isothiocyanates, succinimidyl esters, and sulfonyl chlorides) are available to label amino groups such as lysine residues or the N-terminus of the protein. However, proteins may contain many lysine residues, which makes it problematic to label a specific site selectively. Thiol-reactive fluorescent probes (iodoacetamides and maleimides) offer more selectivity because they can be targeted specifically to cysteine (Cys) residues in proteins. Additional details about the chemistry of labeling functional groups in proteins are available from Molecular Probes (Invitrogen Corp., Eugene, OR) (5).

Often, the greatest challenge in applying fluorescence probes to proteins is the labeling procedure itself. Although the chemical conditions for effective labeling of amino groups and thiols are well established, selectively targeted labeling may require thorough characterization of the reactivities of functional groups. Proteins often have multiple Cys or Lys residues with varying reactivities toward thiol-reactive or amine-reactive probes, respectively. Under the proper conditions, it may be possible to label the most reactive site. Other strategies include site-directed mutagenesis to eliminate Cys residues from protein sites where labeling is undesired and to introduce a Cys residue at the desired location. More selective labeling motifs can also be introduced into proteins by site-directed mutagenesis. For example, a tetra-cysteine motif can be introduced and labeled with a fluorophore derivatized with biarsenical ligands (7).

Figure 2a illustrates the concepts of radiative and nonradiative decay, fluorescence quantum yield, and fluorescence decay. A molecule in an excited electronic state can relax by several channels. Molecules excited to a vibrational level in the excited state undergo vibrational relaxation (cooling, yellow arrows in Fig. 2a) to the lowest vibrational level of the excited state in a
Fluorescence Techniques for Proteins

Figure 2 (a) Jablonski diagram of photophysical processes related to light absorption and emission. One-photon (green arrows) or two-photon (red arrows) exciting a vibronic level in the excited S1 electronic state is followed by rapid vibrational relaxation in the excited state. Nonradiative relaxation of the excited state can occur by internal conversion to the ground state (blue arrow) or intersystem crossing to the triplet state T1 (gray arrow). Other processes that may deplete the excited state include photochemistry and energy transfer (light blue arrow). The excited state can also relax radiatively by emitting a photon (orange arrow). (b) Illustration of fluorescence generated by one-photon excitation (left) or two-photon excitation (right). The one-photon excitation probability is linear in light intensity, whereas the two-photon excitation probability is quadratic in light intensity and therefore occurs mostly in the focal region.

few picoseconds or less. Other nonradiative relaxation channels include internal conversion with rate constant $k_{ic}$ to generate another electronic state with the same electronic spin, or intersystem crossing with rate constant $k_{isc}$ to the triplet state T1. Other decay channels may include excited-state reactions such as electron transfer, proton transfer, or isomerization, with rate constants denoted $k_{pc}$ in Fig. 2a or electronic energy transfer to another molecule, $k_{fret}$. All processes compete kinetically with radiative decay from the excited state. Radiative decay has an intrinsic rate constant (denoted $k_r$) that can be calculated from the integrated absorption coefficient (11).

The total decay rate constant $k_f$ of the excited state is simply the sum of the rate constants for all processes that deplete the excited state:

$$k_f = k_r + k_{ic} + k_{isc} + k_{pc} + k_{fret} + \cdots$$  \hspace{1cm} (1)

where ... includes the rates of any other processes that deplete the excited state. The intensity of fluorescence is proportional to the population of excited states; therefore, after excitation the probability of fluorescence emission decays with the rate constant $k_f$. The inverse of this value is called the fluorescence lifetime $\tau_f$.

The fluorescence decay is then described by:

$$I(t) = I(0) \exp(-k_f t) = I(0) \exp(-t/\tau_f)$$  \hspace{1cm} (2)

The fraction of the decay rate that results in emission of a photon is the fluorescence quantum yield (or quantum efficiency) $\phi_f$:

$$\phi_f = \frac{k_r}{k_f}$$  \hspace{1cm} (3)

Note that if the radiative rate $k_r$ can be calculated, then the fluorescence decay rate and fluorescence lifetime follow from the fluorescence quantum yield $\phi_f$. Of course, the situation is often more complex. As will be described below, fluorescence decays for proteins often do not follow the single exponential decay model of Equation 2. The fluorescence quantum yield and Equation 3 then provide an average fluorescence lifetime.

Excitation sources

The simplest steady-state measurements of fluorescence properties such as the fluorescence emission spectrum or the steady-state anisotropy can be carried out in a standard fluorometer with excitation from a lamp source and a monochromator. Common lamp sources in commercial fluorometers include xenon lamps for excitation from the UV to the near-IR (250 nm to 1100 nm).

For many advanced fluorescence applications, the sample is excited with a laser. Lasers have several advantages over traditional lamp sources: 1) Laser beams often have low beam divergence and therefore are directed readily and focused onto a sample; 2) laser sources can generate narrow excitation bandwidths, which allows excitation of the fluorophore of interest or photoselection of a subset of fluorophores; and 3) through a technique called mode-locking, some lasers can generate extremely short pulses of light, which allows time resolution of protein dynamics on time scales from less than one picosecond ($10^{-12}$ s). The laser source selected for a given technique will depend on the wavelengths needed, the desired beam power, and, for time-resolved experiments, the required pulse width. The properties of some laser sources are summarized in Table 2.
Table 2: Common laser sources for fluorescence excitation

<table>
<thead>
<tr>
<th>Laser Type</th>
<th>Wavelength(s) (nm)</th>
<th>Typical average Power</th>
<th>CW or pulse characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ar ion</td>
<td>457, 476, 488, 496, 502, 534, 529</td>
<td>10 mW to 10 W total for all lines</td>
<td>CW</td>
</tr>
<tr>
<td></td>
<td>(plus other weaker lines)</td>
<td></td>
<td>Pulse can be generated for some lines (476, 496, 502, 514) by mode-locking (pulsewidth 100's of picoseconds)</td>
</tr>
<tr>
<td>Kr ion</td>
<td>531, 568, 647, 676, 752 (plus other weaker lines)</td>
<td>10 mW to 1 W total for all lines</td>
<td>CW</td>
</tr>
<tr>
<td>Diode pumped solid state</td>
<td>460, 473, 488, 532, 561, 635, 660</td>
<td>1 mW - 10 W</td>
<td>CW or pulsed</td>
</tr>
<tr>
<td></td>
<td>(other wavelengths also available)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diode</td>
<td>400-1650</td>
<td>1-100 mW</td>
<td>CW or pulsed</td>
</tr>
<tr>
<td>Dye</td>
<td>400-900</td>
<td>10 mW - 1 W</td>
<td>CW or pulsed</td>
</tr>
<tr>
<td>He-Cd</td>
<td>325, 440</td>
<td>2-200 mW</td>
<td>CW</td>
</tr>
<tr>
<td>He-Ne</td>
<td>543, 594, 633</td>
<td>1-10 mW</td>
<td>CW</td>
</tr>
<tr>
<td>Nd (Nd:YAG, Nd:YVO₄)</td>
<td>1064 and 352, 355, 266</td>
<td>1-10 W</td>
<td>CW or mode-locked (pulse width ~100 ps)</td>
</tr>
<tr>
<td>Ti:sapphire</td>
<td>700-1100 and harmonics: 350-550, 233-366</td>
<td>100 mW - 1 W</td>
<td>CW or modelocked (pulse widths ~10-200 fs)</td>
</tr>
</tbody>
</table>

Two-photon excitation

With intense laser pulses, new nonlinear optical phenomena are possible. The prime example is two-photon excitation (TPE). The peak power in a laser pulse from a Ti:sapphire laser (pulse width ~100 fs) can readily reach 10⁵ W or higher, with a focused intensity of 10¹⁴ W/cm². Under these conditions, excitation can occur with two photons that have half of the energy (twice the wavelength) of the corresponding one-photon transition (see Fig. 2a). The rate of TPE is given by:

\[
\frac{dN^*}{dt} = \delta(2) I^2 N^2 \cdot (hv)^2
\]

where \( N^* \) is the population of the excited states, \( N \) the population of ground states, \( I \) is the peak intensity of the laser pulse in W/cm², \( v \) is the optical frequency of the laser pulse, \( h \) is Planck’s constant, and \( \delta(2) \) is the two-photon cross section. The two-photon excitation probability also depends on the polarizations of the photons, but \( \delta(2) \) is usually given as an orientationally averaged value for linearly polarized light and an isotropic sample. The two-photon excitation spectra and cross sections have been described for fluorescence dyes (12), Trp (13), NADH and flavins (14), and GFP and other fluorescent proteins (15).

Fluorescence Techniques Applied to Proteins

Multiple parameters can be measured for fluorescence photons: count rate (or intensity), wavelength (\( \lambda \)), polarization (\( p \)), arrival time (\( t_a \)), time delay after excitation (\( t_d \)), and location (\( x_t \)) on an imaging detector. These parameters carry information about the fluorophore that includes the nature of its environment, its interactions with other molecules, and its motions. Fluorescence techniques that exploit each of these parameters are described below.
Steady-state fluorescence spectroscopy: intensity and wavelength

The most straightforward fluorescent technique is simply the measurement of the fluorescence intensity. Laser-induced fluorescence (LIF) is one of the most sensitive detection methods known. In LIF, the fluorescence intensity excited by a laser beam is used to quantify the amount of fluorophore present. When combined with a separation technique such as capillary electrophoresis, LIF is a powerful biologic analysis method. Detection limits can reach less than 100 femtomolar and potentially the single-molecule level (reviewed in Reference 17).

The fluorescence intensity resolved by wavelength constitutes the fluorescence spectrum. The wavelengths of fluorescence photons contain information about the environment of the fluorophore and the sample heterogeneity. For example, as described above, buried Trp residues tend to have blue-shifted emission bands ($\lambda_{\text{max}} < 330$ nm), whereas Trp residues partially or fully exposed to water have red-shifted emission bands ($\lambda_{\text{max}} > 340$ nm). Therefore, protein conformational changes or unfolding may be accompanied by shifts in the native fluorescence spectra. Fluorescence spectra can be measured on a standard fluorometer, which is available from many manufacturers.

Time-resolved fluorescence: time after excitation

In this section, we discuss methods that detect the time delay between excitation of a fluorophore and arrival of a fluorescence photon. The distribution of $t_i$ times constitutes the fluorescence decay profile of the fluorophore. The average time lag between the excitation event and the emission is the fluorescence lifetime $\tau$ of the fluorophore. The fluorescence decay contains information about dynamic processes that deplete the excited state (Fig. 2a). In time-resolved fluorescence experiments, the fluorescence decay is measured to gain information about these processes.

Often, experimentally measured fluorescence decays do not follow the simple single-exponential form predicted by Equation 2. In many cases, the fluorescence decays are better described by a multieponential decay:

$$I(t) = I(0) \sum_i a_i \exp(-t/\tau_i)$$  \hspace{1cm} (5)

or by a distribution of decay times

$$I(t) = I(0) \int_0^\infty a(\tau) \exp(-t/\tau) d\tau$$  \hspace{1cm} (6)

In Equation 5, $\tau_i$ is the decay time and $a_i$ is the relative amplitude for the $i$th decay component, and in Equation 6 $a(\tau)$ is the normalized distribution of decay times.

The observation of a nonsingle exponential fluorescence decay contains additional information about the nature of the system of fluorophores. Two general reasons exist for observation of nonsingle exponential decays. First, the sample may be heterogeneous, comprising different ground-state conformations or environments that have different intrinsic fluorescence lifetimes (Equation 5) or a distribution of lifetimes (Equation 6). Second, excited-state reactions may exist that are reversible (which allows return to the fluorescent state) or that generate a new fluorescent state that has a different decay rate.

A great deal of attention has been focused on the origin of nonsingle-exponential fluorescence decays in proteins. Some of the most intensely studied cases involve the fluorescence decays of the amino acids Trp and Tyr in proteins. Fluorescence decays for Trp in proteins are almost always multieponential. The origins of the nonsingle-exponential fluorescence decays continue to be debated. In proteins with multiple Trp residues, multieponential decay can be expected because of different environments and different decay times for each Trp. As with the emission wavelength, the fluorescence lifetime of a Trp residue is highly sensitive to its environment, notably the presence of quenching moieties (for example, histidine residues, disulfide bonds, or the carbonyl group of peptide bonds).

Even proteins that contain only a single Trp residue generally exhibit multieponential decays. Several hypotheses have been proposed to explain why. First, multiple conformational states may exist for the single Trp such as different rotameric configurations (orientations about the Trp $\chi_1$ or $\chi_2$ C-C bond) (18). Even in the absence of multiple rotamers, the electron-transfer quenching rate is extremely sensitive to the local environment, so a distribution of local microconformational states may cause a noneponential fluorescence decay. Other possible sources of noneponential fluorescence decay include the response of the protein and surrounding solvent to the change in dipole moment of Trp on excitation (“solvation”) (19).

Time-resolved fluorescence measurement

Two common methods exist for measuring of fluorescence decays in proteins: time-domain and frequency-domain measurements. Signals are processed by time-correlated single-photon counting (TCSPC) for time-domain measurements, or by phase fluorimetry for frequency-domain measurements. For higher time resolution, both time-domain and frequency-domain methods generally employ the same light sources and the same detectors. Although one or the other technique may have some advantages in certain applications, recent analysis showed that the two methods give essentially identical results for a series of fluorescence lifetime standards (20). More thorough treatments of TCSPC and frequency domain fluorometry are available in books cited in “Further Reading.”

TCSPC is illustrated in Fig. 3a. In addition to a mode-locked laser for pulsed excitation and a detector with high time resolution (usually a micro-channel plate photomultiplier tube capable of time-resolution of 20–30 ps), the required instrumentation includes constant-fraction discriminators to generate electrical pulses triggered by fluorescence photons and by the reference (the excitation pulse), a time-to-amplitude converter or other device to measure the time lag between reference and fluorescence counts, and a multichannel scaler to accumulate.
Fluorescence Techniques for Proteins

Fluorescence Techniques for Proteins

Photodiode
Microchannel
Fluorescence
Excitation
Fixed time delay
Timer
Td
Co
Un
Ts
Time delay histogram
Sample
Td
0 1 0.2 0.3 0.4 0.5 0
Time (ns)
320 MHz
32 MHz
Modulation
Response
(a)
(b)
Figure 3
Time-resolved fluorescence techniques. (a) Time-correlated single-photon counting. The sample is excited with a short pulse of light. A fluorescence photon is detected by a microchannel-plate photomultiplier tube, which generates a voltage pulse. A portion of the excitation pulse is split off from the excitation beam and detected by a photodiode to generate a reference pulse. The time lag between excitation pulse and fluorescence photon is determined by a timer. After many counts are detected, the accumulated histogram of time lags gives the fluorescence decay convoluted with the instrument response function. (b) Illustration of the fluorescence response (solid red lines) to modulated excitation (dashed blue lines), simulated for a 1-ns fluorescence lifetime. At low modulation frequencies, the fluorescence response closely follows the modulation. High modulation frequencies generate a greater phase shift and demodulation of the fluorescence response. The phase shift and demodulation are analyzed as a function of modulation frequency to determine the fluorescence decay properties.

The histogram of lag times. Until recently, each device was a separate component. Now, it is possible to obtain the complete TCSPC electronic processing system on a single computer card. Manufacturers include Becker & Hickl (Berlin, Germany) and PicoQuant (Berlin, Germany).

In the frequency domain, fluorescence lifetime measurements are based on the phase shift and demodulation of a fluorescence signal with respect to a modulated excitation beam (21) (illustrated in Fig. 3b). For routine measurements with low time resolution, a fluorescence lamp or continuous laser source can be transmitted through an intensity modulator such as a Pockels’ cell. For high time-resolution, the modulation frequency achievable by this method is not high enough, and the method employs pulsed excitation, just as in TCSPC. The signal analysis consists in measurement of the fluorescence response at a range of frequencies (harmonics of the pulse repetition rate). The phase shift \( \phi \) and modulation depth \( m \) are given as a function of the modulation frequency \( f \) by (\( \omega = 2\pi f \)):

\[
\tan\phi(\omega) = \omega\tau_f \quad \text{and} \quad m(\omega) = \left(\frac{1}{1 + \omega^2\tau_f^2}\right)^{1/2}
\]

Accurate results require measurement of multiple modulation frequencies and fitting the resulting phase shifts and modulations to characterize the fluorescence decay. At low frequencies, the fluorescence signal follows closely the modulation of the excitation pulse. At high frequencies, the time lag between excitation produces a phase shift in the fluorescence signal, and the distribution of time lags over the fluorescence decay manifests itself as a decrease in modulation. The instrumentation required, in addition to the mode-locked laser source and detector, includes mixing and tuning electronics to detect the fluorescence response as a function of frequency.

Both TCSPC and frequency-domain fluorimetry are limited in time resolution by the response of available detectors, typically >25 ps. For cases in which higher time resolution is needed, fluorescence up-conversion can be used (22). This technique uses short laser pulses (usually sub-picosecond) both to excite the sample and to resolve the fluorescence decay. Fluorescence collected from the sample is directed through a material with nonlinear optical properties. A portion of the laser pulse is used to “gate” the fluorescence by sum frequency generation. The fluorescence is up-converted to the sum frequency only when the gate pulse is present in the nonlinear material. The up-converted signal is detected. The resolution of the experiment therefore depends only on the laser pulse widths and not on the response time of the detectors. As a result, fluorescence can be resolved on the 100-fs time scale. For a recent application of fluorescence up-conversion to proteins, see Reference 23.

Fluorescence anisotropy: polarization

Here, we discuss the polarization of fluorescence emission. The spatial orientations of emitting fluorophores determine the polarization of photons emitted. This relationship is the basis of fluorescence depolarization experiments as illustrated in Fig. 4a. When a sample of randomly oriented molecules (e.g., proteins...
Fluorescence Techniques for Proteins

Fluorescence anisotropy is generated by excitation of the sample with vertically polarized light, which excites molecules preferentially whose transition dipole is aligned with the polarization (red arrows), generating an anisotropic distribution of excited molecular orientations. The fluorescence emission is therefore preferentially polarized vertically. The fluorescence emission is anisotropic. As the fluorophores reorient, the fluorescence emission becomes depolarized, and the fluorescence anisotropy decays. Thus, the rate of fluorescence depolarization measures the rate of molecular reorientation, whether by rotational diffusion, domain motion, segmental reorientation, or local reorientational motion of the fluorophore.

Several parameters have been defined to describe fluorescence depolarization. One is the polarization, \( p \), given by:

\[
p = \frac{I_\parallel - I_\perp}{I_\parallel + 2I_\perp}
\]

Jablonski showed in 1960 that the fluorescence depolarization can be described naturally by another parameter, the fluorescence anisotropy, \( r \), defined by:

\[
r = \frac{I_\parallel - I_\perp}{I_\parallel + 2I_\perp}
\]

where \( r \) denotes the steady-state (time-averaged) anisotropy, \( I_\parallel \) is the fluorescence intensity with polarization parallel to the excitation polarization, and \( I_\perp \) is the fluorescence intensity with polarization perpendicular to the excitation polarization. Francis Perrin had derived an equation in 1926 that related the steady-state polarization \( p \) to the fluorescence lifetime. Cast in terms of the anisotropy, this equation takes on an especially simple form:

\[
r = r_0 \left( 1 + \frac{\tau_f}{\tau_r} \right)^{-1}
\]

where \( \tau_f \) is the fluorescent lifetime, \( \tau_r \) is the rotational correlation time (sometimes denoted \( \phi \)), and \( r_0 \) is the short-time limiting anisotropy. Equation 10 can be used to estimate rotational correlation times from steady-state fluorescence measurements.
One application of fluorescence anisotropy is in fluorescence polarization assays, in which steady-state anisotropy measurements are used to detect binding of ligands to proteins (24). Fluorescence polarizations assays are effective for cases in which a fluorescent ligand binds to a much larger object such as a protein. On binding to the protein, the fluorescent ligand experiences a marked increase in its rotational correlation time, which results in an increased steady-state anisotropy. The technique lends itself to high-throughput assays of ligand binding, which permits rapid screening of binding interactions and sensitive determination of binding affinities.

For time-resolved measurements, Equation 9 can be expressed in time-dependent form:

\[ r(t) = \frac{1}{\tau} \langle F(t) \rangle \cdot \langle F(t) \rangle - 1 \]  

(11)

where now the fluorescence intensity decays are measured as a function of time for parallel and perpendicular polarizations. The importance of Equation 11 lies in the relationship of the fluorescence anisotropy to the reorientation of the transition dipole moment:

\[ r(t) = \frac{1}{\frac{3}{2} d^2} \cdot \langle F(t) \rangle \cdot \langle F(t) \rangle - 1 \]  

(12)

where \( r(t) \) is the angle of reorientation of the emission transition dipole, and the brackets \( \langle \cdots \rangle \) denote the average over parallel absorption and emission transition dipole times at \( t = 0 \) (see Fig. 4b). The maximum value for the anisotropy, obtained for parallel absorption and emission transition dipole times, is 0.4, which corresponds to a rotation \( \gamma \) of 3 to 1.

The time dependence of the anisotropy \( r(t) \) depends on the underlying dynamics of reorientational motion. For rotational diffusion (tumbling) of a spherical object, the expected anisotropy decay is exponential with a rotational diffusion time given in the hydrodynamic limit by the Stokes-Einstein-Derjaguin equation. For nonspherical molecules, more complex time dependence may be detected. (For more on these topics, see the book by Cantor and Schimmel in Further Reading.)

Interesting applications of anisotropy decay for proteins often develop from tumbling of the protein as a whole, but from other reorientational degrees of freedom. These motions may include protein domain motions or segmental motions in proteins and peptides. The anisotropy decay in this case is non-single-exponential (see Fig. 4c) and takes the form:

\[ r(t) = r_0 \sum_i a_i \exp(-t/\tau_i) \]  

(13)

where \( a_i \) is the relative amplitude and \( \tau_i \) the correlation time of the \( i \)th rotational component. The slowest rotational correlation time is usually the global tumbling time of the protein. Faster rotational correlation times represent local motions. The amplitude \( a_i \) contains information about the orientational freedom available for the fast motion of the fluorophore (25).

Time-resolved anisotropy decays can be recorded by time-correlated single-photon counting. The fluorescence signals \( I_{\|}(t) \) and \( I_{\perp}(t) \) are both convolutions of the instrument response function (IRF) with the sample response and can be analyzed by nonlinear regression with iterative reconvolution of the fitting function with the IRF. In contrast, the anisotropy \( r(t) \), as calculated according to Equation 11 directly from the raw fluorescence signals, cannot be written as a convolution with an instrument function. For this reason, it is much more reliable to fit the signals \( I_{\|}(t) \) and \( I_{\perp}(t) \) directly rather than to fit \( r(t) \) to determine the anisotropy decay properties.

Fluorescence anisotropy decay can also be measured by frequency-domain methods. In this approach, the polarized fluorescence intensities \( I_{\|}(\omega) \) and \( I_{\perp}(\omega) \) are measured as a function of the modulation frequency of the polarized excitation beam. Even more information about frequency-domain anisotropy measurement and analysis can be found in the monograph by Lakowicz (see Further Reading).

Fluorescence correlation spectroscopy: photon arrival time

The arrival times of fluorescence photons contain information about correlations in fluorescence signals. Fluorescence correlation spectroscopy (FCS) (26) exploits these correlations to measure the magnitude and time scales of fluctuations in fluorescence. These fluctuations contain information about the dynamic time scales of the system and the concentration of fluorescing molecules. Correlations may span time ranges from nanoseconds to milliseconds, which extends the dynamic time window for fluorescence measurements far beyond what is achievable in fluorescence lifetime measurements. The autocorrelation function is calculated as:

\[ G(\tau) = \langle F(t) \cdot F(t + \tau) \rangle / \langle F(t)^2 \rangle \]  

(14)

where \( \langle \cdots \rangle \) is the fluctuation in the fluorescence intensity \( F(t) \) and \( \langle F(t)^2 \rangle \) here denotes the average over \( t \). The recorded autocorrelation functions are typically fit to a function that describes diffusion through a three-dimensional Gaussian observation volume with adjustable parameters, which include the diffusion time \( \tau_d \), the average number of molecules \( \langle N \rangle \) in the focal volume, and the focal volume dimensions \( r_z \).

\[ G(\tau) = \left( \frac{1}{\langle N \rangle} \right) \left( \frac{1}{1 + \frac{\tau}{\tau_d}} \right) \left( \frac{1}{1 + \left( \frac{\tau}{\tau_d} \right)^2 - \frac{\tau_z}{r_z}^2} \right) \]  

(15)

Because the magnitude of the correlation function in the short time limit is inversely proportional to the concentration of fluorophores, FCS can be used to follow changes in concentration. The diffusion time \( \tau_d \) yields the translational diffusion coefficient. If the fluorophore undergoes intramolecular dynamics or photophysical processes, then Equation 15 must be modified accordingly (27). A analysis of the autocorrelation with a modified fitting function can then provide dynamic information about these processes.
Reviews listed in Further Reading provide excellent introductions to FCS. Related techniques have been developed to detect other molecular properties. These properties include fluorescence cross-correlation spectroscopy (FCCS) (28) to detect codiffusing fluorophores and photon-counting histograms (PCH) (29), or fluorescence intensity distribution analysis (FIDA) (29) to distinguish fluorescent species according to their brightness.

Resonance energy transfer: molecular calipers

Förster resonance energy transfer (FRET) is a remarkable tool for detecting changes in the distance between two fluorophores. FRET is the nonradiative transfer of excitation energy from the excited state of one fluorophore (the donor) to another (the acceptor). It is sensitive to the separation between fluorophores on the scale of tens of Ångstroms, a distance range highly useful for proteins. The theory of dipole–dipole coupling between donor and acceptor (separated by distances that are large compared to the molecular sizes of the fluorophores) was developed by Theodor Förster. In this limit, the rate of energy transfer is given by:

$\kappa_{\text{FRET}} = \frac{1}{\tau_D} \left( \frac{R_0}{R} \right)^6$  \hspace{1cm} (16)

where $\tau_D$ is the lifetime of the donor when no acceptor is present, and $R_0$, known as the Förster radius, is the distance between fluorophores where the FRET rate is equal to the intrinsic donor decay rate (i.e., the efficiency of FRET is 50%). A detailed discussion of the theory and application of FRET to proteins can be found elsewhere in this volume.

Single-molecule fluorescence

The intrinsic high sensitivity of fluorescence detection is illustrated dramatically by the detection of fluorescence from single fluorophores. Many fluorescence techniques described above have been applied with single-molecule sensitivity. In heterogeneous samples, where a distribution of molecular properties exists (structure, dynamics, and environment), single-molecule measurements can map out the distribution, which yields information that is not readily available from bulk or ensemble-averaged measurements. Single-molecule detection sensitivity can also be applied fruitfully in cases where the copy number of molecules is low, for example in detection of proteins in the contents of single cells. Single-molecule detection is often carried out with a fluorescence microscope equipped with a sensitive detector such as an avalanche photodiode or a CCD camera. A detailed discussion about single-molecule measurements are available from several reviews (see Further Reading).

Imaging: location of emitted photons

Another area of rapid growth in application of the fluorescence techniques to proteins is in imaging of cells and tissues. Through use of an array detector such as a CCD camera with wide-field illumination or beam scanning with confocal detection, the spatial position of fluorophores can be recorded with a precision determined by the resolution of the microscope. (Single-particle tracking techniques can provide even greater precision by fitting the intensity profile of the emitted light (30).) A wide range of fluorescence techniques is now applied routinely to biologic imaging. These techniques include two-photon fluorescence microscopy, fluorescence recovery after photobleaching, fluorescence lifetime imaging, FRET imaging, and fluorescence anisotropy imaging. An introduction to topics within the broad area of fluorescence imaging techniques is available in many resources. See, for example the book by Pawley listed in Further Reading.

References


Further Reading

Förster Resonance Energy Transfer (FRET) for Proteins

Lambert K. Chao and Robert M. Clegg, Physics Department, University of Illinois-Urbana-Champaign, Urbana, Illinois
doi: 10.1002/9780470048672.wecb171

Förster Resonance Energy Transfer (FRET) is a spectroscopic technique applied throughout physics, chemistry, and biology to measure quantitatively the distance between selected locations on macromolecules and to determine the close association between interacting molecular components. Because FRET typically occurs over distances from 0.5 to 10 nm, it is especially useful for investigating many interesting biological molecular structures. It is also particularly valuable for following the dynamics and structural fluctuations of biological molecular systems. FRET can be applied in solution or under imaging conditions (such as in fluorescence microscopy, nanoscience, and even macroscopic imaging). In this article, we discuss the fundamentals of FRET. These principles apply to every FRET measurement. We present the basic rudiments and the relevant literature of FRET to provide the reader with the necessary background essential for understanding much of the past and modern literature. At the end of the article, we give a short discussion of several applications of FRET to proteins. The literature for FRET is vast, and many new applications are constantly being developed. We could not do justice to the many practitioners of FRET in such a short space, but armed with the background that is presented, we hope this basic information will help readers follow much of the literature and apply it in their own work.

The description of Förster Resonance Energy Transfer (FRET) in a form that is useful for quantitatively interpreting experimental results was first described in 1946 (1) and was later more quantitatively described by Förster in a series of publications (2–10). It has been popular and extensively used in biochemistry since the early 1950s. Many reviews have been published that cover not only the theory and analysis but also the application to protein structures. For the additional perusal of the reader, we list here some selected classic general overviews and discussions of the theory and analysis (11–52). These reviews contain many references to the literature that deal with specific topics, including proteins. In this article, we will concentrate on a discussion of the physical basis of the FRET mechanism and will present a few applications from the literature to determine macromolecular structures. The initial applications of FRET were by physicists and physical chemists. They dealt mainly with solution studies of freely diffusing molecular chromophores and with solid structures. But already in the early 1960s, the power of applying FRET to biological systems was realized: for instance, applications to proteins (16, 20, 23, 34) and to nucleic acids (53, 54). By this time, the theory had been fully developed and tested; however, the applications were hindered by the limitations of a choice of suitable chromophores that could be attached covalently to specific sites of the structures. Thus, many original applications were carried out using either intrinsic chromophores (tryptophan or tyrosine) or dyes that were known to bind noncovalently to protein or nucleic acid structure. Quantitative interpretations of the early experimental results were thereby complex because the placement of the dyes on the biological macromolecules were usually not well known. However, many ingenious analysis methods were developed to extract structural information from the data. The limitation of available chromophores pairs that can be used to investigate structures of proteins has been removed in the last 20 years; a very large number of available chromophores that can be used as extrinsic labels of proteins can now be purchased commercially. All the research areas in this review are being actively and vigorously pursued, and despite the fact that FRET has been used extensively for over 50 years, new methods of measurement and analysis as well as new areas of application are continually being developed. The literature is extensive and sometimes daunting to the newcomer to the FRET field. The
following is an introduction to the basics of FRET, which enables the reader to read the vast, continually expanding FRET literature. We especially emphasize the aspects of FRET that are critical for determining structural and kinetic information about proteins and the biological structures that incorporate proteins.

Examples Representing the Broad Applications of FRET and Proteins

The introduction of fluorescent proteins has been a great boon for use as FRET pairs that can be inserted into protein structures under genetic control. In vivo FRET studies have benefited greatly from the incorporation of the green fluorescent protein (GFP) gene into a host genome (53) to form protein hybrids. This method eliminates the external labeling of organic fluorescent dyes and allows labeling of specific proteins in vivo. Fluorescence lifetimes and photo-physical properties of fluorescent proteins have been characterized (56–65). It is easier to interpret time-resolved FRET studies quantitatively if the donor has only one fluorescence lifetime; although average lifetimes are often used. The original wild-type GFP and GFP variants exhibit complex (multiexponential) decays from their excited states (66, 67), which limit the reliability of lifetime measurements. Fluorescent proteins better suited for fluorescence lifetime imaging (FLI) and FLI-based FRET studies have been obtained by random and site-directed point mutations (63) for a concise informative review of the development of monomeric fluorescent proteins, see Reference 68).

Energy transfer is an integral part of photosynthetic systems (see review chapters in Govindjee et al., (69)). Excitation energy transfer lies at the heart of the phenomenon and its mechanisms (70). The fluorophore of interest is chlorophyll and a few other intrinsic chromophores. The fluorescence intensity and lifetime of plants are tightly coupled to 1) the competition between the rapid shuttling of the excitation energy by FRET, 2) the dissipation of the excitation energy by quenching mechanisms, and 3) the eventual irreversible transfer of the energy into the reaction center, where it initiates the electron transfer chain of photosynthesis (71, 72). As a matter of fact, photosynthesis was one of the initial motivations for developing the dipole-dipole mechanism of FRET (1, 73), and the role of FRET and its mechanism is still a very active research topic in photosynthesis (70, 74–77).

Single-molecule experiments in a microscope with the macromolecules of interest attached to a surface have made extensive use of FRET in the last several years (78–80). FRET allows one to directly observe conformational changes, and if the kinetics take place in the right time range (tens of microseconds to seconds), then the individual steps can be observed, and the kinetic rate constants can be determined. A nother method that has single-molecule resolution is fluctuation correlation spectroscopy (81, 82). Thereby FRET can be used to measure kinetics of protein conformational changes and noncovalent binding reactions as the macromolecule passes through diffraction limited focused laser light in a fluorescence microscope (83). This technique has been further developed to achieve picosecond time resolution, which allows fluorescence lifetime measurements to be made on the diffusing entities, and FRET to be determined from the lifetimes (60, 84, 85). Additional examples of the application of FRET to proteins are given at the end of the article, after discussing the physical basics of FRET.

What is the Basis for the FRET Phenomenon

Hetero-and homo-FRET

As the name implies, FRET involves the transfer of energy from a molecule in an electronically excited state (this molecule is called the donor) to another molecule (the acceptor) that is within a certain distance from the donor. The transfer of energy is brought about by a dipole-dipole interaction between the donor and acceptor. The acceptor is usually a different molecular species, but it can be the same as the donor. If the donor and acceptor are different molecular species, then the energy transfer is called hetero-FRET. If the donor and acceptor are the same, then it is called homo-FRET. For proteins, most applications in the literature use hetero-FRET to determine information about protein structures and conformational changes. In hetero-FRET, the fluorescence intensity of the donor decreases because a probability exists that the donor loses its excitation energy by transferring excitation energy to the acceptor, instead of fluorescing. If the acceptor can also fluoresce and energy transfer takes place, then emission from the acceptor can usually be observed. With true homo-FRET, the intensity of the fluorescence does not change; the energy is simply transferred from one identical molecule to the other, and the probability of emission remains the same. The only way to observe homo-FRET is to measure the decrease in fluorescence anisotropy (polarization). For measuring homo-FRET, the donor molecules are excited with polarized light, just as when measuring normal fluorescence anisotropy. Donor molecules oriented in a direction best suited for absorbing the polarized light are preferentially excited. Because the donor molecules are selectively excited according to their orientation relative to the excitation light polarization, their fluorescence emission will also be polarized preferentially in a direction related to the excitation light polarization. The extent of polarization depends on the rotational correlation time and time the molecules are in the excited state. If homo-FRET can take place, then those donor molecules excited by homo-FRET are not oriented solely relative to the excitation light, but they depend also on the relative orientation of the originally excited donor molecules and the molecules that accept the energy (as we will see in the general formula for energy transfer). This process will decrease the overall polarization of the sample. The measured anisotropy can then be used to interpret the extent of energy transfer.

The efficiency of FRET

The number of energy quanta transferred from excited donors to acceptors divided by the number of quanta initially absorbed...
by donors is called the efficiency of FRET. The maximum of this efficiency fraction is one. Whether energy transfer is more or less likely to occur in a particular situation will depend on what other paths are available for the excited donor molecule to give up its energy and how likely the donor will de-excite via these alternate paths. In other words, FRET is in direct kinetic competition with all other mechanisms of de-excitation of the donor. Therefore, for FRET to take place, the transfer must occur in the same time range, or faster, than all other de-excitation pathways (such as fluorescence emission, dynamic quenching, intersystem crossing to the triplet state, etc.). In practice, the actual transfer process is not measured directly, but is inferred by measuring its effect on other reaction pathways in kinetic competition with FRET. For example, the efficiency of FRET can be determined by comparing donor or acceptor fluorescence intensities in the presence and absence of FRET. This measurement can be done in steady-state or time-resolved experiments (25, 42, 45, 49, 86).

The fluorescence quantum yield of the donor cannot be zero. On the other hand, the acceptor does not have to fluoresce. Because energy is conserved, the transfer must be resonant; that is, the energy lost by the acceptor must equal the energy gained by the acceptor. The probability that the energy can be transferred depends strongly on the distance between the molecules (see Eq. 1). This dependence makes FRET particularly suited for measuring molecular distances and determining spatial proximities.

The distance dependence of FRET and $R_0$

For two isolated molecules (donor and acceptor), the rate of energy transfer $k_{ET}$ from the excited donor to the acceptor is proportional to the inverse sixth power of the distance between the two molecules $R$ and is equal to

$$k_{ET} = \frac{1}{\tau_D} \left( \frac{R_0}{R} \right)^6$$  \(\text{(1)}\)

where $\tau_D$ is the lifetime of the donor excited state in the absence of the acceptor; $R_0$ is the average distance between donor and acceptor molecules when the rate of transfer is equal to $1/\tau_D$. $R_0$ is often approximately 5 nm.

The rate of transfer shown in Equation 1 has an exceptionally strong $1/R^6$ distance dependence. As discussed above, the probability of energy transfer is in competition with all the other pathways of de-excitation. For instance, according to Equation 1, if the molecules are separated by less than about 0.5$R_0$, then the rate of transfer is greater than 65 $\times$ (1/$\tau_D$). Therefore, on the average, essentially all the excitation energy will be transferred from the donor to the acceptor. If the molecules are separated by 2$R_0$ or greater, then the rate of transfer is less than 0.015 $\times$ (1/$\tau_D$). Essentially, on the average, no energy will be transferred. Thus, as a rule of thumb, because of $R_0 \approx 5$ nm, FRET can only be used to determine distances less than 10 nm. For reasons that will be evident later, the lower limit is approximately 0.5 nm.

The constant $R_0$ is dependent on several parameters: 1) the relative orientation of the transition dipole moments of the two molecules (these dipoles are the spectroscopic transition dipoles), 2) the extent that the fluorescence spectrum of the donor overlaps with the absorption spectrum of the acceptor, and 3) the surrounding index of refraction. We will deal with each of these below (see Equation 8). Because many proteins have diameters less than 10 nm, this distance dependence explains the usefulness of FRET for determining distances inside proteins as well as between interacting proteins, which is the reason that the name “spectroscopic ruler” was coined for FRET (20). FRET is a convenient method for determining the distance between two locations on proteins, or for determining whether two proteins interact intimately with each other. Fluorescence instrumentation is available in many laboratories, and a plethora of dyes and a wide variety of fluorescent proteins are now readily available. Therefore, FRET is a viable option for most researchers. With care, FRET can yield valuable information concerning protein-protein interactions, interactions of proteins with other molecules, and protein conformational changes.

Quantitative expressions for measuring the efficiency of FRET

The efficiency of FRET is the fraction of times that an excited donor molecule will transfer its excitation energy to an acceptor. For instance, if the efficiency of transfer is 0.3, and if a molecule is excited 100 times, then on average it will transfer the excitation energy to the acceptor 30 times. A common way to express the efficiency is by means of the rate of energy transfer (Equation 1). After the donor has been excited from the ground state ($S_0$) into its first excited singlet state ($S_1$), the donor can exit the $S_1$ state by several pathways (see Fig. 1a).

As indicated in Fig. 1a, all pathways that lead away from the excited state of a chromophore (either to the ground state—or by some radiative or nonradiative process—or passage to the triplet state) are in direct kinetic competition with each other. FRET is one of these pathways. Each i-th pathway exists with a certain rate constant ($k_i$, in seconds$^{-1}$).

Figure 1b is a schematic that depicts the kinetic pathways. Because they are in direct competition, the probability of going by any single pathway (the efficiency of that pathway) is the ratio of the rate of that pathway divided by the sum of the rates of all the pathways.

The efficiency of energy transfer, Fig. 2a, can be expressed as

$$E_{FRET} = \frac{\# \text{ times FRET pathway is chosen}}{\text{total \# of times the molecule is excited}} = \frac{k_{FRET}}{\sum k_i}$$  \(\text{(2)}\)

where $k_{FRET}$ is the rate constant for FRET, and the $k_i$s are rate constants (with units of s$^{-1}$), and the only ones that are included in Equation 2 are those that emanate...
Forster Resonance Energy Transfer (FRET) for Proteins

Figure 1 - FRET and competing pathways. (a) A Perrin–Jablonski diagram shows common spectroscopic transitions for de-excitation of the donor fluorophore. Absorption of a photon excites the donor from the S0 to the S1 excited state, where it rapidly relaxes (typically in less than 1 ps) to the lowest vibrational level of the S1 state, which is known as internal conversion (IC). Internal conversion involves energy loss through vibrational interactions with the surroundings. The excited molecule is in a higher vibrational state of S1; thereby releasing heat to the surroundings, finally ending up in the lowest vibrational state of S1. The donor can then undergo de-excitation from the lowest vibrational level of the excited state to the S0 ground state through several pathways. The typical fluorescence lifetime (independent of the other pathways) is on order of 1–10 ns. Through spin–spin or spin–orbital interactions, the S1 state can undergo intersystem crossing (ISC) into the excited triplet state where it spends some time (typically 10–100 s, depending on the concentration of oxygen in the solution) before phosphorescence or conversion to the ground state by internal conversion takes place. The ground state of oxygen is a triplet, and it can easily react with the triplet state of the chromophore, producing reactive oxygen species (radicals) that can collide with the chromophore, destroying it (photolysis). (b) This panel shows a schematic of the various competing pathways for leaving the excited donor state, including quenching, photolysis, fluorescence, and phosphorescence, which compete with FRET. Dexter transfer involves energy transfer to the acceptor molecule by exchange of electrons (see text) when the electronic orbitals of the donor and acceptor overlap. By measuring the effect of FRET on one of the competing processes (e.g., donor fluorescence), one can measure the FRET efficiency. It is not necessary to measure fluorescence to determine FRET efficiency, but this technique is the normal way. It is also possible to measure the fluorescence intensity of the acceptor (if the acceptor can fluoresce) to detect and quantify FRET.

The lifetime of an excited molecule (in seconds) is the inverse of the sum of the kinetic rates (in s⁻¹) for all the pathways for exiting the excited state. Thus, in the absence of an acceptor, the lifetime is

$$\tau = \left( \sum_{i \neq \text{ET}} k_i \right)^{-1}$$  \hspace{1cm} (3)

And in the presence of an acceptor, the lifetime is

$$\tau_{\text{a}} = \left( \sum_{i \neq \text{ET}} k_i \right)^{-1}$$  \hspace{1cm} (4)

Therefore, the efficiency can also be given in terms of the measured fluorescence lifetimes in the presence and absence of an acceptor

$$E_{\text{FRET}} = \frac{1}{\tau_{\text{a}}} - \frac{1}{\tau} = \frac{1}{\tau_{\text{a}}} - \frac{1}{\tau_{\text{a}} + \tau} = \frac{1}{\tau_{\text{a}}} - \frac{1}{\tau_{\text{a}} - \tau_{\text{a}}}$$

Thus one can determine the efficiency of FRET by measuring the nanosecond fluorescence lifetime of the donor in the presence (A) and the absence (A⁻) of the acceptor. If one has the possibility of making this measurement, then it is a convenient, accurate, and robust way to measure the efficiency of FRET because the lifetimes can often be determined accurately, and not many corrections or experimental controls need to be made when calculating the efficiency.

Several ways exist to determine the efficiency by measuring steady-state fluorescence. It is easily observed by inspecting the first equality in Equation 2. The best way to acquire steady-state values of the fluorescence intensity is photon counting. We will not go into the way it is done with hardware and electronics (it is usually done automatically with modern instrumentation), but simply note that by using photon counting one can easily...
Förster Resonance Energy Transfer (FRET) for Proteins

Measure a correct value of the intensity at any wavelength by counting the number of photons emitted. If one does not count photons, then the light detector must be corrected for the sensitivity of the photomultiplier at different wavelengths. We can rewrite Equation 2 as

\[ E_{\text{FRET}} = \frac{\text{# times FRET pathway is chosen}}{\text{total # of times the molecule is excited}} = \frac{\#(\text{photons}_{\text{B,A}}) - \#(\text{photons}_{\text{B,A}})}{\#(\text{photons}_{\text{B,A}})} \]

\[ = 1 - \frac{\#(\text{photons}_{\text{B,A}})}{\#(\text{photons}_{\text{B,A}})} \]

\[ #(\text{photons}_{\text{B,A}}) \text{ means the number of photons emitted by the donor in the presence or absence of the acceptor, respectively. This "counting" quantification is very important when measuring the efficiency of FRET, because the quantities of interest in Equation 6 are the number of photons emitted and the number of photons absorbed, and not their energy. The experiment can be conducted either on an ensemble of molecules (where different donor molecules will be excited by each excitation event) or by repetitive experiments on a single molecule. Thus, the numerator of Equation 6 is just the number of photons that the donor emits in the absence of an acceptor minus the number of photons emitted in the presence of an acceptor. This difference is just the number of "quanta" that are transferred to the acceptor, which would have been emitted as fluorescence photons if the acceptor were not there. The denominator is then simply the number of photons emitted by the donor in the absence of an acceptor. Thus, when the acceptor is far away, the efficiency is zero. When the acceptor is very close to the donor, the efficiency becomes one. Note the exact parallel of Equation 6 to Equation 5 (i.e., \( E_{\text{FRET}} = 1 - \frac{\#(\text{photons}_{\text{B,A}})}{\#(\text{photons}_{\text{B,A}})} \)). This parallel relationship occurs because the lifetime of the donor emission becomes shorter when another pathway (for instance, \( k_{\text{on}} \)) opens up for an escape from the excited state. That is, the average time in the excited state becomes shorter. It reduces the number of photons emitted by the donor by transferring a certain number of energy quanta from the excited donor to the acceptor. Often one measures the intensity of fluorescence. The intensity is proportional to \( \#(\text{photons}) \text{ per second}. \) Then, provided the concentration of the donor is the same in both measurements, and the conditions of measurement remain identical, the efficiency is

\[ \frac{\text{#(photons)}_{\text{B,A}}}{\text{#(photons)}_{\text{B,A}}} = \frac{1}{1 - \frac{\text{#(photons)}_{\text{B,A}}}{\text{#(photons)}_{\text{B,A}}}} \]

\[ = 1 - \frac{\text{#(photons)}_{\text{B,A}}}{\text{#(photons)}_{\text{B,A}}} \]

\[ l_{\text{D,A}} \text{ is the measured fluorescence intensity of the donor in the presence and absence of the acceptor, which is measured under identical conditions and with the same donor concentration. If the concentration of the donor is not identical in both measurements, then it can usually be corrected simply by multiplying by the corresponding concentration ratio.}

Conditions for FRET and the Effect of the Transition Dipoles Orientations

The value of \( R_0 \) for a singular pair of D and A molecules is

\[ R_0 = \frac{9000 \ln(10) \phi_0}{N_A \lambda_0 \eta^2} \]

\[ \left[ \int_{\lambda_0}^{\infty} \frac{F(\lambda)}{\lambda^4} \lambda^4 \left( \frac{\pi n^2}{\lambda} \right)^4 \right] \]

where, \( \lambda \) is wave number (in cm\(^{-1}\) units), \( \phi_0 \) is the quantum yield of the donor, \( N_A \) is Avogadro’s number, \( n \) is the index of refraction pertaining to the transfer, \( \epsilon(\lambda) \) is the molar absorption coefficient of the acceptor (in units of cm\(^{-1}\)mol\(^{-1}\)), \( F(\lambda) \) is the fluorescence intensity of the measured fluorescence spectrum of the donor, and \( e^2 \) is an orientation factor that results from the inner product between the unit vector of the electric near field of the donor dipole and the unit vector of the absorption transition dipole of the acceptor (see the Orientation Factor section below). The ratio of integrals in the bracket has units of cm\(^3\)mol. Using the units given in the paragraph above, we have:

\[ R_0 = 8.79 \times 10^{-25} \Phi_0 \lambda_0 \frac{\pi}{2} \eta^2 \text{ cm}^6 \]
Forster Resonance Energy Transfer (FRET) for Proteins

**Figure 3** The electric field of the donor and the orientation factor. (a) The donor dipole \( p_D \) in the electric field \( E_D \) of the donor \( p_D \) and the donor dipole orientation \( \theta_D \) (the angle between \( p_D \) and \( r \)). (b) This schematic shows the donor and acceptor dipoles and illustrates the 
angles and radial vectors used in the definition of the orientation factor \( \nu^2 \). The coordinate system is chosen such that the \( p_D \) and \( E_D \) vectors are in the \( r \)-y plane, the \( p_A \) vector can be in any direction, and is not supposed to be in either the \( r \)-y or \( r \)-x planes. The dipole field is symmetrical about the azimuthal angle of the \( p_A \) vector.

### Distance dependence

For energy to be transferred from the donor to the acceptor, an interaction must occur between the two molecules, and the energy must be conserved (i.e., the energy lost by the donor equals that gained by the acceptor). Coulomb (charge-charge) interactions between the electron distributions in both molecules are responsible for the FRET interaction, and the energy transfer can be understood as an interaction between two classical oscillating dipoles. Actually, a classical derivation of the FRET equations (87, 26) results in the identical expression as a quantum mechanical derivation (3, 48). The electric field of an oscillating dipole is very large in the “near-field” region where the field can be described by a pure dipole field (Fig. 3a). A dipole field at some point in space a distance \( R \) from the center of the dipole is dependent on the angle relative to the dipole direction, and it varies with distance as \( 1/R^3 \). The interaction energy between the two dipoles therefore varies as \( 1/R^6 \). The interaction energy also depends on the relative orientations of the two dipoles and the orientation of each dipole to the separation vector between the two dipoles (see Figs. 3a and 3b). The dipole-dipole nature of the interaction explains the \( 1/R^6 \) dependence of the rate of energy transfer (Eq. 1).

### Overlap integral

It is important to realize that in this near-field zone of the dipoles (which extends maximally 10 nm for FRET) no photon exists; the lack of photons in the near-field zone is essentially because of the Heisenberg uncertainty principle of quantum mechanics (88). That is, the energy is transferred without the emission or absorption of a photon. Of course, the energy transferred from the donor must equal exactly the energy transferred to the acceptor. However, it turns out that the interaction between the two molecules in the near-field zone can be described as if the acceptor is bathed in an oscillating electric field of the near-field zone of the donor, \( E_D \) in Equation 10. This oscillating electric field (with a frequency equal to that of the fluorescence of the donor emission, \( F_D \) of Equation 8) is in the near-field zone of the donor where no photons exist and not in the far field zone where photons exist. The energy of such a Hertzian oscillating dipole is not emitted as photons until a distance of about one wavelength away from the donor. Energy is conserved during the transfer event; that is, the energy lost by the donor must equal the energy taken up by the acceptor. The change in the energy levels of the donor and the acceptor are the same energy levels that correspond to the spectroscopic transitions of the donor (the emission spectrum) and acceptor (the absorption spectrum). Therefore, to conserve energy, the emission (fluorescence) spectrum of the donor \( F_D \) must overlap with the excitation (absorption) spectrum \( \varepsilon_A \) of the acceptor (Fig. 2b). The more these spectra overlap, the stronger is the energy transfer; that is, a larger spectral overlap leads to a larger \( R_0 \). The \( F_D \) and \( \varepsilon_A \) spectra (Fig. 2b) are those of the separate donor and acceptor components. These spectra must be the appropriate spectra that correspond to the identical conditions as where the FRET measurements are made. Often the spectra can simply be taken from the literature that refers to the separate dyes. But often, the dyes physically interact with the proteins, which changes the dye’s absorption or emission spectra. The overlap integral involves the weighting function \( v^2 \) in addition to the emission spectrum of the donor and the absorption spectrum of the emitter (see Eq. 8). As mentioned above, although no photon is emitted or absorbed in the transfer process, the emission spectrum of the donor and the absorption spectrum of the acceptor are still involved in the overlap integral. This participation of the optical spectra is because FRET involves changes in the energy levels of D and A that are identical to those in normal absorption and emission events. An important condition for FRET to occur is that the interaction between the donor and acceptor is very weak, and the two molecular species retain their separate electronic and vibrational structures and energy levels. In FRET, the donor and acceptor are very weakly perturbed by dipole interactions, which is formally the same type and strength of interaction that describes the interaction of the chromophores with light. Therefore, the energy transitions, which can occur in the donor and acceptor molecules during FRET, are the same as their respective spectroscopic transitions when absorbing or emitting photons. This requirement for the conservation of energy, which is expressed by the overlap integral in the expression for \( R_0 \) (Eqs. 8 and 9), is also the reason for the word “resonance” in FRET. We emphasize once more: Although the emission and absorption spectra of the donor and acceptor are in the overlap integral, the FRET process does not involve the emission or the absorption of a “photon.”
The most difficult factor to control and usually the hardest to action of the donor electric field near-field zone of the donor (by studying the interaction of the acceptor dipole in the description of FRET) is also perfectly consistent with a quantum the near-field of the donor is not propagating, and the field diagonalizes from the oscillating electric field of the donor Hertzian dipole. And this field were equivalent to the oscillating electric field of the donor with the dipole moment of the acceptor. The equations for FRET have been derived classically (87, 26, 48) by assuming that the donor molecule is an oscillating point dipole [a Hertzian dipole (89)]. One assumes that the acceptor is located in the “near-field zone” (which extends much less than one wavelength of light of the corresponding frequency) of the oscillating dipole of the donor (Fig. 3). The acceptor absorbs the energy by interacting with the oscillating near field of the donor (the donor oscillates at the same optical frequency where the acceptor absorbs). The near field is not radiating (no transverse photon emission occurs in the near-field zone of a Hertzian dipole). And the electric field in the near field has both longitudinal and transverse components, in contrast to the far field zone, where the electric field has only transverse components. Nevertheless, the mechanism of absorption of energy by the acceptor from the oscillating electric field of the donor Hertzian dipole is identical to the mechanism of absorption of light of the same frequency. Therefore, the interaction of the “transition moment” of the acceptor with the electric field of the donor obeys the same rules as normal absorption of light by the acceptor (polarization dependence and all the spectral requirements for normal absorption). The rate of absorption is proportional to the square of the vector dot product of the acceptor transition dipole with the electric field of the donor, just as though this field were equivalent to the oscillating electric field of light impinging on the acceptor. However, the electric field in the near-field of the donor is not propagating, and the field direction is longitudinal as well as transverse vector components (propagating light photons has only vector components transverse to the direction of propagation). This electric field is a reflection of the fact that in the near-field zone no photon could even exist according to the uncertainty principle. This simple classical description (which results in the correct theoretical description of FRET) is also perfectly consistent with a quantum derivation.

The orientation factor can be best understood quantitatively by studying the interaction of the acceptor dipole in the near-field zone of the donor (Fig. 3a). The energy of interaction of the donor electric field $E_D$ and the acceptor dipole moment $\beta_A$ is $E_D \cdot \beta_A$. The rate of absorption is proportional to the square of this energy of interaction. Therefore, the rate of energy transfer is proportional to $(E_D \cdot \beta_A)^2$ (Eq. 11).

The field $E_D$ that surrounds an oscillating classical electric dipole $\beta_D$ is shown in Fig. 3a.

$$E_D = \frac{|\beta_D|}{r^2} \left(3 \cos \theta_D \cdot \hat{r} - \beta_D \right)$$

(10)

where $|\beta_D|$ is the time independent dipole strength, $r = |\vec{r}|$ is the distance from the point donor dipole (in FRET it is the distance from D to A), and $\hat{r}$ is the unit vector pointing from the donor dipole to the position $\vec{r}$, where the acceptor is located. $\theta_D$ is the polar angle between $\hat{r}$ and $\vec{r}$, and $\beta_D$ is a unit vector perpendicular to $\hat{r}$ that points in the direction of increasing $\theta_D$ (Fig. 3a). The caps designate unit vectors. Figure 3b shows the juxtaposition of two dipoles, and it defines the parameters used in the Equations 10, 11, and 12. As we said above, $k_{ET}$ in Equation 1 is proportional to $(E_D \cdot \beta_A)^2$, according to classical electrodynamics. We can write

$$k_{ET} \propto \left( E_D \cdot \beta_A \right)^2$$

(11)

where $k_{ET}^2$ has been defined as

$$k_{ET}^2 = \left[ 2 \cos \theta_D \cdot \hat{r} - \beta_D \cdot \sin \theta_D \cdot \hat{r} \cdot \beta_A \right]^2$$

(12)

The orientation factor $k_{ET}^2$ can have values between 0 and 4. We have given $k_{ET}^2$ in two different representations in Equation 12.

Thus, the rate of energy transfer depends on the square of the dipole product between the acceptor dipole (transition dipole) $\beta_A$ and the field $E_D$ of the donor dipole (transition dipole), $\beta_D$ (Eq. 11 and Fig. 3b). For any chosen locations and orientations of the donor and acceptor, the value of $k_{ET}^2$ involves the cosine of the angle between the unit vectors $\beta_D$ and $\beta_A$ (i.e., $\beta_D \cdot \beta_A$) as well as the cosine of the angles between $\hat{r}$ and $\beta_D$ (i.e., $\hat{r} \cdot \beta_D$) and between $\hat{r}$ and $\beta_A$ (i.e., $\hat{r} \cdot \beta_A$). Therefore, for any constant selected angle between the donor and acceptor dipoles (that is, constant $\beta_D \cdot \beta_A$, the value of $k_{ET}^2$ will depend on the position in space where the acceptor dipole is relative to the donor. The strength of the field of the donor molecule for any particular constant values of $\beta_D$, $\beta_A$, and $\beta_D \cdot \beta_A$ changes with the distance $|\vec{r}|$ as $1/|\vec{r}|$, that is, for any particular direction of $\vec{r}$ relative to $\beta_D$. As illustrated in Fig. 3a, for a particular angle between the orientations of the donor and acceptor dipoles $\beta_D \cdot \beta_A$.

---

WILEY ENCYCLOPEDIA OF CHEMICAL BIOLOGY © 2008, John Wiley & Sons, Inc.
the angle between the acceptor dipole and the electric field of the donor depends on the position in space of the acceptor. Also, for constant relative orientations of the donor and acceptor dipoles (constant $\hat{p}_D \cdot \hat{p}_A$), and for constant $\cos \theta_D \cdot \cos \theta_A$, $x^2$ is constant (Eq. 11 and 12). In this case, the rate of FRET is solely a function of the distance between the donor and acceptor. However, if only $\theta_D$ and $\theta_A$ are constant, then the distance dependence is more complicated than just the distance between the donor and acceptor. This intertwining relationship between the distance separating the donor and acceptor and the orientational dependence of $x^2$ illustrates the complexity of determining a value for $x^2$.

Because the orientations and spatial locations of the two chromophores may vary over the ensemble of molecules, and because they can also change during the time the donor is in the excited state, the measured effect of $x^2$ is usually an average over the appropriate spatial/temporal distributions. Whenever $E_D$ and the acceptor dipole moment $\hat{p}_A$ have parallel orientations, the rate of FRET is maximum for that placement in space for the acceptor relative to the donor. For $\hat{p}_A$ oriented parallel to $E_D$, the possible maximum values of $x^2$ are between 0 and 4; that is, the actual maximum value depends on the value of $\theta_A$ (see below). The minimum value of $x^2$ is zero for every position of the acceptor relative to the donor whenever $E_D$ and the acceptor dipole moment $\hat{p}_A$ are oriented perpendicular to each other.

The possible introduction of error in the estimation of the FRET efficiency from the orientation factor is often of concern when measuring FRET in proteins or between proteins. This uncertainty can occur because the actual distribution of dye orientations is often not known, or because the dyes do not rotate freely and rapidly relative to the fluorescence lifetime of the donor. If the donor and acceptor molecules undergo rotational or translational movements during the time the donor is in the excited state, or if an ensemble of different donor and acceptor orientations exist, then the value of $x^2$ (and of course the rate) will be averaged over the corresponding ensemble of configurations. Because $x^2$ can, in principle, range from 0 (e.g., acceptor absorption dipole parallel to the electric field of the donor) to 4 (e.g., end-to-end stacked parallel dipoles), the variation in the rate of FRET can be extensive because of such movements. This possible variation in $x^2$ becomes especially apparent when one notes that the measurement of the rate of transfer (or the efficiency) varies directly as $\kappa^2$ (90), and therefore directly with $x^2$ (see below). If the condition of very rapid (compared with the lifetime of the donor) rotational movements of both donor and acceptor is met, then $x^2$ is rigorously $2/3$, which arises from averaging over all possible orientations. If rapid re-orientation through all angles is not the case, then limitations on the degree of rotational freedom can have significant effects on the measured efficiency of energy transfer. However, as we discuss below, the assumption that $x^2 = 2/3$ is often a justified, and reasonable one to make (39, 48, 86).

It is worthwhile to consider a few simple examples to get a feeling for the values of $x^2$. If the two dipoles have orientations in space perpendicular to each other, and if the acceptor dipole is juxtaposed next to the donor, but in the direction perpendicular to the direction of the donor dipole, then $x^2 = 0$ (that is, because, $\theta_D = \pi/2$, so $\cos \theta_D = 0$, and $\theta_A = 0$). However, this example is only a special case where $x^2 = 0$. As was pointed out above, for any position in space of the acceptor molecule, $x^2$ will equal zero for all orientations of $\hat{p}_A$ where $E_D \cdot \hat{p}_A = 0$. And for most of these positions where $E_D \cdot \hat{p}_A = 0$, the donor and acceptor dipoles are not perpendicular to each other (see Fig. 3a). Conversely, if the donor and acceptor dipoles are perpendicular, then most locations of the acceptor relative to the donor will have $x^2 = 0$. This relationship is easiest to observe by looking at the second equality in Equation 12 or by examining Fig. 3a. Another simple example is when the dipoles are parallel ($\hat{p}_A \cdot \hat{p}_D = 1$). Then, if $\theta_D = 0$ and $\theta_A = 0$ (parallel dipoles, stacked on each other), then $x^2 = 4$. But when $\theta_D = \pi/2$ and $\theta_A = \pi/2$ (again parallel dipoles, but now next to each other), then $x^2 = 1$. In the latter case, $\hat{p}_A$ is parallel to $E_D$ but $\hat{p}_D = \hat{p}_A \cdot r = 0$, and the value of $x^2$ is four times smaller than when $\cos \theta_D = \cos \theta_A = 0$. These few examples demonstrate the complexity of the behavior of $x^2$. In Fig. 3a, we have indicated both the orientation of $\hat{p}_D$ and $E_D$, as well as the angle $\theta_A$ (where $\cos \theta_D = \hat{p}_A \cdot r$). More thorough discussions of $\kappa$ can be found in the literature (35, 48, 51).

Because fluctuations always exist in positions and angles of the D and A molecules, the actual value of $x^2$ is an ensemble average or a time average. The most commonly used average value is $x^2 = 2/3$. As already mentioned, this result is rigorously true if during the excited state lifetime of the donor the orientations of the donor and acceptor can each individually reorient fully in an independent random fashion. However, even when this condition is not met (for instance when the anisotropy of the dyes is not close to zero), it has been found that the approximation $x^2 = 2/3$ is often satisfactory (39, 42, 45, 51, 90). It is often discussed in the literature as though $x^2 = 2/3$ pertains only to the case of very rapidly rotating D and A molecules. This statement is not true, because depending on the placement of the dyes and their relative orientations, it is possible for $x^2 = 2/3$ at every location of the acceptor relative to the donor, even when the dyes cannot rotate at all. Also, many dyes used for FRET have more than a single transition dipole, which can be excited at the same wavelengths. Because different transition dipoles of a fluorophore are usually not parallel to each other (they are often perpendicular to each other), the presence of multiple transition dipoles leads again to an averaging of $x^2 = 2/3$. Finally, if one is interested in detecting a change in structure or extent of interaction (binding), then one may not be interested in exact estimates of $x^2$. For instance, when fluorescent proteins are used in FRET experiments, $x^2$ can become a very important variable, and averages are often not applicable. This situation occurs because the chromophores are fairly rigidly held in the fluorescent protein structure, and the fluorescent proteins may have specific interactions either with each other or with other components of the complex under study (91, 92).
higher indices of refraction in the molecular surroundings (the solvent). A high index of refraction—which is equal to the square root of the dielectric constant—means that the electrons in the molecules of the solvent are freer to respond to an electric field than for solvents with lower indices of refraction. The solvent is assumed not to absorb at the wavelengths in question, so the index of refraction is real (not a complex number). This screening from the response of the solvent molecules leads to a damping of the extent of the field, and therefore the two dipoles (donor and acceptor) must be closer together in a solvent with higher index of refraction to have the same strength of interaction as in a solvent with lower index of refraction. Note that the wavelengths are also shorter in a higher index of refraction; however, remember that FRET does not involve propagating light fields. The actual situation is somewhat more complex, and the reader is referred to an excellent discussion in the recent literature (93-95), which clears up common misconceptions as to the origin of the effect of the index of refraction for FRET. In addition, the index of refraction varies significantly over short distances in and on the surface of proteins (96); for very short distances, the concept of an index of refraction becomes suspect. However, usually an average value is chosen between 1.33 and 1.5, which correspond to values of water and crystals of polypeptides. The value chosen usually depends on whether the dyes are in direct contact with water or are inside the protein, where the dyes are removed from water. Detailed numerical methods (96, 76, 97, 70) can be used to take into account directly the interaction between charges if the protein structures are known and if the position and orientations of the chromophores are known, which extensive numerical calculations, whenever the structures are well known, avoid the use of an “effective” index of refraction.

Quantum yield of the donor

Equation 8 shows that \( R \) is dependent on the quantum yield of the donor \( \phi_D \). This dependence is not strong; \( R \propto \phi_D^{1/6} \). For instance, \( R(\phi_D = 0.1) / R(\phi_D = 0.01) \approx 0.7 \). Therefore, small changes in the quantum yield of the donor will not make large differences in the value of \( R \). Of course, if increasing \( \phi_D \) by a factor allows one to make accurate measurements of distances or to observe conformational changes, then the change can be significant. However, the range of distances observable by FRET are not a strong function of the quantum yield of the donor. However, small displacements in the horizontal placement of the efficiency curve of Fig. 2a can significantly increase or decrease the sensitivity of a FRET measurement. In some circumstances, it is advisable to choose a donor-acceptor pair that have a smaller \( R \) preferably in the range of the distance one is interested in measuring (see section below on “What is important, \( R \) or \( \phi_D \)?”). One can sometimes shorten \( R \) by simply adding a collisional dynamic quencher to the solution, which will lower the quantum yield of the donor in the absence of the acceptor, lowering \( \phi_D \), thereby adjusting \( R \) to lower values (90).

Transfer at very short distances

The Förster transfer mechanism (Eqs. 1, 8, and 10) is valid at all distances in the near-field region where the approximation of point dipoles is valid. At very small distances from the molecular dipoles, the finite extended distribution of the electrons make the point dipole approximation invalid (98, 99). In this case, the higher order terms such as quadrupoles and octapoles would have to be taken into account (14, 15). In addition, if no electric dipole interactions occur between the donor and acceptor (for symmetry conditions, making the transition dipole zero), then these electric quadrupole terms and interactions between the magnetic dipole and electric dipole terms have to be taken into account. The lack of electric dipole transitions is seldom the case, and the use of quadrupole terms has not yet been important for the interpretation of FRET in proteins. However, if the two participating dipoles are very close, then transfer by electron exchange [Dexter transfer (14, 15)] also presents a possible pathway for energy transfer. This type of energy transfer should not be referred to as FRET because it is a different mechanism than Förster Transfer and requires a partial overlap of the electronic orbitals of the donor and acceptor. The transfer probability drops off exponentially, following the exponential decrease in the overlap of the wave functions of the electronic orbitals. However, even in photosynthesis, where the chlorophylls are very close and oriented to maximize energy transfer, it has not been shown unequivocally whether Dexter transfer is an important pathway for energy transfer in the photosynthetic unit. Nevertheless, one should be aware of these other mechanisms of energy transfer.

Noncoherent energy transfer (FRET) or coherent transfer

Lately there has been increased interest in coherent transfer. Förster transfer assumes that interactions with the molecular surroundings, or the internal vibrations of the molecules, are sufficiently rapid so no correlation exists between the act of excitation of the donor and the act of transfer. In other words (to give an anthropomorphic analogy), during the time when FRET can take place the donor shows no memory of how, or when, it was excited. This lack of memory is attributed to the chaotic, randomizing effects take place in subpicosecond times, rapidly relaxing the initially excited molecule to the lowest vibrational state of the excited electronic state (before fluorescence emission or FRET). A similar process takes place after the energy is transferred to the acceptor; the original acceptor state immediately following the transfer is rapidly randomized so that the transfer is irreversible. Whenever this randomization occurs, energy transfer can be analyzed from a probability viewpoint (Förster transfer), as is the case in the derivation of the Equations 2-7, 8, and 11. That is, in Förster transfer we consider the probability per unit time for each pathway where the probabilities are independent of time and independent of each other. This probabilistic treatment of all the rates of deexcitation from the excited state for Förster transfer holds true whether the donor and acceptor are identical (homo-FRET) or different molecules (hetero-FRET); if
molecules (or atoms, with which most of these coherent transfer experiments have taken place) are completely isolated from the environment (such as in a vacuum at very low temperatures, e.g. well below 1 K), or if one is observing the fluorescence emission in the femtosecond time range at low temperatures (\(\sim 77 \, \text{K}\)), then it is possible to observe coherent transfer. Under the right conditions, the energy oscillates between different states (with the energy localized on different molecules or atoms) as a function of time; that is, the location of the excitation energy is not localized on just one chromophore independent of time. Instead, the energy is distributed between different chromophores and oscillates from one to the other. These oscillations can be observed in the time dependence of the fluorescence emission. Analogous oscillations in the transfer of energy between different interacting molecules in resonance (even without emission) were first described by Schrödinger (100). Until recently, it was not considered relevant for protein systems; however, lately these oscillations have been observed for photosynthetic systems at low temperatures (101, 102). How important these coherent oscillations are for the mechanism of photosynthesis at normal temperatures remains to be determined. But the effect has created much interest. The question is whether this coherent transfer, which is very rapid and can therefore transfer energy over large distances before the processes that lead to transfer becomes incoherent, plays a major role in the efficiency of photosynthesis.

What is Important: \(R_0^2\) or \(R_0\), and When?

Whether \(R_0^2\) or \(R_0\) are more important, may sound like a silly question; however, a subtle difference determines when one may be of more interest than the other. Usually, \(R_0\) is stressed. \(R_0\) is the distance where the rate of transfer is half the rate of de-excitation from the excited state (\(1/2\) in Eq. 1).

One usually tries to choose dye pairs (donor and acceptor) such that the expected range \(R_0\) is in the most sensitive part of the efficiency curve (Fig. 2a). Choosing appropriate donor and acceptor molecules can be done by trial and error, researching the literature, or determining the variable parameters in Equation 8 and calculating \(R_0\). The variable experimental parameters that define \(R_0\) are: \(|\text{D}|, \alpha, \kappa, F, n, x, \varepsilon\), and \(v_0\). To make calculations of \(R_0\), one usually chooses \(x^2 = 2/3\), and then corrects the \(x^2 = 2/3\)-calculated \(R_0\) by the appropriate factor if it is suspected that \(x^2 \neq 2/3\). Because \(R_0\) is proportional to the \(1/6\) power of \(|\text{D}|, \alpha, \kappa, F, n, x^2\), and \(v_0\), the value of \(R_0\) is not a strong function of these parameters; therefore, variations in their values do not change the value of \(R_0\) very much. As we pointed out above, \(R_0\) is proportional to the \(4/6\) power of the index of refraction, \(n\). So a fractional change in the index of refraction can have a more significant effect on \(R_0\). For instance, a 20% change in \(n\) will change \(R_0\) by only 3%. However, a 20% change in \(n\) will change \(R_0\) by 13%. So, all the variations in \(R_0\) are smaller than the corresponding variation in the parameter itself. Even if the parameters change by a factor of 2 (except for \(n\)), \(R_0\) will only change by 12%, which is the reason that the \(R_0\) of many dyes are in the same range. So, one might get the idea that variations in these parameters will be insignificant compared with the expected (or searched for) changes in distance.

However, the measured variation in the range of transfer (Eq. 1) is proportional to \(R_0^2\), which means that variations in these parameters could have a much more significant effect on the measured efficiency. If \(R\) is not approximately within \(R_0 \pm 0.5R_0\), then the measured efficiency (Eq. 2) will not depend on variations or changes in these parameters significantly (Fig. 2a), because the efficiency is within about 8% of the minimum or maximum value, and such small changes may be hard to measure. However, if one is in the range \(R \approx R_0 \pm 0.5R_0\), then the changes in any of the parameters \(|\text{D}|, \alpha, \kappa, F, n, x^2\) and \(v_0\) can lead to significant changes in the efficiency that could be confused with changes in \(R\). The sensitivity to such conformational changes (if the dyes are chosen with judicious values of \(R_0\)) is one of the valuable characteristics of FRET. This possibility should be carefully considered, especially whenever one is using FRET to measure changes in protein structure and wants to interpret the results as a change in \(R\) (see, for instance Reference 90). In any case, as is clear from looking at Equation 1, one achieves maximum efficiency for observing small changes in \(D\)—A distances when \(R\) is approximately equal to \(R_0\). The take-home message is that when choosing dyes for specific cases, it is worthwhile to consider carefully the choice of dyes, investigate their spectroscopic properties, and calculate the \(R_0\). It is not necessarily the best choice to select dye pairs that show the maximum values of \(R_0\).

Examples of FRET and Proteins

For over 50 years, FRET has been applied to protein and peptide structures (103, 104). It would be impossible to review even a small selection of the vast literature, and the following list of applications does not attempt to present details of the methodologies or the results. We will present just a few selected publications to give a flavor of the type of problems where FRET is applied, and from which the interested reader can obtain references.

FRET has been shown to acquire reliable estimates of protein and other macromolecular structures (27, 34, 53, 90, 103-108). Attention must be paid to the relative orientations of the transition dipoles of the donor and acceptor (28), but it has also been shown that for many biological systems, the approximations \(x^2 = 2/3\) is reasonable and gives correct distances despite relatively high fluorescence anisotropies of the donor and/or acceptor emissions (39, 109, 90). Reasons for this approximation have been discussed (48). Conformational changes based on singular positions of the donors and acceptors as well as donor-acceptor distributions can be observed in the nanosecond time range by measuring the fluorescence lifetimes of the donor or by following steady-state fluorescence of the donor or acceptor at longer times (110, 113-117). FRET is also very useful as the measurement basis of biochemical
The stoichiometry and formation of macromolecular assemblies and aggregated protein structures can often be quantitated by well-planned and correctly analyzed FRET measurements. A common problem in FRET studies can occur with the use of multi-donors and multi-acceptors. But analysis methods have been developed to take this problem into account, and it can also be an advantage when determining structures (125–127).

FRET has been widely used to determine the distribution of proteins in membranes and their interactions in two dimensional systems, as well to detect communication across membranes (128–137). The introduction of fluorescent proteins has been a great addition to the repertoire of FRET measurements of protein systems in solution and especially in fluorescence microscopy (138–141). Indeed, the number of applications of FRET in fluorescence microscopy have exploded in recent years, and novel analysis methods have also been developed in order to extract quantitative parameters from image data (142–146).

In bioluminescence resonance energy transfer (BRET), the excitation of the donor is carried out with a biological/chemical reaction instead of light excitation. It was discovered as a naturally occurring phenomenon (147). The photoprotein aequorin in Aequorea euryp挝es emits blue light when alone; however, when GFP and aequorin are associated in vivo, GFP accepts the energy from aequorin and emits green light. The use of bioluminescence as a donor in vitro was realized early (148), and it has undergone a renaissance because of the availability of the biological systems used for donor and acceptor excitation (149, 150). In general, the donor is replaced by luciferase, which becomes excited chemically in the presence of a substrate and can transfer its energy to an acceptor, which is usually (although not necessarily) a fluorescent protein (149, 150). BRET is very useful in microscopy and the background fluorescence is essentially zero. BRET has also been used in high-throughput screening (151).

References


86. Ketskemety I. Zwischenmolekulare Energieübertragung in flor-
87. Ketskemety I. Zwischenmolekulare Energieübertragung in flor-
88. Ketskemety I. Zwischenmolekulare Energieübertragung in flor-
123. Chaz E, Armb-Join D, Javin T. Proximity of lectin receptors on the cell surface measured by fluorescence energy transfer in a flow system. J. Histochem. Cytochem. 1979;27:56-64.
Mass Spectrometry, Applications in Phosphoproteomics

Francesco Giorgianni, Charles B. Stout Neuroscience Mass Spectrometry Laboratory, The University of Tennessee Health Science Center, Memphis, Tennessee

Sarka Beranova-Giorgianni, Department of Pharmaceutical Sciences, The University of Tennessee Health Science Center, Memphis, Tennessee
doi: 10.1002/9780470048672.wecb304

Proteins are biological macromolecules whose structure and functions are essential to every biological process within cells. Protein phosphorylation is one of the most important post-translational modifications and it has a profound effect on protein function. Recently, concurrent advances in bioanalytical technologies and informatics enabled studies of proteins and phosphoproteins on a global scale. These large-scale approaches represent an integral component of systems biology, which is an area of scientific inquiry that focuses on a biological system as a whole. State-of-the art mass spectrometry is a key technology for global-scale protein and phosphoprotein analyses. Phosphorylated proteins from diverse biological systems can be probed with a combination of separation methods, tandem mass spectrometry, and bioinformatics, to reveal the identity of the phosphorylated protein and the exact localization of the site(s) of phosphorylation. Characterization of the phosphoproteomes in cells, tissues and biological fluids provides an excellent foundation on which to build new knowledge of living systems.

Proteins are the final products of gene expression, and while the genome provides the “blueprint” for the molecular components of a living cell, proteins are the essential molecules responsible for cellular structure and function. Decades of studies of proteins in a one-by-one fashion have generated a wealth of knowledge on proteins as individual parts of the cellular machinery. Recently, interrogation of proteins in biological systems on a global scale gave rise to a new area of scientific inquiry, termed proteomics. Proteomics focuses on the study of the proteome, which is defined as the array of proteins that are present in a cell, organ, or biological fluid at a specific time, under a specific set of conditions. The goals of proteomics are diverse and include elucidation of basic molecular mechanisms that regulate cell function in physiological and pathological state, discovery of novel targets for the development of improved drug treatments, discovery of biomarkers for early detection of a disease and for design of tailored therapies, and many other objectives.

From the analytical standpoint, large-scale, comprehensive analysis of proteins is an extremely challenging undertaking because of the enormous complexity of proteomes and their dynamic nature. In fact, the development of proteomics as a scientific discipline was made possible through the concurrent advances in separation sciences, mass spectrometry, and informatics. Mass spectrometry has been the essential technology that enabled interrogating proteins on a global scale, with a high degree of sensitivity and accuracy. The purpose of this chapter is to describe the basic principles of mass spectrometry in the context of proteomics. Specifically, the review focuses on the use of mass spectrometry for large-scale analysis of specific subsets of proteomes – the phosphoproteomes. The chapter includes discussion of the basics of gas-phase behavior of peptides and phosphopeptides, and shows the role of mass spectrometry as a component of a general analytical strategy for phosphoproteome analysis. Because of the diversity of the analytical platforms that are being used, this review is not intended as a comprehensive description of all approaches. Rather, this article includes an overview of selected methods, a sampling of relevant references, and an example of a mass spectrometry-based phosphoproteomics methodology used in the authors’ laboratories.
Biological Background

Proteins are high molecular weight organic molecules that are essential to every biological process within living systems. Proteins are structurally and functionally diverse. Some proteins are assembled in multi-unit complexes to form the cytoskeleton of cells or other mechanical structures, while others are enzymes that catalyze biochemical reactions, or they participate in signal transduction within a cell or in cell-to-cell communication. Post-translational modifications play a key role in regulatory cellular processes, and in particular, protein phosphorylation is central to most of the signaling events that ultimately determine the biological status of all eukaryotic cells. The intracellular regulation of protein phosphorylation within cells occurs via a very complex system of positive and negative feedbacks with the surrounding environment. Protein phosphorylation regulates critical protein functions such as protein-DNA, protein-RNA, and protein-protein interactions, enzyme activity, protein trafficking, protein intracellular localization, and protein degradation. Aberrations in protein phosphorylation can have deleterious consequences and have been linked to various diseases, including cancer. It is estimated that approximately 30% of all proteins in a mammalian cell are phosphorylated at any given time (1).

A proteome represents the complete repertoire of proteins present in a cell at any given time. The term phosphoproteome refers to a specific subset of the proteome that includes all the phosphorylated protein species. Phosphoproteomics focuses on the comprehensive characterization of phosphorylated proteins in biological systems, including identification of phosphorylated proteins, assignment of their exact sites of phosphorylation, and quantification of changes in protein phosphorylation. The expansion of proteomics and phosphoproteomics in recent years has been driven by technological developments. The greatest challenge for proteomics is the inherent complexity of cellular proteomes, which is due to the dynamic nature of the proteome, the large number and wide abundance range of cellular proteins, and their diverse physicochemical properties. It is recognized that the diversity and extent of proteome complexity cannot be solved by a single technology. Instead, the trend in proteomics is to develop an array of methodologies from which a method or a set of methods can be selected to tailor the analytical strategy to suit a specific study. Chromatography and electrophoresis are the central separation technologies for proteomics. High performance mass spectrometry in combination with bioinformatics tools are key components for protein identification and characterization.

Mass Spectrometry in Phosphoproteomics

Proteins are made of twenty “standard” amino acid residues joined together through peptide bonds. Each of the twenty amino acids has unique physico-chemical properties stemming from the size of the side chain, and the possible presence of an acidic or basic ionizable group. Protein phosphorylation most commonly occurs on serine, threonine, or tyrosine residues. The task to characterize phosphorylated proteins on a proteome-wide scale includes determination of protein identities and localization of the phosphorylated amino acid residues in these proteins. Mass spectrometry is the central technology for these tasks. Although there have been major advancements in the mass spectrometry analysis of intact proteins in proteomics (2), most approaches still focus on characterization of peptides and phosphopeptides from protein/peptide digestion of proteins. Therefore, the discussion in this section will concentrate on these strategies.

Mass spectrometry has several inherent characteristics that make it an excellent choice for peptide analysis. The technique is rapid, versatile, highly amenable to automation, and it requires low-to-mid femtomole sample quantities to yield reliable information about the amino acid sequence of a peptide. (The field of mass spectrometry is continuously moving towards improved detection limits, and cutting-edge instruments provide sensitivity in the attomolar range). For phosphoproteome analysis, the general analytical strategy (Fig. 1) includes isolation of the proteins from the biological system under study; protein or peptide fractionation and enrichment of phosphorylated protein/peptides; mass spectrometry measurement of specific attributes of phosphopeptides, including their mass and fragmentation patterns; searches of protein sequence databases to identify the proteins and to assign phosphorylation sites.

Protein extraction

The first step in the analysis of proteomes and phosphoproteomes involves extraction of proteins from the biological system under study; the objective is to solubilize the proteins and to prepare them for subsequent analysis. Obviously, this step is critical for the overall success of the analysis, and choice of

---

**Figure 1** Basic components of a general analytical strategy for phosphoproteome analysis.
methods should be tailored to the characteristics of the biological system and to the goals of the study. Depending on the biological system, protein extraction may involve disruption of cells, removal of contaminants such as salts (e.g., by dialysis or ultrafiltration), and/or overabundant proteins (e.g., by immunoadfinity columns). For phosphoproteomics, particular care must be taken to preserve phosphorylation of the proteins, i.e., to prevent the action of protein phosphatases. This is achieved by controlling the temperature of the sample and by addition of phosphatase inhibitors to the extraction buffer.

**Separation and enrichment**

Complexity of the analyte mixtures is a critical issue for phosphoproteome analysis. The challenge is to probe a specific subset of molecules (phosphopeptides) among an enormous number of other peptides: diverse physicochemical properties and low abundance of many phosphoprotein/phosphopeptides further contribute to the challenge. To address this issue, multidimensional fractionations of analyte mixtures, often in conjunction with phosphoprotein/phosphopeptide enrichment, are necessary prior to the mass spectrometry analysis itself. Fractionation of the original analyte mixture produces multiple mixtures, each with reduced complexity than the starting material. The necessity to analyze all the fractions decreases the overall throughput of the analytical platform but results in an improved coverage of the phosphoproteome. The fractionation steps can be performed at the protein level or at the peptide level, using electrophoretic and/or chromatographic methods. At the protein level, the traditional approach for proteome fractionation involves two-dimensional gel electrophoresis (2D-PAGE) that combines isoelectric focusing (IEF) in the first dimension and SDS-PAGE in the second dimension. For phosphoproteomics, specific stains or immunoblotting may be used to selectively visualize phosphorylated proteins in 2D gels. One-dimensional electrophoretic separations using IEF or SDS-PAGE provide protein fractions of moderate complexity. Alternatively, separations may be performed at the peptide level after proteolytic digestion of the entire proteome. In the context of phosphoproteomics, the most widely used methodology involves strong cation exchange chromatography (3, 4).

The electrophoretic or chromatographic separation methods are often combined with enrichment strategies to specifically enrich for phosphorylated species. This enrichment can be employed at the protein level, preceding protein fractionation, and/or at the peptide level following protein fractionation and digestion. Enrichment is accomplished by various types of affinity chromatography. Affinity chromatography methods may involve immunoadfinity using antibodies specific for phosphorylated amino acid residues in proteins or peptides (5), immobilized metal ion affinity chromatography (IMAC) (6) or metal oxide affinity chromatography (MOAC) (7). Some analytical platforms include esterification to reduce non-specific binding of non-phosphorylated peptides to the IMAC column (8). Chemical derivatization methods have also been developed that replace the phosphate group with a chemically different moiety, e.g., to introduce a tag that allows subsequent capture of the modified peptides (reviewed in ref 9).

**Mass spectrometry of peptides and phosphopeptides**

The basic goal of the mass spectrometry measurement in the context of peptide analysis in proteomics and phosphoproteomics is to determine specific attributes that are then used in subsequent database searches to provide: 1. the identity of the proteins present in the sample; 2. location of the sites of phosphorylation in these proteins. Both pieces of information are derived from the mass of the peptide and, most importantly, from the gas-phase dissociation patterns that are diagnostic of the peptide’s amino acid sequence and phosphosite location. The gas-phase dissociation patterns are obtained via tandem mass spectrometry (MS/MS). On a phosphoproteome-wide scale, the analysis includes measurement of the attributes for many thousands of individual peptides.

**Basics of gas-phase behavior of peptides**

Mass spectrometry deals with measurements of gas-phase ions. Therefore, the first step in peptide analysis by mass spectrometry is the conversion of the analytes into charged species in the gas phase. Peptides are relatively large, non-volatile biomolecules. In today’s world, the two ionization methods that are used in mass spectrometry of peptides and other biological macromolecules are matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). MALDI and ESI are based on different principles (10, 11); however, both methods ionize peptides via protonation in an acidic environment. MALDI produces mainly singly-protonated peptide ions, while in ESI protonation occurs on all available basic sites in the peptide, thus yielding multiply-charged species. For tryptic peptides, which possess an amino group at the N-terminus, a basic amino acid (arginine or lysine) at the C-terminus, and possible internal basic amino acid residues, doubly- and triply-charged peptide ions usually predominate. Multiple charging in ESI has important implications for subsequent dissociations of the (activated) peptide ions. The “mobile” proton originally localized on the N-terminus migrates along the peptide backbone by internal solvation, producing a heterogeneous population of peptide ions that have the same sequence but different sites of protonation (12, 13). This heterogeneity of protonation sites drives the dissociations of peptide ions in MS/MS that are discussed below.

Mass analysis determines the mass-to-charge ratio (m/z) of ions derived from the analyte. For peptide ions, two characteristis can be obtained. The first characteristic is the molecular weight of the peptide, which can be calculated from the measured m/z of the source-generated intact peptide ion (the so-called molecular ion). The second characteristic is structural information that is obtained via an MS/MS analysis. A n M S/MS experiment measures gas-phase dissociations of an activated molecular ion of the peptide to yield product-ion data that are diagnostic of the peptide sequence. The basic sequence of events in MS/MS includes: 1. mass selection of the peptide ion of interest (that is population of ions of a single m/z) as a so-called precursor ion; 2. activation of the precursor ion, most commonly through collisions with an inert gas, followed by dissociation of the activated precursor and formation of product ions; 3. mass
analysis of the product ions and recording of the MS/MS spectrum. The process of collisional activation and dissociation is termed collision-induced dissociation (CID). Depending on the instrument type, the MS/MS events can be separated in space (tandem-in-space) or in time (tandem-in-time).

In MS/MS, protonated peptide ions in the gas phase dissociate via cleavages along the peptide backbone; fragmentations can occur at any of the three types of bonds that make up the backbone of the peptide (Fig. 2). The nomenclature for peptide dissociations distinguishes six major series of sequence-determining product ions (14, 15). The N-terminal series encompass the a-, b-, and c-ions; and x-, y-, and z-ions. In addition to the six basic series, other types of product ions may also be observed (16) under certain conditions. The relative abundance of the different product ions depends on the amino acid sequence of the peptide, on the internal energy of the dissociating precursor ion, and on additional variables that affect the CID process (17). Under low-energy CID regime, used for example in ion trap mass spectrometers, the predominant types of product ions are the b-ions and y-ions that form by cleavages of the peptide bond. As shown in Fig. 2, adjacent (singly-charged) product ions from a series have a difference in mass that determines the amino acid present at that position of the peptide. For example, if the amino acid in position 3 of the tetrapeptide in Fig. 2 is a serine (R3=CH2OH), the mass difference between the y2 and y3 product ions will be 18 Da, corresponding to the mass of the serine residue -NH-CH(C2H5)OH-. Therefore, when an MS/MS spectrum of a peptide ion contains high quality data for one complete or several partial overlapping product ion series, then the sequence of the peptide can be deduced from the MS/MS data.

**Specifics of phosphopeptides**

Phosphorylated peptides are modified peptides and therefore most of the basic concepts discussed above also apply to mass spectrometry of phosphopeptides.

In terms of ionization, the majority of large-scale phosphoproteomics strategies utilize ESI. The difficulty to effectively analyze phosphopeptides by ESI-based approaches is often attributed, among other factors, to selective suppression of phosphorylated peptides in the presence of unmodified peptides, and to decreased ionization efficiencies of phosphopeptides relative to their non-phosphorylated counterparts. However, this notion has been challenged in a recent study [18] underscoring the complexity of the phenomena associated with analyses of highly complex peptide/phosphopeptide mixtures.

Modification by phosphorylation adds 80 Da to the mass of the corresponding peptide. The principles of gas-phase dissociations of protonated phosphopeptide ions into sequence-determining product ion series are analogous to those of non-phosphorylated peptides, with an additional issue that must be taken into consideration. Under CID conditions phosphorylated peptide ions undergo a facile neutral loss of H3PO4, corresponding to the loss of 98 Da. The mechanisms that underlie this dissociation behavior have been studied for phosphoserine-, phosphothreonine-, and phosphotyrosine-containing peptides (19, 20). The loss of phosphoric acid from the molecular ion produces a non-sequence-specific product ion (M + H)−. This product ion can serve as a marker ion, indicating the presence of a phosphorylated peptide. An example of an MS/MS spectrum of a phosphorylated peptide is shown in Fig. 3. This spectrum illustrates the typical fragmentation behavior of protonated phosphopeptide ions in ion trap MS/MS. The spectrum is dominated by the intense (M + 2H-H3PO4)2+ product ion; the spectrum further contains product ions of the y- and b- series that determine the amino acid sequence of the phosphopeptide and the location of the phosphorylation site. Often, the scenario is not so favorable. The loss of phosphoric acid dominates and not enough other product ions are observed for an unequivocal sequence determination. One way to remedy this unfavorable outcome is to perform an additional dissociation, an MS/MS/MS, where the primary product ion (M + H)− H3PO4)2+ is mass-selected and then dissociated via CID.

The LC-MS/MS experiment

This section discusses a typical mass spectrometry experiment used for phosphopeptide analysis. Typically, after peptide fractionation and enrichment, the peptide digest will contain phosphopeptides in mixture with nonphosphorylated peptides. This mixture will be of a high complexity. Commonly, the mass spectrometry measurement itself is preceded by a separation of this peptide digest by reversed-phase liquid chromatography, interfaced online to MS. This LC step, typically performed in nanoflow mode and using a shallow mobile phase gradient, will separate the peptides according to their hydrophobicities. The peptides eluting from the LC are ionized by nanoESI, and peptide mass and MS/MS data are measured.
Two issues have to be addressed. First, because of the complexity of the starting mixture, even after LC separation, multiple peptides will elute and therefore at any given time there will be more than one peptide ion present. Second, the characteristics, including the mass of the precursor ion that is needed to set the precursor selection in MS/MS are not known. These issues are dealt with through data-dependent acquisition mode, in which the mass spectrometer automatically cycles through a sequence of measurements of MS and MS/MS data. For peptide and phosphopeptide analysis, the instrument measures an MS spectrum to obtain masses of the analytes eluting from LC at that particular time. Based on the information from these MS data, subsequent MS/MS events are set – for example, 5 MS/MS measurements of 5 of the most intense ions from the MS spectrum, provided they are above a specified intensity threshold. The cycle is repeated many times during the LC-MS/MS analysis. To maximize information gained in the MS/MS steps, some strategies are incorporated such as a permanent exclusion from MS/MS of known contaminants throughout the entire analysis; temporary exclusion of peptides whose MS/MS have already been measured for the duration of the expected time that it takes for the peptide to elute from the LC column. These strategies that decrease redundancy and maximize the number of peptides surveyed in the analysis are particularly important for phosphopeptides that are frequently minor components in a peptide digest. In a typical LC-MS/MS analysis, a large number of MS and MS/MS spectra are acquired, for example >10,000 MS/MS spectra on state-of-the-art ion trap instruments.

Bioinformatics

In proteomics and phosphoproteomics applications, search programs are used that utilize minimally processed MS/MS data without the need for manual interpretation (21). Development of these programs, for example SEQUEST, that allow the integration of mass spectrometry data with database searching has been one of the enabling developments in proteomics. In SEQUEST-based searches, the experimentally measured peptide mass is used to locate in the database peptide sequences whose masses match the measured mass, and then experimental product ion patterns are compared to theoretical patterns for each candidate peptide, and a correlation score is calculated. The highest scoring peptide sequences are reported. For phosphopeptide characterization, the search considers possible addition of 80 Da to serine, threonine, and tyrosine residues. After completion of the search, it is imperative that the spectra and the database search outputs are inspected before an ultimate decision about the correctness of the match is reached. For phosphopeptides, this examination includes verification that the correct amino acid sequence was retrieved, and verification of the assignment of the phosphorylation site. Once the phosphopeptides are identified and their sites are characterized, additional bioinformatics resources are available for in silico analysis and functional integration. For example, with the program ScanSite (scansite.mit.edu), sequences of the identified proteins are searched to locate motifs that would suggest phosphorylation by a specific kinase or a phospho-specific binding interaction.
Information on protein phosphorylation is compiled in several databases, for example Phosphosite (www.phosphosite.org), Phosida (www.phosida.org), and others.

Additional approaches and current developments

Mass spectrometry-based phosphoproteomics is characterized by a great diversity of bioanalytical workflows and by continuous developments of new and improved strategies. Some of these new approaches are summarized in this section.

To address some of the issues associated with CID of phosphopeptides, new strategies for ion activation/dissociation have been introduced recently in the context of phosphoproteomics. In particular, electron transfer dissociation (ETD) is emerging as a promising new strategy for MS/MS-based phosphopeptide analysis (22, 23).

It should also be noted that besides the conventional product ion scanning, specialized MS/MS functions have been adopted for phosphopeptide analysis (9). These include precursor ion scanning for monitoring the precursors of the product ion (或许是9) in the negative mode, or precursors of the phosphotyrosine-specific immonium ion at m/z 216.043 in the positive mode.

Finally, quantitative information in phosphoproteomics may be obtained through the use of stable isotope labeling. Stable isotope labeling is a proven approach for mass spectrometry-based quantification. Recent examples of stable isotope labeling methods that have been successfully adapted for large-scale quantification of protein phosphorylation include the iTRAQ methodology that involves chemical tagging at the peptide level (24), and the Stable Isotope Labeling of Amino Acids in Culture (SILAC) methodology that involves metabolic labeling (4).

Chemical Tools and Techniques

The study that serves to illustrate a possible bioanalytical strategy for characterization of protein phosphorylation in a human pituitary tissue (25) is included. The analytical methodology involves in-gel IEF for protein separation, IMAC for enrichment of phosphopeptides after digestion, LC-MS/MS and database searches for phosphopeptide identification and localization of the phosphorylated amino acid residues.

The tissue sample is obtained from surgery or autopsy, and it must be frozen immediately to prevent protein degradation. Prior to analysis, the proteins including phosphoproteins are extracted via homogenization in the Trizol reagent that isolates proteins from RNA and DNA; a phosphatase inhibitor cocktail is added to minimize dephosphorylation. Proteins are obtained in the last step of the Trizol-based extraction in the form of a pellet that is then dissolved in a buffer suitable for isoelectric focusing. This buffer typically contains chaotropes such as urea and thiourea, CHAPS detergent, ampholytes that aid in the IEF, and dihydrothreitol as a reducing agent. IEF that is performed in a commercially available immobilized pH gradient (IPG) strip separates the proteins based on their charge. After IEF, the strip is divided into sections. Each section still contains multiple proteins but the complexity of these mixtures is greatly reduced compared to that of the initial mixture. The proteins in each section of the IPG strip are digested with trypsin to produce mixtures of peptides that include phosphorylated and non-phosphorylated peptides. IMAC is used to enrich for phosphorylated peptides. The steps in the IMAC procedure involve: 1. selective binding of the phosphopeptides via interaction of their phosphate groups with the immobilized metal ion (e.g., Ga+3) under carefully controlled acidic pH conditions; 2. washing of unbound and non-specifically bound material; 3. elution of phosphopeptides from the column under alkaline conditions. Following IMAC, desalting and volume reduction of the samples is performed with a C18 minicolumn, and the samples are analyzed by LC-MS/MS. The nano-LC setup includes a combined capillary column/spray needle packed with a C18 stationary phase. The i.d. of the column is 75 µm, the i.d. of the spray tip is 15 µm, and the flow-rate is on the order of 50-150 nl/min. Mobile phases typical for reversed-phase chromatography that are compatible with mass spectrometry are used, such as water/acetonitrile/formic acid or water/methanol/formic acid (26). Peptides and phosphopeptides eluting from the nano-LC are ionized by nanoelectrospray to produce multi-protonated ions in most cases (doubly or triply charged). MS and MS/MS spectra are acquired in the data-dependent mode. MS/MS/MS may be performed in non-data-dependent mode in a separate LC-MS/MS experiment. Alternatively, this step, where MS3 is triggered when a (M + nH-H3PO4)n product ion is present in the MS/MS spectrum, may be incorporated into data-dependent scanning (3). The set of data is used to search a protein sequence database such as SWISSPROT or NCBI (nr). The search parameters include modifications for C-terminal carboxylic acid residues that are added upon phosphorylation is expected to occur. The search yields lists of phosphopeptide matches with scores indicating the quality of the match. The matches are evaluated manually. This evaluation has two objectives: confirmation of the correct amino acid sequence of the phosphopeptide which establishes the presence of the phosphorylated form of the corresponding protein in the pituitary; and assignment of the exact phosphorylation site(s) in the peptide. This validation includes inspection of the MS/MS data and the scores. Finally, the phosphorylated proteins are put into context of current scientific knowledge, using databases such as Phosphosite that extract and compile published information on protein phosphorylation.

Acknowledgments

The authors’ research activities are funded by the University of Tennessee College of Pharmacy, and by Chiesi Pharmaceuticals. Funds for mass spectrometry instrumentation have been provided in part by NIH grant 1S10 RR16679.

References


See Also
Post-Translational Modifications, Roles in Regulating Protein Function; Proteins, Chemistry and Chemical Reactivity of...
Membrane Protein Structure, Techniques to Study

Gavin King and Ann M. Dixon, University of Warwick, Coventry, United Kingdom
doi: 10.1002/9780470048672.wecb653

The biological importance of membrane proteins has been recognized worldwide for many years, but historically these proteins have proved difficult to characterize structurally because of a variety of experimental challenges. Recently, technological advances across several disciplines have prompted considerable progress in three-dimensional structure determination of membrane proteins. This review describes the state-of-the-art methods that have successfully produced high-resolution membrane protein structures to date. Most notably, X-ray crystallography will be discussed, as this technique has made by far the largest contribution to our current knowledge of membrane protein structure. This is followed by discussion of nuclear magnetic resonance spectroscopy and cryo-electron microscopy, both techniques that have also been successful in producing high-resolution structures for membrane proteins, albeit to a lesser degree than X-ray crystallography. Finally, we will discuss atomic force microscopy. Although this technique cannot be used for atomic level structural determination, it offers distinct advantages for investigation of membrane protein oligomerization, dynamics, and large-scale conformational changes. Recent notable membrane protein structures are included throughout to illustrate progress in the field as well as the strengths and weaknesses of each method.

Structural, biophysical, and biochemical studies of membrane proteins have revealed the importance of this class of proteins in fundamental biological processes such as the import and export of nutrients and waste into and out of cells, cell division, and signaling, to name a few. With approximately one third of sequenced genomes encoding integral membrane and membrane-associated proteins, and two thirds of all drugs in development targeting membrane proteins, their importance is now well accepted. However, three-dimensional (3-D) structures of the majority of known membrane proteins still elude us. Of the 47,000 protein structures deposited in the Protein Data Bank (PDB), only about 1% are for membrane proteins. Solving the structures of these proteins has proved to be highly challenging, and it still represents the leading edge of protein structural biology. The slow progress is caused, in part, by difficulties in protein production and purification (especially for eukaryotic membrane proteins) and in part by experimental difficulties encountered because of their large size and the requirement for detergent/lipid solubilization. Despite these obstacles, significant advancement has been made over the past decade. Careful production of membrane proteins and major breakthroughs in instrumentation have yielded stunning results for several classes of membrane proteins, including channels, pores, and receptors. The resulting structural data have given us an insight into the architecture of both simple and complex membrane proteins as well as their function [for an excellent summary of known structures, see http://blanco.biomol.uci.edu/MembraneProteins_xtal.html]. In this review, we will cover the four most successful and commonly used experimental methods for determination of membrane protein structures, namely X-ray crystallography, nuclear magnetic resonance spectroscopy, electron microscopy, and atomic force microscopy, with particular emphasis on recent advances. We note that molecular dynamics simulations have also contributed significantly to our current understanding of membrane protein structure, and closely accompany the above.
Membrane Protein Structures by X-Ray Crystallography

X-ray crystallography is by far the most successful method for solving structures, having provided approximately 80% of the membrane protein structures currently held in the PDB. The first membrane protein structure solved by X-ray crystallography (XRC) was that of the photoreaction center from Rhodopseudomonas viridis, originally solved at 3 Å resolution in 1985 (1). Since that time, XRC has been applied to a wide variety of membrane protein families. This technique typically produces 3-D structural data with resolution ranging from 1.5 to 3.5 Å, making it the highest resolution technique currently available for structure determination.

XRC determines the precise arrangement of atoms within a crystal by analyzing the scattering (diffraction) of X-rays by electrons to produce an electron density map. The rate-limiting step in structure determination by XRC is the production of well-ordered 3-D crystals, which has proved to be highly challenging as illustrated by the scarcity of membrane protein structures. Obtaining sufficient quantities of highly purified protein for numerous crystallization trials is often difficult, as membrane proteins can be toxic to heterologous hosts. Recent advances in automation and miniaturization have greatly reduced the quantity required, currently enabling investigators to test up to 100,000 conditions in parallel per day using less than 300 mg of protein (2). Despite these advances, however, crystallization remains a "trial-and-error" process, involving numerous variables and exacerbated for membrane proteins by the need for solubilizing agents such as detergents, which can greatly destabilize protein folds and shield crystal contacts.

The size and quality of the crystals greatly impacts the resolution of XRC (X-ray crystallography) and, therefore, is of utmost importance. The X-ray source also plays a key role in the quality of XRC data. Traditional X-ray sources can suffer from poor spatial coherence and low beam strength. For these reasons, synchrotron sources are now commonly used to collect data. A synchrotron source produces high-energy X-rays that more effectively penetrate the crystal to interact with atoms in the protein. The X-ray beam from a synchrotron source is more coherent, has a higher concentration, and can be accurately focused on very small targets.

3-D crystallization of membrane proteins

The first, and often most difficult, step in XRC is the production of 3-D crystals of purified protein. It is important to emphasize the importance of protein purity for successful crystallization because, although it is true that a nonhomogenous mixture may crystallize, it has been noted by some researchers that difficulties in crystallizing a protein may be negated by further purification. Crystallography of membrane proteins is a complicated process, and various approaches have been used, including vapor diffusion, microdialysis, batch crystallization and the recently developed lipid-phase methods. Detailed information on membrane protein crystallization can be found in several excellent reviews in the "Further Reading" section. The general principle of protein crystallization involves the supersaturation of a protein solution. The addition of precipitating agents such as salts, organic solvents, or polymers triggers the crystallization process.

Detergents play a vital role in the crystallization of membrane proteins. However, as mentioned, they can also have a destabilizing effect on the protein. For successful crystallization, the detergent micelles must be accommodated in the crystal lattice with minimum effect on the formation of crystal contacts. The choice of detergent and its associated properties (polarity of head group, aliphatic chain length, size) must be explored for each protein (3). Generally, charged detergents should be avoided because of the risk of repulsion between protein-detergent complexes. The length of aliphatic chains should be balanced between the need to cover the protein’s hydrophobic surfaces and the need to have as low a detergent volume as possible to maximize protein-protein contacts. The phase behavior of detergents and lipids is also a pertinent consideration. Phase separation from a micellar phase (detergent-rich) toward a nonmiscible phase (detergent-poor) can occur at high-detergent and precipitant concentrations. These phase boundaries have been exploited to enhance crystallization of membrane proteins (4). The use of additives such as small amphiphiles (e.g., heptane-triol, LDAO) can also enhance crystallization by reducing the volume of detergent in the protein-detergent complex, leading to increased numbers of crystal contacts (5). However, the overriding problem with the use of detergents is the increase in the number of variables that need to be optimized in crystallization trials, which is already considerable. Coupled with the use of additives, this presents a huge number of possibilities. The absence of a general set of rules to guide detergent/additive choice is currently a significant problem for crystallization, and it increases the timescale for each structure.

A significant landmark in membrane protein crystallization has been the development of lipidic cubic phase technology (6). This technically challenging approach is capable of producing crystals that diffract beyond 2 Å while providing an environment resembling that of a natural membrane. Lipidic cubic phases are gel-like materials into which the protein is embedded. Crystallization is then initiated by the addition of precipitants. If successful, the resulting crystals contain ordered layers of protein-lipid sheets, with contacts formed within and between the sheets. Lipidic cubic phases have been successfully used to crystallize several membrane proteins, including archeal seven transmembrane proteins (6, 7). Co-crystallization with antibody fragments (Fv fragments of Fab domains) has also been used to enhance membrane protein crystallization (8, 9). The fragments form bridges between protein-detergent complexes, thus increasing the number of protein-protein contacts. Using this method, crystallization was achieved for the cytochrome bc1 complex (8), the KcsA potassium channel (10), and more recently the human β3 adrenergic receptor (Fig. 3A) (11-13). A caveat to this approach is that antibody production must be
Membrane Protein Structure, Techniques to Study

The process of crystallization is still largely a process of trial and error, but in an attempt to formulate general rules, the cumulative experiences of X-ray crystallographers are being compiled into online databases such as the Marseille Protein Crystallography Database (14) and the Biological Macromolecule Crystallography Database (15). These and similar databases will provide useful starting points in crystallization trials and may greatly accelerate the process.

Collection of X-ray diffraction data

A detailed examination of the crystallographic method is beyond the scope of this article. Sources of more detailed information can be found in the “Further Reading” section. The general method of XRC involves five steps: 1) crystallization of purified protein (discussed above); 2) measurement of crystal diffraction; 3) phase determination; 4) phase and electron density calculations; 5) model building. Once crystals are formed, they are exposed to X-rays and diffraction data are collected. Exposure to X-rays can damage the crystals, and this can be reduced by freezing the crystal under a nitrogen stream at 100 K. This approach is not always favorable for membrane proteins as crystal contacts can be quite weak and, therefore, highly sensitive to temperature. Cooling can further destabilize the crystal by causing the detergents to undergo phase transitions that perturb crystal contacts. For these reasons, cooling may be limited to 4°C, which often proves sufficient to collect diffraction data. Smaller crystals are more amenable to freezing but often yield poor-quality data, being best suited for use with synchrotron beams (16). After a collection of diffraction data, phase determination is commonly carried out for soluble proteins using the method of molecular replacement. This method is less useful for membrane proteins because of the lack of known structures. A related method for phase determination is the “heavy-atom” method, which involves soaking crystals in heavy-atom solutions. Unfortunately, this method is also not as effective for membrane proteins because of the reduced binding of heavy atoms, which is a direct result of the decrease in hydrophilic surfaces, lack of accessibility, and nonspecific interactions of hydrophobic heavy-atom compounds with detergents. An alternative approach to the phase problem involves multi-wavelength anomalous diffraction phasing using selenomethionine (17). Diffraction and phasing data are used to create an electron-density map and to build a preliminary model. The model is further refined by multiple rounds of energy minimization and manual manipulations. Finally, simulated annealing is performed to produce the final structure of the protein.

Recent notable structures

Human β-2 adrenergic receptor

In 2007, this structure was solved to a resolution of 2.4 Å (11) (Fig. 3a), and it represents the first structure of a ligand-activated G-protein-coupled receptor. Crystallization was achieved using the antibody co crystallization approach. This work revealed structural features not seen before in this protein family, including cholesterol-mediated intermolecular associations and an extracellular loop containing a helix.

Sodium-potassium pump

The structure of the sodium-potassium pump was solved in 2007 to a resolution of 3.5 Å (Fig. 3b), in complex with two bound rubidium ions (13). Structures were also solved for other members of the P-type ATPase family (18, 19) around the same time revealing the structural similarity between family members as well as the conformational changes that occur during function.

Acid-sensing ion channel

This high-resolution structure (1.9 Å) solved in 2007 was the first structure of an acid-sensing ion channel (19). It belongs to the degenerin/epithelial sodium channel family of ion channels, which play essential roles in diverse biological processes such as mechanotransduction and ion homeostasis. The structure (Fig. 3c) revealed the trimeric state of the protein channel as well as a possible mechanism for activation, involving long-range conformational changes triggered by proton binding.

Membrane Protein Nuclear Magnetic Resonance Spectroscopy

Solution-state nuclear magnetic resonance (NMR) spectroscopy has also proved very useful in membrane protein structure determination, particularly for smaller proteins. Like XRC, solution NMR requires the use of high concentrations of protein as well as solubilizing agents like detergents, and therefore, it requires a similar level of optimization. However, when optimal conditions are achieved, NMR provides high-resolution structural data and has the added advantage of providing information on function, dynamics, topology, and interactions of the protein with the membrane, lipids, and other proteins. NMR determines the arrangement of atoms in a protein by exploiting the quantum magnetic property that certain atomic nuclei will align their magnetic moment with an external magnetic field. Perturbation of B₀ by a second perpendicular magnetic field leads to a response of the atomic nuclei that is detected as the NMR signal. NMR signals provide information on both short- and long-range interatomic distances that are used with simulated annealing methods to determine the 3-D structure of a protein. The lifetime of the NMR signal is related to the size of the protein, with signals for large, slowly tumbling proteins relax very rapidly. This relaxation leads to severe broadening of the signals, and it places a practical limit on the size of proteins that can be analyzed. Thus far, solution-state NMR has yielded the global fold of an 82-kDa soluble protein (20); however, for membrane proteins, the practical limit is around 40 kDa, making it most useful for proteins or individual domains. NMR studies of several membrane proteins have been carried out in the solid-state (in lipid bilayers), and this will be discussed briefly in the next section. However, a larger number of membrane proteins has been studied in solution, and this will be the main focus here.
Membrane Protein Structure, Techniques to Study

Figure 1  Ribbon representations of three membrane protein structures determined using X-ray crystallography. (a) The human β2 adrenergic receptor was crystallized using the technique of antibody co-crystallization (11) (PDB ID: 2R4R). The structures of the receptor (i) and the antibodies (ii) are shown. (b) The structure of the sodium-potassium pump (12) (PDB ID: 3B8E) is shown, indicating the three subunits of the protein. (c) The structure of the acid-sensing ion channel 1 at low pH (13) (PDB ID: 2QTS), which is composed of three chains (i, ii, and iii) that assemble to form a trimer.

Solid-state NMR

Solid-state NMR will only be considered in brief because of the limited number of novel membrane protein structures produced. For comprehensive reviews of this technique, the reader is referred to the “Further Reading” section. Around 17 structures (mainly small peptides) solved by solid-state NMR have been deposited in the PDB, with the first being gramicidin A in 1997 (21) and the most recent being the backbone structure of Influenza A M2 proton channel transmembrane domain (22). Solid-state NMR differs from the solution-state in that the molecule under investigation is static or slowly tumbling, resulting in broad NMR peaks with low resolution and sensitivity. This fundamental problem has been tackled by development of MAS (magic-angle spinning) (23) and REDOR (rotational echo double resonance) (24) experiments. Advantageously, solid-state NMR methods enable membrane proteins to be studied in lipid bilayers resembling native membranes; a disadvantage is that proton detection is challenging, with most studies focusing on isotopically labeled samples.

Membrane protein samples for solution-state NMR

Three-dimensional structure determination by NMR requires complete assignment of spectral peaks (or resonances), and the success of this process is highly dependent on sample preparation. Similarly to XRC, obtaining purified protein is a rate-limiting step in NMR structure determination. Resolution and sensitivity can be increased through isotopic labeling of proteins with 15N and 13C, either by expression in labeled media or by incorporation of labels during synthesis. Once a labeled protein has been obtained, solution conditions (e.g., temperature, pH, and ionic strength) must be optimized for NMR. For membrane proteins, this also involves selection of a solubilizing agent, which can solubilize the protein at a sufficiently high concentration for NMR (typically ~1 mM), produce high-resolution spectra, and maintain the protein in its native conformation. Detailed protocols for protein and sample preparation can be found in the “Further Reading” section.

NMR data collection and structure determination

Once a suitable membrane protein sample is prepared, spectral assignments are made. A assignment is another rate-limiting step in NMR structure determination, and it has greatly benefitted
Membrane Protein Structure: Techniques to Study

Figure 2  Membrane mimetics used in NMR studies of membrane proteins. (a) Small, spherical detergent micelles composed of amphiphilic detergent monomers. Dodecylphosphocholine (DPC) micelles were employed to determine the structure of OmpA (25) (PDB ID: 1G90) (b) Bicelles, composed of lipid/detergent mixtures, have been used recently to solve the structure of the Bnip3 transmembrane domain dimer (26) (PDB ID: 2J5D) (c) Synthetic lipid bilayers closely approximate the natural environment of membrane proteins and have been used with solid-state NMR to determine the structure of the fd bacteriophage ph2 coat protein (27) (PDB ID: 1MZT). (d) Amphipols are amphiphilic polymers consisting of a hydrophilic backbone onto which numerous hydrophobic chains are attached that can coat the hydrophobic region of membrane proteins. (e) Nanodisks are self-assembling structures composed of a phospholipid bilayer disk encircled by an engineered membrane scaffold protein. (f) Lipid cubic phases are primarily used in X-ray crystallography studies but recently have shown promise for use in NMR studies as well. (g) NMR structure of the phospholamban pentamer (PDB ID: 1ZLL), looking down on the central cavity. (h) Side view of the phospholamban pentamer shown in (g) overlaid with a refined version of the structure published in 2006 (PDB ID: 2HYN).

From the development of the TROSY (transverse relaxation optimized spectroscopy) triple resonance experiments, which improve spectral resolution and sensitivity (36). Although TROSY experiments will not be covered in detail here, briefly these experiments reduce relaxation of NMR signals at high magnetic field strengths, thus increasing the molecular mass of proteins that can be studied to ~50 kDa. Additional benefits can be realized by using TROSY in conjunction with a selectively deuterated protein and deuterium decoupling, acting to further reduce broadening of signals (37). TROSY experiments have been used successfully to assign the resonances of both β-barrel (25, 38) and α-helical membrane proteins (39).

After completion of the spectral assignment, NMR-derived short- and long-range distances can be determined by measurement of nuclear Overhauser enhancements (or NOEs). An NOE is an interaction between a pair of atoms ≤5.0 Å apart, and the intensity of the NOE can be related to the distance (r) separating the pair. The use of NOEs can be problematic for helical proteins as they tend to have few long-range NOEs. The use of methyl protonation can increase the number of NOEs; however, because of the poor chemical shift dispersion of methyl groups in helical membrane proteins, this approach is generally ineffective. An alternative is the use of residual dipolar couplings (RDCs) to obtain long-range distances. RDCs are derived from the difference in coupling constants in an aligned and unaligned state, providing information on the orientation of internuclear vectors relative to the external magnetic field (40). Several methods have been developed to produce weak alignment of membrane proteins, thus enabling the measurement of RDCs. Polyacrylamide gels can be used for small-to-medium proteins to produce an anisotropic environment where protein orientations are limited, resulting in weak alignment (40).
Membrane Protein Structures, Techniques to Study

DNA nanotubes are also a promising alignment method with the added advantage that they are resistant to detergent and do not reduce achievable protein concentrations (41). Alignment can also be achieved by incorporation of paramagnetic lanthanide ions into diamagnetic proteins (42). Once aligned samples are prepared, RDCs are accurately measured using a set of TROSY-based experiments (43). Importantly for helical proteins, a plot of backbone amide RDCs versus residue number produces a wave pattern with a periodicity of 3.6 residues per cycle for helical regions, allowing detection of helical secondary structure and the tilt of the helix relative to the magnetic field (44). The use of RDCs for structure refinement has been demonstrated for several membrane proteins, including OmpA (45), Vpu (46), and MreF (47).

Finally, NMR-derived distance information as well as information about dihedral angles (obtained from chemical shifts) is incorporated into structure calculations performed using molecular dynamics and simulated annealing programs such as CNS (48) and XPLOR-NIH (49) to calculate the protein structure.

Recent notable structures

Mistic

Despite its uncertain status as a membrane protein, the structure of Mistic from Bacillus subtilis determined in detergent micelles highlights a promising approach to solving the 3-D structure of multispanning helical membrane proteins by solution-state NMR (50). Backbone and side-chain assignment was achieved by partial deuteration and full 13C/15N labeling of the protein. The conformation of its four u-helices was determined using 13C chemical shift-derived angle restraints, NOEs, and hydrogen bond restraints. Importantly, 497 long-range restraints from 5 paramagnetic spin labels and 29 long-range NOEs allowed the fold of the protein to be determined.

Outer membrane protein a (OmpA)

In 2001, the structure of the transmembrane domain of OmpA (Fig. 2a) was solved in DPC micelles, and it represented the largest membrane protein structure (19 kDa) to have been solved by NMR at that time (25). TROSY triple-resonance experiments enabled a large number of the residues to be assigned for the deuterated protein. Relatively limited structure restraints (NOEs, dihedral angles, interstrand hydrogen bonds) revealed an eight-stranded β barrel structure for the transmembrane domain of this ion channel that was very similar to the existing X-ray structure.

Phospholamban pentamer

Phospholamban is a homopentameric membrane protein involved in muscle contraction through regulation of the calcium pump in cardiac muscle cells. The structure of the unphosphorylated protein solved in DPC micelles reveals a symmetric pentamer of phospholamban monomers (Fig. 2b) stabilized by leucine/isoleucine zipper motifs along the transmembrane domains (51). Notably, another structure was produced for phospholamban (Fig. 2h) that used a variant of the traditional simulated annealing and molecular dynamics protocol that reduced the chances of entrapment in local minima (52).

Membrane Protein Structures by Electron Microscopy

Electron microscopy (EM) has also made major contributions to the field of membrane protein structure determination, and could potentially lead the field before long. This technique provides major advantages in that the structural features of membrane proteins can be measured in their native membranes or reconstituted lipid bilayers, and not in detergents as required by (and which often limits the success of) XRC and solution NMR techniques. Although electron microscopy has historically produced lower resolution structures than XRC and NMR, recent improvements in instrumentation and sample preparation have produced data with resolution that rivals even the best X-ray structures (53).

The first reported 3-D model of a membrane protein was obtained using EM. This ground-breaking work was reported by Henderson and Unwin in 1975, who analyzed two-dimensional (2-D) crystals of bacteriorhodopsin within the purple membrane to produce a density map with 7Å resolution and a structural model of the protein (54). This data provided our first view of the architecture of a membrane protein, revealing key features such as membrane spanning u-helices. Fifteen years and many instrumental and methodological improvements later, a model of bacteriorhodopsin at atomic resolution was reported, marking the first structure solved by electron microscopy (55). Since that time, structures of important membrane protein classes, including pores (53, 56), receptors (57–59), channels (60, 61), transporters (62), and enzymes (63), have been solved using EM.

EM can provide structural information for a protein at a variety of different resolutions depending on the microscope used, the method of staining, and the condition of the sample. State-of-the-art electron microscopes can provide resolution of up to a few angstroms for good quality samples. In order to obtain high-resolution structural information from EM, one important factor is the electron source. Electrons emitted from a heated metal cathode suffer from poor spatial coherence and yield only a small amount of signal at high resolution. An alternative electron source that is rapidly gaining widespread use, the field emission gun (FEG), produces a much more coherent electron beam by limited heating in combination with an electric field. Certainly most of the current high-resolution EM structures have been acquired using an EM equipped with an FEG.

Another important factor in obtaining high-resolution data is the minimization of radiation damage to the sample. Radiation damage results from exposure of the protein sample to the electron beam as it is being imaged, and it greatly limits the resolution of the data. This loss of resolution is one of the primary reasons that XRC has had more success in obtaining structural data for proteins sensitive to radiation. One method that greatly reduces radiation damage and represents a major milestone in the field is electron cryomicroscopy (cryo-EM), in which samples are prepared in a frozen hydrated state and imaging takes place at liquid nitrogen or liquid helium temperatures under high vacuum. Cryo-EM is now the method of choice for the freezing and imaging of biological macromolecules.
Preparation of samples for cryo-EM analyses

In order to produce 3-D reconstructions of membrane protein structures effectively, cryo-EM is most successful for samples present as single particles or in 2-D crystals (as opposed to the 3-D crystals used in XRC). Single-particle cryo-EM is used primarily for large proteins and complexes that do not form crystals (64). In this approach, images are obtained for particles fixed to a carbon-film surface and coated in a layer of heavy metal salts, a process called negative staining, that acts to protect the sample from beam damage and improve image contrast, or particles suspended within a layer of vitreous ice (cryo-EM). These images are then averaged after alignment of the various particle orientations.

Although single-particle EM has been most successfully and routinely used for soluble proteins, structural information for several membrane proteins including the nandine receptor ion channel (58), the L-type Ca\(^{2+}\) channel (61), the voltage-gated K\(^{-}\) channel (60), and the inositol (1,4,5)-triphosphate receptor (65) has also been obtained using this method. However, single-particle EM has thus far resulted in low-resolution structural information, typically yielding data with resolutions in the range. A further disadvantage of single-particle cryo-EM is the lower limit for molecular weight, which is currently approximately 250 kDa. This limit results from the need to distinguish between individual particles and molecules in the surrounding medium, and it is unlikely to change in the future (66).

The highest resolution cryo-EM data has thus far been obtained for proteins present in 2-D crystals (a technique known as electron crystallography). This technique measures the structural features of membrane proteins reconstituted into 2-D crystals in the presence of lipid bilayers. In contrast to the 3-D crystals used in XRC, in which proteins are solubilized in detergent micelles that can disrupt crystal formation and reduce crystal quality, formation of 2-D crystals forces the membrane proteins to pack within a lipid bilayer, thus restoring their native environment. A further advantage of 2-D crystallization is that it requires very small amounts of protein (as opposed to NMR, which requires milligram quantities).

As with the growth of 3-D crystals, the growth of 2-D crystals is often the most difficult step in cryo-EM analysis of a membrane protein. Two-dimensional crystallization of membrane proteins can be achieved by mixing the protein with detergent and lipids, and then slowly decreasing the detergent concentration (e.g., by dialysis). Two-dimensional crystals can also be formed by vesicle fusion, addition of crystallizing agents, addition of detergents to reduce the lipid-protein ratio, or adsorption onto lipid monolayers (for a review of 2-D crystallization techniques, see the “Further Reading” section). Because the fundamental mechanisms of 2-D crystal formation are thus far very poorly understood, extensive screening of suitable lipids and lipid-protein ratios is required (67) to obtain a high-quality crystal. However, in some cases, membrane proteins that have proved very resistant to 3-D crystal formation have readily formed 2-D crystals in lipid membranes.

Two types of 2-D crystal have been analyzed by cryo-EM: sheet-like (planar) crystals and tubular crystals. Planar crystals are tilted in the electron beam in order to obtain different projection maps of the crystal at a variety of angles. Maps collected at different angles are then averaged together to produce a 3-D reconstruction (68). However, the planar crystal cannot be tilted through all angles in the microscope (for example, a sample rotated through 90° would then be parallel to the electron beam) and typically can only be rotated through 70°–75°. The resulting resolution is, therefore, spatially heterogeneous and appears to have what is known as a “missing cone” of data. Alternatively, membrane proteins can also crystallize into tubular crystals where, in a single crystal “tube” one can potentially see all orientations of a protein without the need to tilt the crystal. Tubular crystals overcome the “missing cone” of information and provide resolution that is equal in all directions (59).

Collection of cryo-EM data

High-resolution cryo-EM data can be collected in two forms: electron images (69) or as electron diffraction patterns. Cryo-EM images contain information on both amplitude and phase, which can be analyzed after Fourier transformation. The quality of the amplitude data can be improved if combined with electron diffraction data, which contains only amplitude information. In this way, EM overcomes one of the main difficulties in XRC. In XRC, only diffraction patterns are obtained. X-rays cannot be used to form an image of the crystal; therefore, the phase information is lost. In contrast, electron microscopes contain electron lenses that can capture phase information.

Structure determination and other applications of cryo-EM

A 3-D structure is determined by fitting an atomic model to a 3-D density map at a given resolution. This map is determined by analyzing EM images and electron diffraction patterns. Because the membrane protein is embedded within a lipid bilayer, 2-D crystals more closely resemble its native environment than 3-D crystals do. The protein Aquaporin 1 (AQP1) embedded in a 2-D crystal was shown to maintain its biological activity (70). Because they are more likely to adopt their native fold in a lipid bilayer, 2-D crystals are thought to provide more biologically relevant structural information about protein-protein and protein-lipid interactions. This is illustrated by the 3-D structure of the transporter protein EmrE, which has been solved by both XRC (71) and cryo-EM (62). Comparison of the two structures reveals important differences between the structure of EmrE in a 2-D crystal (formed in detergent) and its structure in a 2-D crystal (embedded in a bilayer) that result from the presence of lipids.
Membrane Protein Structure, Techniques to Study

conformational changes of a protein over time, thus yielding critical insight into function. In these time-resolved studies, conformational changes are induced in the protein (by changing pH, adding ligand, etc.) and then trapped by freezing the crystal. Freezing can take place at several time points after induction of a conformational change to produce a series of structures that describe a reaction pathway. Time-resolved studies have been performed to investigate the photocycle of bacteriorhodopsin (72), the gate-opening mechanism of the acetylcholine receptor (57), and more recently, the pH-dependent mechanism for a Na\(^+\)/H\(^+\) antiporter (73).

**Recent notable structures**

**Aquaporin-0**
The structure of this water-selective membrane pore protein (Fig. 3a and 3b) represents the highest resolution structure obtained from electron crystallography to date (53). Data were obtained for Aquaporin-0 in double-layered 2-D crystals, and its staggering 1.9-Å resolution clearly reveals water molecules within the pore. The data also reveal associated lipids, allowing key protein-lipid interactions to be modeled.

**Aquaporin-4**
This 3.6-Å structure of Aquaporin-4 (Fig. 3c) was determined by electron crystallography of double-layered 2-D crystals (56). Features in the structure show that Aquaporin-4 can form membrane junctions, and they suggest for the first time its role in cell adhesion. This structure is of additional interest in that it is the first structure of a multispanning mammalian membrane protein obtained by purely recombinant methods.

**Glutathione transferase-1**
At 3.2 Å, the resolution of this recent structure also rivals that obtained in XRC (63). This structure provides a strong insight into the function of the protein, and it reveals key differences between the glutathione binding site of this membrane-spanning enzyme and that of its soluble counterparts.

**Imaging Membrane Proteins by AFM**
The 2-D crystals of membrane proteins discussed in the above section can also be analyzed using atomic force microscopy (AFM) (74). Although it is a low-resolution technique restricted to surface contouring, and as such cannot provide atomic level structural information, AFM has developed into a powerful tool for investigation of oligomerization, dynamics, and large-scale conformational changes. AFM also provides several advantages over EM. As mentioned, high-resolution EM analyses require coating of the sample, low temperatures, and high vacuum. In contrast, AFM images (or topographs) can be obtained under physiological conditions allowing investigation of structure and conformational changes involved in function (e.g., gating of channels). AFM is also the only imaging technique that can collect data on liquid samples, and it is therefore able to provide, in addition to images of 2-D crystals, images of membrane bilayers in an aqueous environment.

AFM was first used to analyze a membrane protein in 1990, when images were collected on hydrated purple membranes (75). In these images, the global arrangement of bacteriorhodopsin molecules was observed for the first time at near-physiological conditions (75). The resolution of AFM images has improved steadily over the past 20 years, and now the technique is regularly used to observe individual proteins and macromolecular assemblies in both 2-D crystals (74–78) and in native membranes (75, 81, 82). From AFM images, information has been obtained for a wide variety of membrane proteins, including Class A G-protein-coupled receptors (81), pumps (78, 83), channels (76), enzymes (84), and pores (80). Larger protein complexes, such as the light harvesting complexes I and II, have also been studied in both 2-D crystals (79) and in their native membranes (82).

Currently, the resolution of AFM is lower than that of NMR, XRD, or cryo-EM. The highest resolution AFM data thus far has been collected on samples in 2-D crystals, producing structural data with resolutions between 9 and 14 Å (77–80). In general terms, AFM works by raster scanning (scanning line by line) a very sharp tip attached to a flexible cantilever over a sample
Membrane Protein Structure, Techniques to Study

in order to produce a contour map of its surface. In order to
maximize resolution, vertical fluctuations of the tip must be
minimized. This is achieved using a self-regulating feedback
system (or servo system), which keeps the cantilever deflection
approximately constant by making small adjustments in the
vertical displacement of the sample. A 2-D crystal provides
an ideal sample for AFM as it is hard and flat, with the protein
embedded in a densely packed array, thus restricting any lateral
or vertical movement of the protein. Another prerequisite for
high-resolution data is minimal interaction between the tip and
the sample, and this is greatly influenced by the size and the
geometry (or sharpness) of the tip. The tip geometry is most
commonly an inverted pyramid. Depending on its sharpness, as
the tip moves across a surface, structures may interact with both
the end of the tip as well as the sides of the tip. Any interactions
with the sides of the tip will cause a broadening of the signal,
thus reducing the lateral resolution of the image. This effect,
also known as tip convolution, is most noticeable on surfaces
with considerable height differences. Most of the tips used on
high-resolution instruments are now commercially fabricated,
with the best tips having a radius of curvature of >5 nm.

Structure determination and other
applications of AFM

As mentioned, individual membrane proteins and complexes
in 2-D crystals and in densely packed, noncrystalline arrays can
be imaged by AFM at sub-nanometer resolution (76–80, 84).
This makes AFM ideally suited to provide information on the
conformation and oligomeric state of membrane proteins in their
native membranes. One very nice example of this is the study of
rhodopsin in its native membrane (81). AFM was used to image
native disk membranes in aqueous solutions, and revealed the
organization of rhodopsin into dimers and higher oligomers.
By imaging membranes in aqueous solutions, AFM also allows
us to observe conformational changes of biomolecules as they
function.

Because deflections of the cantilever can be detected in the
10-50 pN range, making single-molecule force measurements
possible, AFM can also be used to study protein folding. In
these studies, the AFM tip is displaced toward the sample until a
protein is attached by contact adhesion, the tip is then retracted.
Force versus distance curves are recorded before and after
adhesion in order to determine the forces involved in unfolding
or “unzipping” the protein. In addition to the force versus
distance curves, images can also be acquired before and after
an unfolding event to observe the effects on the surrounding
environment (85).

Recent structural studies

KirBac3.1

AFM is a powerful method of providing information on larger scale structure, especially when combined with higher resolution structural data. This fact is beautifully illustrated in the investigation of the KirBac3.1 potassium channel (76). AFM imaging of KirBac3.1 embedded in a lipid bilayer has revealed the tetrameric assembly of the protein (Fig. 4a and 4c) as well as large conformational changes upon ligand binding (Fig. 4b and 4d), thus providing insight into the gating mechanism of this channel.

F0F1-ATP synthase

The F0F1-ATP synthase is responsible for synthesizing ATP in many organisms. This enzyme is composed of two rotary motors (F0 and F1) connected by a central stem, and the stoichiometry of these two motors is of critical importance to energy conversion. F0F1-ATP synthases are very large, making them unsuitable for study by NMR or X-ray crystallography. However, AFM is ideally suited to such applications and has been used to determine the oligomeric states of the F0 and F1 motors in several species of bacteria and plants to shed new light on how the enzymes function (84, 86, 87).

Summary

The aim of this review has been to summarize the key tech-
niques currently used for determination of membrane protein
structures. Thus far, XRC has produced the largest number of
membrane protein structures, and it has achieved the highest resolu-
tion of any other technique. However, this technique still re-
quires a high-quality crystal, which is currently the rate-limiting
step in the process. Steady progress in our understanding of
crystal growth and the development of new crystalization meth-
ods show great potential for future studies. NMR has proved to
be a valuable technique for the study of small membrane pro-
teins and individual domains of larger proteins. In addition to
3D structural information, NMR can also provide information
on dynamics and lipid binding. NMR also has the advan-
tage that it does not require crystals, which are difficult and
time-consuming to produce, and instead uses a large range of
membrane mimetics and solubilizing agents. The major limi-
tations of NMR remain the upper limit on protein size and
difficulties in resolving resonances for assignment purposes.

Continued developments in instrumentation, experimental meth-
ods, and improved membrane mimetics will greatly advance the
field. EM has also made a large impact on our knowledge of
membrane proteins and promises to equal or surpass the more
static structural methods of XRC and NMR. Because of recent
new developments in technology, cryo-EM has caught up with
XRC and NMR in terms of resolution and is well positioned to
lead the field in the future. Both EM and AFM offer the unique
advantage that membrane proteins can be studied in synthetic
lipid bilayers as well as their native membranes, producing more
biologically relevant structural data. Furthermore, these methods
allow study of conformational changes over time, in response to
such events as lipid binding or unfolding. However, the most
exciting prospect for future studies of membrane proteins comes
when data from all these methods are used in concert, alongside
molecular dynamics simulations, to describe the structures and
functions of these essential proteins.

References
1. Deisenhofer J, Epp O, Miki K, Huber R, Michel H. Structure of
the protein subunits in the photosynthetic reaction centre of
3. Wiener MC. The development of membrane protein crystalization
screens based upon detergent solution properties. Biophys. J.
2002;82:29a.
4. Arnold T, Linke D. Phase separation in the isolation and purifica-
5. Wiener MC. Existing and emerging roles for surfactants in the
dimensional crystallization of integral membrane proteins.
lipidic cubic phases: a case study with bacte-
8. Huntre C, Kropje T, Lange C, Rossmann T, Michel H. Structure at 2.3 Å
resolution of the cytochrome bc(1) complex from the yeast, Saccharomyces cerevisiae co-crystallized with an
istry of ion coordination and hydration revealed by a K–
channel-Fab complex at 2.0 angstrom resolution. Nature 2001;414:
43–48.
10. Doyle DA, Cabral JM, Pluttimer RA, Kuo AL, Gulbis JM, Cohen
SL, Chai BT, MacKinnon R. The structure of the potassium chan-
nel: Molecular basis of K+ conductance and selectivity. Science
11. Chenazov V, Rosenbaum DM, Hansen MA, Ramussen SG,
Thon FS, Koballa TS, Choi HJ, Kuhn P, Weiss W, Koballa
BE, Stevens RC. High-resolution crystal structure of an engi-
neered human beta2-adrenergic Gn protein-coupled receptor. Sci-
ence 2007;318:1258–1265.
12. Jardt F, Furukawa H, Gouaux E. Structure of acid-sensing ion channel 1 at 1.9 Å resolution and low pH. Nature
13. Morten J, Pedersen BP, Toudrup-Jensen MS, Sorensen TL, Pe-
tersen J, Andersen JP, Vilsen B, Nissen P. Crystal structure of the
on-line crystallization data bank for screening strategies. Acta
15. Gilliland GL, Tung M, Blaikiees DM, Lader J. Biological
macromolecule crystallization database, version 3.0: New fea-
tures, data and the NAs archive for protein crystal-growth data.
16. Dauter Z. Efficient use of synchrotron radiation for mac-
romolecular diffraction data collection. Prog. Biophys. Mol.
17. Hendrickson WA, Horton JR, Lennard DM. Selenomethionyl pro-
teins produced for analysis by multielevation anomalous diffrac-
tion (MAD) - a vehicle for direct determination of 3-dimensional
18. Olsen C, Picard M, Witten A, Gyurcs C, Mort J, Ovigs C,
Møller J, Nissen P. The structural basis of calcium transport by the
P. Crystal structure of the plasma membrane protein pump. Nature
20. Tugarinov V, Choy WY, Orndhoff VF, Kay LE. Solution
NMR derived global fold of a monomeric E2-A20 enzyme. Proc.


Membrane Protein Structure, Techniques to Study


69. Baldwin JM, Henderson R, Unwin PN. Three-dimensional structure of the type 1 inositol 1,4,5-trisphosphate 


86. Popoyev V, Yu, J, Meier T, Vorick J, Dimmock P, Muller DJ. The c15 ring of the Spirillum platensis F-ATP synthase. FBPase symme-


Further Reading


Engel A. Robert Fernig Lecture: Micriscopic assessment of mem-


Jayabal H, Kedrov A, Cipriani DA, Sapra KT, Struckmeier J, Muller DJ. Imaging and detecting molecular interactions of single transmem-


Mossar G. Two-dimensional crystallization of transmembrane pro-


Prive GG. Detergents for the stabilization and crystallization of mem-


See Also

- Protein Expression, Systems for
- Solubilize Membrane Proteins, Techniques to
- Crystallization of Proteins, Overview of Applications in Chemical Biology
- NMR for Proteins
- Imaging Techniques for Proteins
NMR for Proteins

Angela M. Gronenborn, Department of Structural Biology, University of Pittsburgh Medical School, Pittsburgh, Pennsylvania
doi: 10.1002/9780470048672.wecb394

Solution NMR methods for determining three-dimensional, atomic resolution structures and dynamic properties of biologic macromolecules are well established. Currently, an ever-increasing number of experimentally solved protein NMR structures is available and contributes to our understanding of biology. The development of novel NMR approaches, labeling techniques, and calculational algorithms allows the study of larger and more complex systems. Here I summarize the basic NMR approaches, discuss labeling and assignment strategies, and provide examples for their uses.

Introduction

Understanding protein function relies on studying protein structures at the atomic level. Among the various techniques that can be applied to investigate proteins, only two have provided many high-resolution structures: single-crystal X-ray diffraction and nuclear magnetic resonance (NMR) spectroscopy. Currently, the Protein Data Bank (PDB) contains 36,329 crystal and 5757 NMR protein structures, with 1685 and 134 complexes determined by X-ray or NMR, respectively. Although X-ray crystallography has contributed protein structures since the 1950s, with NMR only reaching this stage in the 1980s, the number and size of macromolecular structures that have been investigated by NMR has increased dramatically over the past 20 years (1). Thus, NMR now is firmly established as a methodology that has moved beyond its initial applications in analytical chemistry into the structural biology field.

In the early years, the development of a wide range of two-dimensional (2-D) NMR experiments (using 1H as the active nuclei) culminated in the determination of three-dimensional (3-D) structures of several small proteins (2, 3). Under favorable circumstances, 2-D NMR techniques can be applied successfully to the structure determination of proteins up to ∼100 residues (4, 5). Beyond 100 residues (∼10 kDa), however, 2-D methods tend to fail, principally because of the spectral complexity that cannot be resolved in two dimensions. In the late 1980s and early 1990s, a series of major advances took place, and the introduction of three- and four-dimensional (4-D) experiments that use 13C and 15N nuclei significantly extended the attainable spectral resolution (see 6, 7, for reviews). The combination of multidimensional with heteronuclear methodologies also alleviated problems associated with large linewidths, mainly because they exploit heteronuclear couplings that are large compared with homonuclear proton couplings. (Note that with increasing molecular mass, a molecule tumbles slower in solution, resulting in increased linewidth; consequently, the efficiency of magnetization transfer through bonds decreases dramatically, rendering 1H through-bond correlation experiments unsuitable for large systems.) Following these spectroscopic advances, further progress occurred with the introduction of residual-dipolar-coupling (RDC)-based methodologies, transverse relaxation-optimized spectroscopy (TROSY) experiments, and specialized isotopic labeling schemes. Exploiting all the recent developments, it is possible now to determine the structures of proteins in the 50 kDa range at a resolution comparable with ∼2 Å resolution crystal structures. The upper application limit for the current NMR methodology probably is around 50–60 kDa. The largest single-chain proteins solved to date to high resolution comprise 240–260 residues with molecular masses of 25–30 kDa (8–14). The largest overall protein structures solved to date are multimers (15–17) or protein complexes (14, 18, 19). In addition, beyond the elucidation of three-dimensional structures, investigations of even large systems are possible (20–23). These studies include the identification of secondary structures, the mapping of intermolecular binding sites, and the study of conformational changes. Determinations of the relative orientation of domains or proteins within multiprotein complexes based on RDCs are feasible, and dynamical characterizations of very large biomolecular complexes can be carried out.

In addition to methodological advances in protein NMR that make structure determination of small systems routine and of large systems possible, the availability of commercial high-field spectrometers equipped with cryogenic probes has played a major role as well. Currently, spectrometers used for biomolecular NMR employ magnets that range in field strength from 14.1 T (600 MHz) to 22.3 T (950 MHz). The improved resolution and increased sensitivity afforded by these state-of-the-art instruments allows the detection of small signals, in particular NOEc: permits measurements on low solubility proteins; and aids in the investigation of large molecular weight complexes.
General Strategy

The power of NMR over other spectroscopic techniques results from the fact that every NMR active nucleus gives rise to an individual resonance in the spectrum that can be resolved by using multidimensional (i.e., 2-D, 3-D, or 4-D) techniques.

To measure distances between hydrogen atoms via NOEs, the corresponding resonances have to be identified unambiguously. Thus, each resonance has to be assigned to a unique atom in a specific monomer unit (amino acid for proteins) in the macro-molecule. This assignment is achieved via a combination of through-bond and through-space correlations. Subsequent identification of NOE through-space interactions yield the principal experimental parameters—distance constraints—that are used for NMR structure calculations.

The principles of structure determination by NMR thus can be summarized as follows: 1) sequential resonance assignment, 2) torsion angle determination and stereospecific assignment at chiral centers using three-bond scalar couplings combined, where appropriate, with intraresidue and sequential interresidue NOE data, 3) identification of through-space NOE connectivities between protons separated by less than 5 Å, 4) determination of orientational constraints extracted from ratios of heteronuclear relaxation (T_r/T_s) or residual dipolar couplings (RDCs), and 5) calculation of 3-D structures on the basis of all the above experimental NMR restraints by using one or more of several computational algorithms (24).

Sequential Resonance Assignment

Conventional sequential resonance assignment has been applied to proteins up to about 100 residues, albeit with considerable effort. It relies on 2-D homonuclear 1H–1H through-bond correlation experiments to identify amino acid spin systems. These spin systems are linked via 2-D 1H–1H NOE experiments that identify through-space (5 Å) sequential connectivities of the type Cα(i)–HN(i) + 1, Cα(i)–HN(i) + 2, and Cα(i)–HN(i) + 3 along the backbone (25). Cα and Cβ represent main chain and side chain carbons, respectively; i represents the reference amino acid residue, with i + 1 indicating the next residue in the chain. As stated earlier, for larger proteins the spectral complexity is such that 2-D experiments no longer suffice, and it is essential to increase the spectral resolution by increasing the dimensionality of the spectra (26).

In some cases, it is still possible to apply the same sequential assignment strategy based solely on through-bond correlations that involve well-defined heteronuclear one-bond (1JCα, 1JNα, 1JNCO, 1JCαCO, and 1JCC) and two-bond (1JNC, 1JCC) scalar couplings along the polypeptide chain (17, 28). Figure 1 illustrates the magnetization transfer paths in four types of experiments. Pulsed-field gradient methodologies (29, 30) reduced the required measuring times and it became possible to acquire only a few scans per increment without introducing artifacts. This allows each 3-D experiment to be recorded in as little as a few hours. In most cases, however, signal-to-noise requirements necessitate 24 hours of measuring time. For proteins larger than ~25 kDa, the assignment of the backbone and side chain carbons is facilitated by using a sample in which the non-exchangeable (carbon-attached) protons are deuterated (see also below). Perdeuteration results in narrower lines caused by a substantial increase in the relaxation times of carbon and proton spins in the proximity of the substituted deuterons. The approximately six-fold lower gyromagnetic ratio of deuterons relative to protons accounts for this phenomenon. Thus, for example, in the case of a 30 kDa protein, the average transverse relaxation time (T2) for the backbone amides is increased from ~13 ms in the protonated sample to ~28 ms in the perdeuterated sample (8).

In addition to perdeuteration, it is possible to narrow the linewidths significantly via the so-called transverse relaxation-optimized spectroscopy (TROSY). In these experiments, transverse relaxation is suppressed by the use of destructive interference between dipole–dipole coupling and chemical shift anisotropy (31). The TROSY sequence can be incorporated into numerous multi-dimensional experiments, and its effectiveness increases with increasing field strength up to an optimal value. For example, for a 150 kDa protein at a spectrometer frequency of 750 MHz, the residual linewidths of the 13C and NH resonances in a TROSY experiment will be ~10 Hz and ~45 Hz.
respectively, in a protonated sample, and -5 Hz and 15 Hz, respec-
tively, in a perdeuterated sample. TROSY is not limited to 1H, 1H-correlation experiments only; it also can be applied to aromatic 13C, 1H spin systems (32), methyl groups (either 13CH3 or 13CH2D), and methylene (13CH2) groups (33). The successful applications to very large protein complexes with molecular masses of up to 800 kDa (33, 68) render 13C, 1H-correlation experiments as important alternatives to the 15N, 1H ones. In particular, such carbon ex-
periments are ideally suited to study protein dynamics because methyl groups are appealing probes that are distributed uni-
formly throughout the protein structure (34). TROSY, in conjunction with appropriate levels of perdeuter-
amination, removes one major impediment to extending the NMR method to larger proteins, namely broad lines that arise from slow molecular tumbling. However, it still is essential to assign all the side chain resonances, a task that becomes problematic for perdeuterated systems where protons have been replaced by deuterons and are therefore of low abundance, although as-
signments of the side chains carbons are easy. Through-space NOE interactions between protons provide the main source of geometric information and protons are necessary to calculate structures. Through-bond HCHC-correlated and total correlated spectr-
troscopy experiments can be employed for proteins up to 25 kDa to transfer magnetization from protons to carbon through 
neighboring carbons and back to protons. For proteins larger than 25 kDa, however, the sensitivity of these two experiments is markedly reduced as the 13C linewidths approach the value of the 1JCC coupling (1-30-35 Hz). An alternative, power-
ful strategy is to rely on through-space correlations between adjacent carbon atoms (i.e., 13C-13C NOE) (35). In contrast to through-bond correlations, the efficiency of transfer of the 13CC, 13CC NOE increases with both increasing molecular weight and increasing field strength. This increase has been exploited in a suite of 13C heteronuclear direct detection experiments that are suitable for sequential protein backbone assignment (36).

Torsion Angles and Side Chain Rotamers

Coupling constant data allow the extraction of torsion angle restraints given their simple geometric, Karplus-type (37) rela-
tionships. For small molecules, the coupling constants can be measured directly from the in-phase or antiphase splitting of a particular resonance in the 1-D or 2-D spectrum. For large sys-
tems for which the individual linewidths exceed the coupling, it becomes difficult to extract accurate couplings in this manner. As an alternative, exclusive correlated spectroscopy (ECOSY) ex-
periments are employed to generate reduced cross-peak mul-
tiplets (38). Although this permits one to measure accurate couplings, the sensitivity of ECOSY experiments, unfortunately, is quite low. In addition, the usefulness of such experiments is curtailed by the fact that the couplings in multi-dimensional experiments have to be measured in the indirectly detected fre-
quency dimensions and hence are influenced by any limitation in digital resolution. To overcome these shortcomings, a series of highly sensitive quantitative J correlation experiments have been developed (39). In some of these experiments, the coupling is obtained from the ratio of cross-peak to diagonal-peak inten-
sities, and in others, it is obtained from the ratio of the cross 
peaks obtained in two separate experiments (with the coupling active and inactive) that are recorded in an interleaved manner. The most precise coupling values are obtained from experi-
ments that allow their extraction from the difference of two 
large multiple quantum splittings (40, 41). Particularly useful couplings are 3JC, 3JHC, 3JHH, and 3JCH, which are related to the back-
bone torsion angle φ, the 3JC, 3JCH, 3JHH, 3JHH, and 3JCH, couplings, which report on the χ1 side chain torsion angle, and the 3JCH, 3JHH, couplings, which are linked to the χ2 and χ3 side chain torsion angles of leucine, isoleucine, and methionine. For smaller proteins, it is often possible to obtain stereospecific assignments of χ-methylene protons on the basis of a qualita-
tive interpretation of the homonuclear 3JCH coupling constants 
and the intraresidue NOE data that involve the NH, CαH, and 
CαH protons (35). A more rigorous approach, which also per-
mits one to obtain χ, φ, backbone, and χ1 side chain torsion angle restraints, involves the application of a conformational grid search in φ, φ, χ1 space on the basis of the homonu-
clear 3JHH, and 3JCH coupling constants (which are related to χ1 and χ3, respectively) and the intra- and sequential intraresidue NOEs that involve the NH, CαH, and CαH protons (42, 43). This information can be supplemented and often supplant ed by measuring heteronuclear couplings using quantitative J correla-
tion spectroscopy. For larger proteins, the most useful couplings in this regard are the 3JC, 3JHH, and 3JCH, couplings that involve the aromatic, methyl, associated and methylene C atoms. These couplings suffice to derive the associated χ1 side chain rotamer. An alternative method that uses cross-correlated relaxation al-
lo ws one to measure angles between bond vectors directly and 
thus does not rely on coupling constants. It defines the back-
bone torsion angle φ, either by measuring the dipole-dipole cross-correlated relaxation of double-quantum and zero-
quantum coherences that involve the intraresidue dipolar 
field of the 1H-N(1) and 13C-H(1) bond vectors (44) or by measuring the cross-correlated relaxation between 
13C-H(2) dipolar interaction and 13C(carbonyl) chemical shift 
anisotropy (45).

NOE Assignment in Proteins

For the purposes of spectral assignment, the complement of 3-D heteronuclear experiments outlined above is sufficient in most cases. However, because of spectral overlap, greater increases in resolution are required for reliable identification of NOE through-space interactions, even for medium-size proteins. This increase is achieved by extending the dimensionality to four di-

dmensions (6). In this manner, each 1H-1H NOE interaction is specifed by four chemical shifts, the two proton frequencies of the atoms that give rise to the NOE and the resonance frequen-
cies of the heavy atoms to which they are attached (Fig. 2). The number of NOE interactions present in each 2-D plane of a 4-D
NMR for Proteins

Figure 2  Schematic illustration of 2-, 3-, and 4-D heteronuclear-edited NOE spectroscopy.

\[ 1^{13}C/^{15}N, \quad 1^{13}C/^{13}C, \quad \text{or} \quad 1^{15}N/^{15}N-\text{separated NOE spectrum is relatively small compared with a homonuclear proton 2-D NOESY spectrum. This yields an inherently high-resolution 4-D spectrum, despite low digitization. Indeed, it has been born out after more than a decade that both good sensitivity and resolution for proteins up to 400 residues can be achieved using this methodology. Once complete 1H, 1^{15}N, and 1^{13}C assignments are obtained, analysis of the 4-D 1^{15}N/^{13}C, 1^{13}C/^{13}C, and 1^{15}N/^{15}N-separated NOE spectra permits the assignment of numerous NOE interactions in a relatively straightforward manner. The first successful application of these methods to the structure determination of a protein greater than 15 kDa occurred in the early 1990s with the determination of the solution structure of interleukin-1, a protein of 18 kDa and 153 residues (46). Equivalent methodology now is used routinely for medium-size proteins.

For proteins larger than 30 kDa, it is not always possible to obtain complete side chain proton assignments, either because of spectral overlap or the lack of through-bond connectivities caused by large linewidths. In such cases, selective or defined deuteration (see below) is extremely helpful. One of the simplest approaches is to use a sample of a perdeuterated protein in which only the amides are protonated by back-exchange with water. Recording a 4-D 1^{15}N/1^{15}N-separated NOE spectrum (47) allows one to obtain NOEs with exceptional sensitivity because many spin-diffusion pathways are removed and long mixing times can be employed. In addition, labeling strategies in which the amides and methyl groups are protonated, while all other groups are deuterated, have become very popular and can aid in structure determination (48, 49). Such labeling regimes retain the advantage of deuteration in triple resonance assignment experiments while ensuring the presence of a reasonable number of protons in strategic locations for the purpose of NOE detection and structure determination. Indeed, model calculations have shown that it is possible to obtain low-resolution global folds on the basis of NH–NH, NH–methyl, and methyl–methyl distances alone (50). A striking example of a successful fold determination obtained with the tools described above is that of the 723-residue protein Malate Synthase G (51).

Protein Complexes

The concept of heteronuclear, multidimensional spectroscopy outlined above can be extended and modified for applications to protein complexes. To that end, complexes are prepared so that one component is labeled with 15N and 1^{13}C and the other one is unlabeled (i.e., one partner remains at natural isotopic abundance). Provided that one constituent in the complex (e.g., a peptide, an oligonucleotide, a drug, or other ligand) presents a relatively simple spectrum that can be assigned by 2-D methods, the most convenient strategy for dealing with such protein complexes involves using a 1^{15}N- and/or 1^{13}C-labeled protein and an unlabeled ligand. It then is possible to use a combination of heteronuclear filtering and editing to design experiments in which correlations that involve only protein resonances, only ligand resonances, or only through-space interactions between ligand and protein are observed (52). These experiments have been employed successfully in several laboratories and for a wide range of systems, including protein–drug, protein–peptide, and protein–nucleic acid complexes.

For protein complexes or multimeric proteins, additional, specifically tailored labeling schemes can be used to facilitate the identification of intermolecular NOEs. Note that multimeric proteins can be viewed as protein complexes between identical subunits. For such complexes, 1:1 mixtures of 1^{13}C/1^{15}N/1^{1}H, 1^{15}N/1^{13}C/1^{1}H, 1^{13}C/1^{15}N/1^{1}H, and 1^{15}N/1^{13}C/1^{1}H-labeled proteins enable one to record high sensitivity 2-D, 3-D, and 4-D experiments to observe specifically NOEs.
between 1) protons attached to $^{13}$C or $^{15}$N and protons attached to $^{12}$C or $^{14}$N, 2) $^{13}$C-attached and $^{15}$N-attached protons, and 3) $^{15}$N-attached protons and protons attached to $^{12}$C or $^{14}$N. In this manner, correlations that are specific for intermolecular contacts can be identified unambiguously from either side of the interface by employing two complex samples with alternating labeled partners. An example of strips from a 3-D $^{13}$C-edited, $^{12}$C/14N-filtered NOE spectrum and a depiction of the responsible contacts in the final structure of the protein–protein complex are illustrated in Fig. 3.

### Additional Structural Restraints

The traditional interproton distance and torsion angle restraints that initially allowed three-dimensional structure determination of proteins can be supplemented by several other NMR observables. These observables include three-bond coupling constants (related to torsion angles), amide deuterium isotope effects on $^{13}$C$_\alpha$ shifts (related to the backbone $\psi$ angle (53)), $^{13}$C secondary chemical shifts (related to the backbone $\phi$ and $\psi$ angles), and $^1$H chemical shifts (influenced by short-range ring current effects from aromatic groups, magnetic anisotropy of C=O and C–N bonds, and electric field effects that arise from charged groups). These parameters are easily measured and therefore represent a useful source of additional information. Indeed, it was found that the agreement between observed and calculated values for these various parameters was frequently better for high-resolution X-ray structures than for the corresponding high-resolution NMR structures that were determined without these restraints. Inclusion of such additional NMR restraints has little impact on the precision of NMR structures but clearly improves their accuracy.

Greater improvements in the quality of structures generated from NMR data can be obtained by using a conformational database potential derived from dihedral angle relationships in databases of high-resolution protein crystal structures (54, 55). The rationale for including such “packing” potentials stems from uncertainties in the description of the nonbonded contacts, which present a key limiting factor in the attainable accuracy of protein NMR structures. The nonbonded interaction terms currently used in most algorithms have poor discriminatory power between high- and low-probability local conformations. It therefore seems reasonable to bias sampling during simulated annealing refinement to conformations with dihedral angles close to those that are known to be physically possible and energetically favorable. In this manner, the variability in the structures is primarily a function of the experimental restraints rather than an artifact of a poor nonbonded interaction model. Therefore, this procedure does not compromise the agreement with the experimental restraints, and deviations from idealized covalent geometry remain within experimental error.

### Orientational Restraints

Structure determination by NMR traditionally has relied exclusively on restraints whose information content is entirely...
local and restricted to atoms close in space. Specifically, NOE-derived short (~5 Å) interproton distance restraints, supplemented by coupling constants, 1H secondary shifts, and 1H long-range shifts are used. The success of the NOE-centered NMR methodology is based on the fact that short interproton distances highly restrict the conformational space of the molecule. However, numerous cases exist in which restraints that define long-range order would be highly desirable or necessary. In particular, the relative positioning of structural elements that do not have many short interproton distance restraints is critical for the accurate determination of the molecular structure.

Anisotropy should be greater than an order of magnitude. This typically means that the diffusion anisotropy should be greater than 34 m3/molecule, which is about 6 · 10−12 m3/molecule. For molecules in isotropic solution, this typically means that the magnetic susceptibility anisotropy should be greater than 0.2 Hz make the potential routine use of dipolar couplings for biomolecular structure determinations. Sufficiently high degrees of alignment are achieved in this manner, resulting in one-bond dipolar couplings of 5–40 Hz that are easily detectable by simply measuring the splittings in 2D- and 3D-coupled HSQC spectra. Additional experiments for extracting other types of RDCs and the development of more media that allow tuning the degree of molecular alignment followed suit. Limitations with respect to the long-term stability of the original DMPC/DHPC phospholipid-based liquid crystal phases prompted the replacement of the diacyl phospholipids by non-hydratable dialky analogs (64). Other alternative alignment media include quasi-ternary systems of surfactant/salt/alcohol, which are known to form Helfrich lamellar phases (65, 66), as well as lyotropic liquid crystals formed by alkyl-poly (ethylene glycol)/alcohol mixtures in water (67). Importantly, suspensions of charged, rod-shaped viruses and filamentous phages (68, 69) have proven very robust and popular over the last few years. Indeed, more than half of the NMR structures in the PDB that are solved to date with RDCs were aligned in colloidal phase solutions. Methodologies that do not involve any liquid crystalline media but rather exploit the anisotropy of strained polymeric gels also are employed increasingly. Vertical (70) or radial compression of polycrylamide gels (72) are introduced into the gel matrix simply by diffusion. Such gels are extremely stable and inert and even allow the study of proteins under denaturing conditions.

Several important reasons exist for having different alignment media available. First, not every medium is compatible with the properties of the protein or systems under investigation. Proteins that interact with membranes are clearly not compatible with phospholipid-based media, and very flexible or partially folded proteins have a tendency to strongly interact with bicelles. Likewise, negatively charged molecules, such as nucleic acids, tend to bind to positively charged lamellae, whereas positively charged proteins potentially can interact with negatively charged phosphate groups at neutral pH values. This interaction can result in an increase in the electrostatic component of the alignment, causing large linewidths or a collapse of the liquid crystalline phase. Second, because the alignment tensors in two different alignment media will exhibit different correlations, changing the alignment medium commonly results in different orientations of the solute molecule with respect to the magnetic field. The presence of different orientations is an important property that resolves the degeneracies in the orientation of a given interatomic vector in the relationship between dipolar coupling and intermolecular vector orientation. A RDC measured in a given alignment medium positions the vector between the two coupled partners on one of the two possible, oppositely oriented cones on a unit sphere. If the alignment tensor in the second medium has a different orientation relative to the molecular frame of the molecule, the same vector now will reside on two different cones. Thus, the true orientation of this particular interatomic vector will lie at the intersections between the two cones. Therefore, RDCs measured in several independent media allow one to define uniquely the associated vector orientations.
Naturally, using heteronuclear T\textsubscript{J}/T\textsubscript{J} ratios or RDCs in NMR structure determination and refinement assumes that internal motions are negligible. The magnitude of RDCs depends on the generalized order parameter S for internal motions of the interatomic vector (72), thus different contributions have to be considered, at least in principle. In practice, however, it has been found satisfactory to assume a uniform S value for all those residues for which heteronuclear relaxation measurements indicate a well-ordered conformation, as evidenced by experimental S\textsuperscript{2} values of 0.7–0.9 (corresponding to S values of 0.85–0.86), rather than to use individual, residue-specific S values. Therefore, dipolar coupling constraints for residues that experience either slow conformational exchange or low order parameters (S\textsuperscript{2} < 0.6) need to be excluded from the data set used for structure determination.

A key aspect in the use of either \textsuperscript{15}N/T\textsubscript{J}/T\textsubscript{J} ratios or RDCs for structure refinement is the need to determine the magnitude (D\textsubscript{0}) of the anisotropy and rhombicity (R) of the diffusion or the alignment tensor. These quantities can be obtained by examining the distribution of the measured \textsuperscript{15}N/T\textsubscript{J}/T\textsubscript{J} ratios (73, 74) from the powder pattern appearance of their distribution. Alternatively, a maximum-likelihood method for extracting D\textsubscript{0} and R (75), singular value decomposition for calculation of the Saupe order matrix (76), or prediction of the alignment tensor based on the shape of the solute molecule by using an obstruction model (77) can be employed.

**TROSY**

As discussed above, spectral overlap of resonances can be overcome to a certain degree by increasing the dimensionality of the spectra and by employing isotope-labeling approaches. However, limitations imposed by transverse relaxation pose a severe technical challenge, and deuteration alone cannot extend the application range of solution NMR to beyond ~60 kDa. For very large systems, TROSY methods (31) alleviate linewidths challenges. Such experiments work best with deuterated proteins that carry protonated amide groups or selectively protonated methyl groups.

Transverse relaxation increases proportionally with molecular mass because of the slower tumbling or decreased rotational correlation time \(\tau_2\). This increase leads to a broadening and a loss of signal intensity. T\textsubscript{J} for the amide nitrogen is dominated by dipole-dipole (DD) relaxation between \textsuperscript{1}H and \textsuperscript{15}N spins and by the chemical shift anisotropy (CSA) of each individual spin. A fluctuating, time-dependent magnetic field, B\textsuperscript{CSA}(t), is induced by CSA and modulated by Brownian motion. Similarly, DD relaxation induces a motion-dependent magnetic field that, together with CSA, perturbs the local magnetic field of the \textsuperscript{15}N spin and causes transverse relaxation and line broadening of the cross peak. Although the B\textsuperscript{CSA}(t) field is proportional to the overall B\textsubscript{0} magnetic field strength, B\textsuperscript{CSA}(t) is independent of B\textsubscript{0}, but its sign depends on whether the two spins \textsuperscript{1}H and \textsuperscript{15}N are parallel or antiparallel. Because B\textsuperscript{CSA}(t) and B\textsuperscript{DD}(t) both influence the relaxation of the \textsuperscript{15}N spin with the same angular and time dependence, the two fields either add or subtract. For one multiplet coherence, B\textsuperscript{DD}(t) opposes B\textsuperscript{CSA}(t), resulting in a narrow linewidth, whereas the other multiplet coherence relaxes fast because B\textsuperscript{CSA}(t) and B\textsuperscript{DD}(t) are additive, causing broad lines. The TROSY experiment aims for optimal compensation of B\textsuperscript{CSA}(t) with B\textsuperscript{DD}(t) by increasing B\textsubscript{0}; even if the optimal field strength of 1 GHz is not quite attainable yet, excellent compensation already is observed at 700 MHz. Technically, transverse relaxation optimization is achieved by eliminating the inversion (decoupling) of the scalar coupled \textsuperscript{1}H spin during \textsuperscript{15}N evolution and by eliminating the inversion (decoupling) of \textsuperscript{1}H during \textsuperscript{15}N acquisition. Alternatively, the scalar coupled \textsuperscript{13}N spin can also be eliminated during polarization transfer in so-called \textsuperscript{13}N, \textsuperscript{1}H CRIP-T (cross-correlated relaxation induced polarization transfer)-TROSY experiments (78). Basically, in all TROSY experiments, favorable (sharp) multiplet components are selected by varying schemes along the evolution (\textsuperscript{15}N or \textsuperscript{13}C) and acquisition (\textsuperscript{1}H) dimensions.

**Specific Labeling Approaches**

Isotope labeling methodologies have played pivotal roles in solution NMR studies of proteins from the earliest days. In many cases, labeling has greatly improved the quality of spectra by both reducing the number of resonances and by concomitantly narrowing linewidths. A ready 40 years ago, deuteration was used to eliminate the complexity of the one-dimensional \textsuperscript{1}H spectra of proteins by growth in D\textsubscript{2}O media, supplemented either with uniformly or selectively protonated amino acids. This process allowed monitoring the remaining resonances and helped to identify conformational changes that occurred during ligand binding (79, 80). Selective amino acid protonation in otherwise deuterated proteins or vice versa has been used regularly for spectral simplification and residue type assignment over the years (see Reference 81 for a review). Random fractional deuteration improves the quality of homonuclear proton 2-D spectra by removing contributions to proton linewidths from proton-proton dipolar relaxation and \textsuperscript{1}H-\textsuperscript{1}H scalar couplings. The decrease in sensitivity that results from the loss of protons is largely offset by a reduction in peak linewidths at deuteration levels of 50–75%. Superior spectral quality also ensues in many heteronuclear experiments for highly deuterated proteins, first noted in a \textsuperscript{13}C, \textsuperscript{15}N HSQC spectra of larger proteins (81). Deuteration in the context of triple-resonance spectroscopy first was employed for correlating sequential amides in a 4-D HN(COCA)NH experiment (82), followed by a suite of triple-resonance experiments for backbone assignments (83). The larger the protein, the more deuteration is necessary. Unfortunately, this comes at the cost of reducing the number of protons available for NOE-based interproton distance restraints.

A different technique used for incorporating isotopically labeled peptide segments within a protein sequence is based on protein splicing (84, 85). In this manner, signals of selected N- or C-terminal regions along a peptide chain can be made NMR observable. However, very few applications of segmentally labeled protein have been reported to date in structural NMR, and more robust methodologies along these lines are needed.

An extremely powerful labeling strategy was introduced by R. Rosen et al. (86) who developed a method that produces high
levels of deuterium at all positions in a protein with the exception of a select set of methyl groups. This approach relies on the fact that pyruvate is the metabolic precursor of Ala, Val, Leu, and the methyls (12 only). Therefore, the methyls of pyruvate can be used as the sole carbon source in D₂O-based minimal medium. Unfortunately, the presence of methyl isotopomers is a drawback both in terms of resolution and sensitivity. Therefore, an alternative approach was devised in which bacteria are grown in a highly deuterated, glucose-containing, medium supplemented with amino acid precursors and amino acids that have useful patterns of protonation and deuteriation (87). Despite the benefits of deuterium substitution that causes reduced heteronuclear transverse relaxation rates, associated deuterium isotope shifts for 1H and 13C resonances are unfortunate drawbacks. Fortunately, given the additivity and only weak structural dependence of deuteron isotope effects, chemical shift assignments between fully protonated and perdeuterated molecules can fairly easily be transferred. (88). This is not the case for partially deuterated systems because the presence of multiple deuterium-containing isotopomers at each site can complicate the spectra significantly. Although spin gymnastics has been applied to this problem, the best solution at present seems to be the elimination of all but CH₃ isotopomers during protein expression (87).

As an alternative to expression with defined precursors, Kainocho et al. (89) developed an elegant cell-free translation system for assembling proteins from synthetic stereoisotope-labeled amino acids. In this so-called SAIL (stereo-array isotope labeling) approach, each 13C(1H)₂ unit is replaced stereospecifically with 13C(2H)₃ and each 13C(1H)₃ unit is replaced stereospecifically with 12C(1H)₃. In addition, redundant methyl groups are converted to 12C(1H)₃, and six-membered aromatic rings are labeled with alternating 12C(2H)₂ and 13C(1H)₂ moieties. This labeling pattern preserves through-bond correlations for backbone, side chain assignments, yields sharper lines by decreasing long-range couplings and dipolar relaxation pathways, simplifies the coupling patterns for the easy extraction of 3J couplings and yields unambiguous stereo-specific assignments. As an additional bonus, the methyl and methylene labeling patterns make the analysis of side chain relaxation measurements more straightforward. The SAIL approach is particularly useful for high-throughput protein structure determination in structural genomics, where automated resonance assignments based on a limited number of proton shifts feed into automated structure determination algorithms.

Quality and Validation of NMR Protein Structures

The traditional representation of NMR structures as ensembles frequently provokes the question about the overall quality of NMR structures, compared to that of X-ray structures. Although this contribution is not focused on the algorithms used for NMR protein structure determination (see Reference 90 for a review) and their influences on the resulting ensembles of structures, it maybe instructive to briefly comment on this question. The validation of structures typically involves two aspects: 1) how well do the structures agree with the experimental data and 2) how do the structures compare with statistics derived from a reference database of high-quality protein structures in terms of standard geometric values (e.g., PROCHECK (91) and WHAT IF (92)). The latter aspect is common to both X-ray and NMR structures. With respect to experimental data, the main quality indicator for a crystal structure is the free R-factor that quantitatively measures the agreement between the observed and the calculated reflections. An equivalent NOE R-factor is problematic because NOEs generally are implemented in a qualitative manner and direct refinement against NOEs is not commonly used. Therefore, it is important for NMR structures to use a large set of experimental restraints and ensure that no violations of these restraints are present in the final ensemble of structures. In addition to the quality of the experimental input data, the energy function used in the final refinement step also influences the quality of protein NMR structures. Currently, the best quality indicator for NMR structures is an R-factor for NOE couplings (93). The dipolar R-factor expresses the ratio of the observed rms difference between measured and calculated values with that expected for a totally random distribution of vectors. In this manner a quantitative and readily interpretable measure of the agreement between measured and calculated dipolar couplings is obtained. R (dipolar) is readily amenable to complete crossvalidation, an important consideration in assessing quality.

Conclusions and Outlook

The development of highly sensitive multidimensional heteronuclear NMR experiments at high fields in conjunction with specifically tailored labeling approaches has propelled the field of protein structure determination by NMR towards larger molecular weight ranges. At present, proteins and protein complexes up to 50–60 kDa are amenable to detailed structural analysis in solution. In addition, although not covered in this review, NMR offers a unique means of probing molecular motions on the picosecond to the second time scales (94). Despite these advances, it needs to be stressed that several key requirements have to be satisfied to permit a successful structure determination of larger proteins and protein complexes by NMR. The system in hand must be soluble and should not aggregate up to concentrations of about 0.3–0.5 mM; it must be stable at room temperature (or at least ~10°C) for considerable periods of time; it should not exhibit significant conformational heterogeneity that may manifest itself in extensive line broadening; and finally, it must be amenable to the various labeling methodologies. Currently, still only relatively few examples exist in the literature of single chain proteins in the 30–50 kDa range whose structures have been solved by NMR. Likewise, only a handful of protein-DNA, protein-peptide complexes, and protein oligomers have been determined to date using these methods. One can anticipate, however, that over the next few years many more high-resolution NMR structures of proteins and protein complexes will become available. In addition, methods for studying large protein complexes are still being developed, and questions related to dynamics, binding, and conformational (allosteric) changes can be probed (95). There is no doubt that
NMR as a methodology has come a long way and will continue to provide structural and dynamical answers for biologic questions that are complementary to those obtainable by X-ray crystallography.

References


44. Reif B, Hening M, Grisargier C. Direct measurement of angles between three dihedral factors in high-resolution NMR. Science 1997;276:1230–1233.


One goal of chemical biology is to document and understand the macromolecular interactions that take place in the cell. In this review, we discuss the classic yeast two-hybrid method (and its derivatives) and how this method continues to provide insight into the number of protein–protein partnerships in a cell. In turn, these platforms have been adapted to explore a variety of interactions, such as those between proteins and small molecules. These systems collectively are called three-hybrid assays, and they provide new opportunities for the discovery of potential binding pairs. However, two- and three-hybrid assays also have significant disadvantages, such as incomplete coverage and high rates of false positives. With these issues in mind, we discuss emerging ways of minimizing the impact of these limitations using microarrays and mass spectrometry. Finally, chemical probes related to the three-hybrid concept are going beyond observation and providing active, rational control over individual protein–protein contacts. In these systems, bifunctional compounds are used to homo- or heterodimerize target proteins reversibly, thus altering their colocalization. By purposefully controlling protein–protein contacts, these chemical dimerization methods have progressed beyond observation and towards synthetic manipulation of protein function.
The two-hybrid method is robust and lends itself well to automation; thus, high-throughput applications of the two-hybrid approach have become increasingly common over the last decade. These projects encompass two phases: an initial phase of reagent development and a subsequent phase of high-throughput, two-hybrid screening. Methods for recombination-based cloning (4) and increased experience with large-scale cloning projects have enabled the generation of many successful bait and prey collections in numerous organisms (5-11). Subsequent high-throughput screening procedures typically employ array and pooling strategies driven by simple robotics (Fig. 1a).

In one approach, haploid yeast carrying bait and prey constructs, respectively, are mated to generate a diploid strain with both bait and prey for analysis of protein-protein interaction. This screen can be accomplished either by arraying the haploid strains before mating (generating an interaction "matrix") or by using bait/prey libraries (6, 7). The efficiency of these methods is improved by pooling; small pools of baits and preys are tested first, and individual interactions are subsequently confirmed through analysis of a particular pair (12-14).

Although the two-hybrid approach has clearly proven to be a powerful tool for biological discovery, it is not without its caveats. For example, some genuine interactions may be missed in a two-hybrid screen because a given protein may not be expressed properly as a binding domain- and/or activation domain-fusion (15); also, certain interactions may be too weak to be detected by two-hybrid methods, although the minimum affinity threshold for detection has been difficult to determine. Moreover, the two-hybrid method is prone to identifying false-positive interactions, and many possible sources of these artifacts exist (16). For example, a given protein pair may interact by yeast two-hybrid analysis but be precluded from interacting in vivo because of the distinct subcellular localizations of the two proteins (i.e., one protein may be nuclear, whereas the other may be cytoplasmic). In the two-hybrid method, both putative interacting proteins are forced to the nucleus by virtue of nuclear localization signals incorporated in the chimeras, which yields an artificial colocalization and correspondingly artificial potential for interaction. In general, this situation exemplifies a class of false-positive results by two-hybrid analysis, wherein the two hybrid-associated proteins possess the potential to interact in vivo but are restricted from doing so in their cellular context. A reflection of the yeast two-hybrid assay itself are also evident, which results in observed transcriptional activity independent of protein-protein interaction. This additional class of false-positive results is exemplified by bait/prey proteins that are capable of activating reporter transcription in the absence of a second hybrid (15).

Figure 1 The classic yeast two-hybrid method and derivatives. (a) Schematic diagram of the yeast two-hybrid approach, describing an interaction between protein X and protein Y. Protein X is fused to a transcription factor DNA-binding domain (the "bait" construct), and protein Y is fused to a transcription factor activation domain (the "prey" construct). (b) High-throughput applications of the yeast two-hybrid method use mating of haploid strains carrying bait and prey, respectively. The resulting diploid strain serves as a host for a reporter Gene expression. (c) In one approach, haploid yeast carrying bait and prey constructs, respectively, are mated to generate a diploid strain with both bait and prey for analysis of protein-protein interaction. This screen can be accomplished either by arraying the haploid strains before mating (generating an interaction "matrix") or by using bait/prey libraries (6, 7). The efficiency of these methods is improved by pooling; small pools of baits and preys are tested first, and individual interactions are subsequently confirmed through analysis of a particular pair (12-14). Although the two-hybrid approach has clearly proven to be a powerful tool for biological discovery, it is not without its caveats. For example, some genuine interactions may be missed in a two-hybrid screen because a given protein may not be expressed properly as a binding domain- and/or activation domain-fusion (15); also, certain interactions may be too weak to be detected by two-hybrid methods, although the minimum affinity threshold for detection has been difficult to determine. More significantly, the two-hybrid method is prone to identifying false-positive interactions, and many possible sources of these artifacts exist (16). For example, a given protein pair may interact by yeast two-hybrid analysis but be precluded from interacting in vivo because of the distinct subcellular localizations of the two proteins (i.e., one protein may be nuclear, whereas the other may be cytoplasmic). In the two-hybrid method, both putative interacting proteins are forced to the nucleus by virtue of nuclear localization signals incorporated in the chimeras, which yields an artificial colocalization and correspondingly artificial potential for interaction. In general, this situation exemplifies a class of false-positive results by two-hybrid analysis, wherein the two hybrid-associated proteins possess the potential to interact in vivo but are restricted from doing so in their cellular context. A reflection of the yeast two-hybrid assay itself are also evident, which results in observed transcriptional activity independent of protein-protein interaction. This additional class of false-positive results is exemplified by bait/prey proteins that are capable of activating reporter transcription in the absence of a second hybrid (15).
Two-hybrid data validation

Considering the caveats indicated above, yeast two-hybrid methods are often coupled with secondary, high-confidence assays to define a truer interaction data set. Toward this end, Li et al. (8) undertook a large-scale yeast two-hybrid analysis of protein–protein interactions in Caernarvon’s elegans and used co-affinity purification assays to validate 143 putative interactions. Co-immunoprecipitation provides a high-confidence measure of protein interaction; however, the approach is labor intensive and presents a greater challenge for automation. Other related approaches have also been employed to verify protein interactions that are suggested by two-hybrid studies. For example, Tong et al. (17) integrated results from two-hybrid analysis with protein interaction data sets derived from phage display experiments to yield a subset of high-confidence interactions associated with Sk3 domain-containing proteins in the budding yeast. Protein localization data have also proven useful in identifying biologically relevant interactions within large two-hybrid data sets because interacting proteins should localize within the same subcellular compartment (18). Collectively, these and other approaches (see the section titled “Combining Two- and Three-Hybrid Screens with Other Technologies”) offer significant promise toward the construction of high-confidence protein interaction maps on a genome-wide scale.

Reverse two-hybrid approaches

The classic two-hybrid method has been modified over time to yield a variety of derivative approaches that offer complementary data. For example, the “reverse” two-hybrid method was developed to identify drugs, mutations, and/or proteins capable of disrupting a known interaction (Fig 3c). Briefly, the reverse two-hybrid approach employs a reporter gene that encodes a counter-selectable marker (e.g., the yeast gene URA3) (19). The URA3 gene product is required for growth on medium lacking uracil; however, this product is toxic to yeast cells grown on medium containing 5-fluoroorotic acid (5-FOA). By the reverse two-hybrid method, dissociation or inhibition of a hybrid interaction results in loss of URA3 expression, which permits cell growth in the presence of 5-FOA. Similar loss-of-interaction detection schemes have been devised using the Escherichia coli tet repressor (tetR) as a reporter (20). Expression of tetR represses expression from a second reporter containing an upstream tetR-binding site and, conversely, disruption of the interaction restores expression from this reporter. Although the affinity requirements and dynamic range of these methods have yet to be fully described, they have proven useful in proof-of-principle studies.

Membrane-associated two-hybrid systems

In the classic two-hybrid system, protein interaction is assayed in the nucleus, a subcellular environment for the detection of many protein–protein interactions—particularly those involving membrane proteins. As an alternative to nuclear two-hybrid assays, several groups have developed membrane-associated two-hybrid systems that use signaling components, such as Ras (21). The membrane-associated Ras protein binds and hydrolyzes GTP, alternating between an active GDP-bound form and an inactive GDP-bound state. These states are modulated by GTase-activating proteins that promote GTP hydrolysis and guanine nucleotide exchange factors (GEFs) that stimulate GDP release/GTP uptake. Aronheim et al. (22, 23) used Ras signaling as the basis for a two-hybrid system, substituting the human GEF Hsos for the native yeast RAS GEF, Cdc25p, in yeast cells carrying a mutant allele of cdc25. Using a hybrid consisting of a target protein fused to Hsos and a second putatively interacting protein localized to the cell membrane by a myristylation tag, protein–protein interaction recruits Hsos to the cell membrane, activating Ras, which, in turn, promotes cell survival and growth.

Takoff and colleagues (24) presented an extension of this system for the identification of proteins binding 3-phosphoinositide second messengers, PtdIns(3,4)P2 and PtdIns(3,4,5)P3, in vivo. In this system, the putative PtdIns(3,4)P2/PtdIns(3,4,5)P3 binding protein (or specifically just its predicted pleckstrin homology, PH, domain) is fused to constitutively activated GTβ-bound Ras. This hybrid protein is coexpressed with phosphatidylinositol 3-kinase (PI3K), which generates the 3-phosphoinositides in a temperature-sensitive Ras mutant. Binding of the target protein’s PH domain to the 3-phosphoinositides draws the hybrid protein to the cell membrane where the activated Ras substitutes for the nonfunctional mutant, which enables cell growth. Medici and colleagues (25) developed an additional membrane-based two-hybrid assay with bait and prey constructs consisting of the yeast G protein α subunit Gpa1p and the α-factor receptor Ste2p. By this approach, interaction between a Gpa1p chimera and Ste2p hybrid drives a signaling cascade that permits cell growth in a gap1 mutant background.

Imaging protein–protein interactions in live cells

In contrast to the classic yeast two-hybrid method and its close derivatives, chimeric protein–protein interactions may also be assessed in many organisms using readouts easily visualized in live cells. For example, Sanjiv Gambhir’s group developed a modified mammalian two-hybrid system using the Gal4p-DNA-binding domain, herpes simplex virus VP16 activation domain, inducible promoters, and the firefly luciferase gene as a reporter (26). In a proof-of-principle study, the helix-loop-helix proteins Id and MyoD were shown to interact, which results in luciferase activity visualized by standard methods of bioluminescence imaging in cell culture and in 293 T cells implanted in mice (26).

Protein–protein interactions may also be detected in living cells using hybrids that consist of fluorescent proteins capable of undergoing fluorescence resonance energy transfer (FRET). FRET refers to an exchange of energy occurring between two fluorescent molecules, with fluorons satisfying the following two criteria: 1) The emission spectrum of one fluorophore (the donor) must overlap with the excitation spectrum of the second fluorophore (the acceptor), and 2) the fluorors must be within approximately 10 nm of each other (reviewed in Reference 27).
Two- and Three-Hybrid Systems

Protein X
Protein Y
440 nm
500 nm
440 nm
545 nm
480 nm
545 nm
500 nm
480 nm

(a) Acceptor: Donor: Protein Y Protein X YFP CFP
(b) Acceptor: Deep Blue C Donor: Rluc
GFP

Figure 2 Imaging protein–protein interactions by FRET and BRET. (a) Diagram illustrating FRET between the donor CFP-fusion and the acceptor YFP-fusion. (b) Detection of protein–protein interaction between X and Y by BRET using the Renilla luciferase (Rluc) as a bioluminescent donor and GFP as a fluorescent acceptor. The substrate for Rluc is deep blue coelenterazine. Constructs not drawn to scale.

If these requirements are met, the donor fluorophore (e.g., cyan fluorescent protein, CFP) can transfer the resonance energy of its excited state to the acceptor (e.g., yellow fluorescent protein, YFP), which causes it to fluoresce (Fig. 2a). The efficiency of this energy transfer depends on the inverse sixth power of the distance between donor and acceptor; therefore, the detection of FRET is a sensitive measure of the mutual proximity of two fluoros. Thus, FRET can be used to detect protein–protein interactions with high confidence and is amenable to implementation on a fairly large scale, using YFP- and CFP-chimeras. Recently, Müller et al. (28) used FRET to identify core proteins of the yeast spindle pole body, thus highlighting the applicability of this approach toward defining a target subset of protein–protein interactions.

Along similar lines, bioluminescence resonance energy transfer (BRET) also presents a technology for the detection of protein–protein interactions based on energy transfer between two proteins in close spatial proximity. As presented by De and Gambhir (29), BRET can be implemented using the Renilla luciferase (Rluc) as a bioluminescent donor and mutant green fluorescent protein (GFP) as a fluorescent acceptor (Fig. 2b). Proteins predicted to interact are fused to luciferase and GFP; interaction between the Rluc and GFP chimeras results in excitation of GFP by resonance energy transfer from the reaction of Rluc with its substrate, deep blue coelenterazine. De and Gambhir used this approach to detect Id–MyoD and FKBP12FRAP interactions in both cultured cells and in deeper tissues of mice with implanted cells overexpressing the fusion constructs (29).

Although FRET- and BRET-based technologies offer many advantages associated with the ability to image protein interactions in live cells, both approaches are substantially limited by the spatial constraints of energy transfer. Many genuine protein interactions may be missed in FRET/BRET approaches because donor and acceptor are not brought in sufficiently close proximity for resonance energy transfer by target protein dimerization (27). The false-negative rate associated with these technologies is, therefore, presumed to be high. Conversely, however, these approaches are prone to a very low rate of false-positive results, and detected interactions can be considered genuine with few exceptions.

Split-reporter strategies

Since the mid-1990s, numerous studies have advanced clever alternatives to transcription factor-based two-hybrid methods through technologies that use various split-reporter systems (30). In particular, Johnson and Varshavsky (31) developed an interesting split-protein sensor of protein interactions using the small protein ubiquitin (Ub). Ubiquitin is a 76-amino acid protein best characterized for its role in protein degradation: Covalent attachment of Ub to a target protein marks that protein for degradation by the 26 S proteosome (32, 33). In eukaryotes, newly formed Ub fusions are cleaved by site-specific proteases after the last Ub residue at the Ub–target protein junction, provided that the Ub protein is properly folded. The Varshavsky group took advantage of this fact, splitting Ub into amino- and carboxy-terminal halves and fusing each to one of two putatively interacting membrane proteins (Fig. 3a). The amino-terminal Ub fragment was modified additionally by the introduction of a point mutation that decreased its affinity for the carboxy-terminal Ub fragment. This carboxy-terminal fragment was fused to a transcription factor, which provided a convenient readout for this system. During target protein interaction, the amino- and carboxy-terminal Ub fragments are brought into close proximity, which drives partial reassociation of Ub. This reconstituted Ub is recognized by its proteases, which results in...
cleavage and release of the transcription factor; protein interaction thus can be monitored by the activity of reporter genes (e.g., H15 and lacZ) recognized by the released transcription factor. This split-Ub system has been used successfully in several studies in organisms ranging from yeast to humans (34, 35) and, in fact, was used recently to isolate proteins interacting with human Bap31 and Erb3 (36). This method has also been applied to membrane proteins (37).

The Blau group has presented an alternative protein fragment complementation approach for detecting protein-protein interactions using the well-studied reporter β-galactosidase (β-gal) (38). The lacZ gene is known to exhibit intracistronic complementation, wherein pairs of inactive β-gal deletion mutants are capable of complementing each other in trans, assembling to yield an active enzyme. To use β-gal fragments as the basis for a protein interaction system, Rossi et al. (38) selected β-gal mutants (ΔN and ΔC) with low affinity for each other and fused each mutant to one of two target proteins. Target protein interaction drives the formation of active β-gal, which provides a convenient and quantitative monitor of protein dimerization. Recently, Wehman et al. (39) have modified this system, using a truncated low-affinity ω-peptide that weakly complements the ω fragment, such that recognition between the two fragments in the absence of forced dimerization is insufficient to maintain a complemented enzyme. Accordingly, reversible interactions can be monitored by this method, and the activity of the enzyme is directly related to the local concentration of the enzyme fragments. This approach has been used successfully as an assay of nuclear import/export (39) and as a monitor of G-protein-coupled receptor internalization (40). As with other ω-gal readout systems, this complementation approach allows for convenient signal visualization, which offers a method for the direct assessment of protein interaction in the context of native cell compartments and environments. The Michnick group has also developed an elegant split-reporter system, based on reconstitution of dihydrofolate reductase (DHFR) (41). In this system, cell survival requires functional DHFR; thus, a successful protein-protein interaction can be identified readily by selection. This system has been applied to the identification of peptide sequences that bind the Ras-binding domain of Raf (42).

The β-gal luciferase enzyme has also served as the basis of an informative protein fragment complementation approach for detecting protein-protein interactions in cells and living animals (43). By this method, amino- and carboxy-terminal fragments of luciferase are fused to target proteins of interest, such that protein-protein interaction reconstitutes active luciferase (Fig. 3c). Luker et al. have used this approach to investigate STAT1 activity and yielded data suggesting that STAT1 proteins homodimerize in the absence of ligand-induced phosphorylation (43). In addition to the firefly enzyme, Renilla luciferase has also been used, for example, to explore protein kinase A interactions (44). These luciferase complementation imaging strategies share several advantages of the systems described above and, in particular, provide near real-time detection of protein-protein interactions in intact cells and living animals. Tom Keppola’s group has developed a distinct and powerful variation on split-protein detection strategies, by using fragments of a fluorescent protein as part of a technique termed bimolecular fluorescence complementation (BIFC) (45). In this method, nonfluorescent fragments of enhanced yellow fluorescent protein are fused to putatively interacting target proteins; protein interaction brings these fragments into close proximity, which results in detectable yellow fluorescence (Fig. 3d).

**Figure 3** Protein fragment complementation strategies. (a) Schematic representation of the split-ubiquitin yeast two-hybrid system. The amino-terminal fragment of ubiquitin is designated “Nub;” and the carboxy-terminal fragment is designated “Cub.” Interaction of proteins contains a deletion of residues 11–41. Fragments not drawn to scale. (c) Firefly luciferase complementation by protein-protein interaction results in fluorescence at 545 nm. Again, fragments are not drawn to scale. (d) Bimolecular fluorescence complementation using split mutants of YFP, as applied in mammalian cells, results in fluorescence at 545 nm. Again, fragments are not drawn to scale.
For genuine in vivo interaction partners, the BiFC approach is typically quite successful, although the fluorescence intensity produced by target protein dimerization is routinely less than 10% of levels generated by intact yellow fluorescent protein (45). To date, BiFC has been used to identify interactions among many types of proteins in numerous cell types and organisms (reviewed in Reference 46) and provides a method not only to identify protein interactions but also to define the subcellular localization of protein complexes in living cells (47). The BiFC approach is amenable to many additional applications (47); however, it is not without its limitations. Specifically, the fluorescent protein fragments in a BiFC system associate irreversibly under some conditions and exhibit an intrinsic affinity for each other in the absence of target protein interaction, which limits both the sensitivity and scope of BiFC-based studies. Nonetheless, the approach holds significant potential as a generally applicable means of imaging protein interactions and complexes in single live cells, with future promise for even broader applications in a wide range of organisms.

Three-Hybrid Screens

The classic three-hybrid screen

Although the two-hybrid is an effective option for protein–protein interactions, this platform is not suitable for other types of macromolecular contacts, such as those involving small molecules or nucleic acids. The three-hybrid screen was developed specifically to address this need; for reviews see References 48–50. The third component in these systems is a small molecule or RNA that serves as the “bait” (Fig. 4). A general feature of three-hybrid “baits” is that they have two distinct, nonoverlapping domains; the anchoring end has a well-characterized binding partner, whereas the other side is used to query a collection of potential “preys.” For example, the first compound used in a three-hybrid was composed of a dexamethasone (Dex) anchor attached to FK506 (2). Dex has high affinity for the glucocorticoid receptor (GR), and, therefore, it is bound to the promoter via its tight association with this well-characterized binding partner, whereas the other side is used to query a collection of potential ligands.

References 48–50. The third component in these systems is a small molecule or RNA that serves as the “bait” (Fig. 4). A general feature of three-hybrid “baits” is that they have two distinct, nonoverlapping domains; the anchoring end has a well-characterized binding partner, whereas the other side is used to query a collection of potential “preys.” For example, the first compound used in a three-hybrid was composed of a dexamethasone (Dex) anchor attached to FK506 (2). Dex has high affinity for the glucocorticoid receptor (GR), and, therefore, it is bound to the promoter via its tight association with this well-characterized binding partner, whereas the other side is used to query a collection of potential ligands.

A general feature of three-hybrid “baits” is that they have two distinct, nonoverlapping domains; the anchoring end has a well-characterized binding partner, whereas the other side is used to query a collection of potential “preys.” For example, the first compound used in a three-hybrid was composed of a dexamethasone (Dex) anchor attached to FK506 (2). Dex has high affinity for the glucocorticoid receptor (GR), and, therefore, it is bound to the promoter via its tight association with this well-characterized binding partner, whereas the other side is used to query a collection of potential ligands.

Although the two-hybrid is an effective option for protein–protein interactions, this platform is not suitable for other types of macromolecular contacts, such as those involving small molecules or nucleic acids. The three-hybrid screen was developed specifically to address this need; for reviews see References 48–50. The third component in these systems is a small molecule or RNA that serves as the “bait” (Fig. 4). A general feature of three-hybrid “baits” is that they have two distinct, nonoverlapping domains; the anchoring end has a well-characterized binding partner, whereas the other side is used to query a collection of potential “preys.” For example, the first compound used in a three-hybrid was composed of a dexamethasone (Dex) anchor attached to FK506 (2). Dex has high affinity for the glucocorticoid receptor (GR), and, therefore, it is bound to the promoter via its tight association with this well-characterized binding partner, whereas the other side is used to query a collection of potential ligands.
with a LexA–GR fusion (Fig. 4a). The query end of this divergent compound recruits a transcriptional activator that is fused to FK506-binding protein (FKBP). The reporter gene is typically a selection marker, such as LEU2, or a directly detectable signal, such as fluorescence or luminescence. Thus, this relatively straightforward adaptation of the two-hybrid platform has the potential to identify binding partners for a variety of macromolecules.

Features of the three-hybrid

Several recent studies have examined the strengths and limitations of the three-hybrid. For example, this system has been successfully adapted for use in yeast (51) and bacteria (52). In addition, extensive analyses of the kinetic and affinity requirements have been reported (53, 54). These studies have revealed that, although the dynamic range is small (\( \sim 1 \) order of magnitude), the three-hybrid signal is dependent on the association strength of the interaction, with a minimal \( K_D \) of \( \sim 50 \) nM (55). This in-depth work has helped refine our understanding of the key interactions, and, through these studies and others (48), the features of three-hybrid screens that influence the signal have begun to be revealed.

Target identification for small molecules

In forward chemical genetics, a system is treated with a collection of compounds and a candidate molecule that produces the desired phenotype identified (56–58). One main challenge in this field is that each new chemical probe requires follow-up studies to identify its cellular protein target (59). Accordingly, one primary use of the three-hybrid system is to find these targets. However, the challenge is that the molecule of interest must be covalently coupled to the anchoring compound without a loss of binding affinity (Fig. 4a). One of the most extensive synthetic studies was reported by the Verdine group, in which they generated a library of 320 tetrahydrooxazepines attached to an analog of FK506 on solid phase (60). This study validates that large collections of bifunctional ligands are possible, but, in practice, only a few query molecules have been screened. For example, the Kley group used a three-hybrid approach, coupled with affinity chromatography and in vitro kinase assays, to identify CDK1, CDK2, and other kinases as the protein targets of the kinase inhibitors, purvalanol B, roscovitine, and indeno[1,2-c]pyrazole, IP-1 (51).

Reverse three-hybrid

The downside to the three-hybrid as a target validation tool is that a suitable attachment point must be found on the query molecule. Oftentimes, selection of an appropriate site requires structure-activity relationship (SAR) studies to avoid damaging the compound’s binding affinity. Moreover, each chemical entity requires individual synthetic effort because the properties of

![Figure 5](image-url)
the linker can have important effects on the efficiency of the system (51, 54). One way to avoid the need for synthetic manipulation of the query molecule is to conduct a reverse three-hybrid screen. Similar to the reverse two-hybrid discussed above, a library of compounds is applied to a two- or three-hybrid reporter (Fig. 4f), and if the molecule interrupts the reconstituted complex, the reporter is deactivated. In one example, this approach was used to identify inhibitors of the helicase-prime interaction in Bacillus stearothermophilus (61). A related system involves expression of potential binding proteins in the presence of an intact three-hybrid contact (Fig. 5e) (62). These reverse three-hybrid systems do not require extensive synthetic manipulations and are ideal for those examples in which little SAR is available or no suitable attachment point is found.

**Screens for enzymatic activity and enzyme inhibitors**

The Cornish group has reported new variations of the three-hybrid, which all involve an enzyme as a necessary fourth component (Fig. 5b). In their first report, they developed a way to study cephalosporinase activity in yeast; they installed a cephalosporinase-sensitive site within a Dex–MtX conjugate such that the enzymatic activity would cleave the bifunctional component (63). Later variations of this platform have been used to evolve glycosynthase activity (64). Briefly, formation of a glycosidic link between detached sugar analogs reconstituted the bifunctional "bait" that subsequently activated the DHFR–GR-based LEU2 selection marker (Fig. 5g). Using this method, they evolved a glycosynthase with fivefold better enzymatic activity. Related methods have been used to study phosphorylation-dependent protein–protein contacts (65, 66), which demonstrates the versatility of this approach. These systems all take advantage of the signal-amplifying properties and high sensitivity of three-hybrids to monitor enzyme-dependent interactions in living systems.

**RNA–RNA interactions and RNA-binding proteins**

A rather adaptation of the three-hybrid replaces the bifunctional small molecule with an RNA molecule (first described by SenGupta et al. in 1996 (67) and reviewed more thoroughly by Jaeger et al. (68)). The bifunctional RNA is composed of two distinct domains: one part has a well-characterized affinity for an RNA-binding protein (typically MS2 coat protein or Hiv-1 Rev-MtX), whereas the query RNA is used to pan for possible "preys" (Fig. 4b). This system has been used to verify aptamer interactions (69) and to identify RNA–RNA interactions (70); also it has been used successfully to characterize over 20 RNA-binding proteins (68). Additionally, this method has been quite successful in revealing the key interacting residues in known RNA-binding proteins, including p53, HIV Tat, FBF-1, and histone hairpin binding protein (HHP). By random or directed mutagenesis (68, 71-74). Like other three-hybrid systems, the kinetics and binding constants are important for the efficiency of this system (75). As evident by the more than 60 published reports of successful three-hybrid screens (68), this method has proven to be a powerful addition to the arsenal of molecular and chemical biologists.

**Membrane-associated three-hybrid screens**

As mentioned above, advantages are gained by screening in non-nuclear compartments, but adapting the two- and three-hybrid technologies for different reporter strategies can be challenging. The Kley group has developed a system that they call MA SPIT as one solution (76). In this system, erythropoietin receptor is expressed as a fusion with DHFR in mammalian cells, and methotrexate-coupled bait molecules are used to recruit the prey (Fig. 5c). This membrane-proximal juxtaposition triggers Jak2-mediated phosphorylation of STAT3 and activation of a downstream reporter. This system was used to identify a family of ephrin receptor tyrosine kinases as targets of the inhibitor, PD173955 (76). Although this system is relatively complex, it represents an important advance because it allows screening in the cytosol and opens the possibility of finding contacts in their native subcellular environment.

**Combining Two- and Three-Hybrid Screens with Other Technologies**

Both two- and three-hybrid screens are limited by technical problems, such as high false-positive rates and incomplete coverage. This last problem is best illustrated by comparing the hits in otherwise identical screens. Parrish et al. compared the hits from three screens of 24,000 protein–protein interactions (~54% of the estimated total) in Drosophila and found remarkably little overlap (only 2 gene common to all three, and only 61 shared by at least 2 experiments) (77). Similar trends were seen in human two-hybrid systems, and, together, these findings strongly suggest that the coverage of possible interactions is incomplete (77). Insufficient data prevents similar comparisons in three-hybrid systems, but it is presumed that similar problems will be encountered. Minimizing the impact of these limitations might be best achieved by coupling two- and three-hybrid screens with other methods. Specifically, recent advances in mass spectrometry and microarrays are being used to investigate and confirm protein–protein interactions. High-throughput mass spectrometry is being used to both identify protein–protein contacts and to investigate these interactions on a proteome-wide scale (78-80). Thus, any interactions identified by two- or three-hybrid screens might be confirmed by this means. Finally, advances in microarray profiling offer promise to those screening for novel protein–protein or protein–ligand interactions. In these techniques, potential binding partners are labeled (e.g., with fluorescence or other reporter) and passed over patterned proteins, small molecules, or carbohydrates arrayed on a chip. After removing nonspecifically adsorbed material, the remaining "hits" represent interacting pairs. Technical advances over the past 10 years have positioned mass spectrometry and microarrays as viable tools for performing large-scale protein interaction...
Chemical Dimerization: Three-Hybrids for Conditional Regulation of Protein Function

Chemical inducers of dimerization

Although two- and three-hybrid screens are useful ways to identify new protein–protein and protein–drug contacts, the next set of questions require one to control these interactions rationally and to confirm their importance in a biological process. Toward this goal, the chemical dimerization systems developed in the mid-1990s by Crabtree and Schrieber are ideal tools (89–93). In these systems, a chemical inducer of dimerization (CID) is used to force the conditional homo- or hetero-dimerization of two proteins (for more in-depth reviews of these systems, see References 94–97). The target proteins are expressed as fusions with known drug binding domains such that addition of the bifunctional molecule causes their tight association (Fig. 6a). Although other pairs have been used (98–102), the most common systems are the homodimerization of FKBP by dimers of FK506 and the heterodimerization of FKBP with the FK506-rapamycin binding (FRB) domain by analogs of rapamycin. The shared feature of these CIDs is that they have two nonoverlapping domains; thus, they can bridge two proteins. A further common property of these systems is that they are reversible; thus, they offer the opportunity to regulate protein function conditionally. Whereas two- and three-hybrid screens are used to identify binding partners, CID systems are typically used to test whether protein colocalization is necessary and sufficient to instigate a biological response.

CIDs are used in many biological applications

Chemical dimerization has been remarkably successful at addressing previously inaccessible biological questions (Fig. 6a). For example, recent studies have explored the mechanisms of complex pathways, such as chromatin disassembly (103) and learning and memory (104). In addition, this method has proven quite adept in studies aimed at understanding signaling through the T-cell receptor (89), vasopressin receptor (105), GTPases (110), and phosphoinositides (111, 112). Similarly, this procedure has been used to gain regulatory control over important cellular processes, such as apoptosis (113, 114), preRNA splicing (115), translation initiation (116), glycosylation (64, 117), pathogen recognition (118, 119), and secretion (120, 121). This list of applications is necessarily brief, because greater than 300 manuscripts have been published across numerous scientific areas using CID systems (http://www.ariad.com/wt/page/ktb_references). This expansive literature illustrates the widespread acceptance of this method in the scientific community. In addition, this breadth makes a comprehensive overview of the area difficult, and the interested reader is encouraged to pursue several excellent reviews (49, 93–97, 122).

Inhibition of protein–protein interactions

A recent use of CIDs is to inhibit protein–protein interactions. One reason that this class of interactions has been difficult to block is that they typically involve a large and complex surface (123). Therefore, typical drug-like molecules are often too small to block these contacts effectively. This problem is particularly evident in amyloid formation, in which high-affinity protein–protein interactions lead to formation of cytotoxic, oligomeric structures (124). To alleviate this problem, we used a bifunctional CID that binds tightly to amyloid on one end and to FKBP on the other (Fig. 6b) (125). Recruitment of FKBP to the amyloid interface blocked the interaction 40-fold better than “regular” inhibitors (125, 126). These studies established that CIDs could be used in unconventional ways to engage biological systems and achieve desired outcomes.

Conditional protein stability

The chemical dimerization platform has been adapted for several other new applications. One approach is to recruit the target protein to the proteolytic machinery using CID-directed engagement leads to proteosomal degradation and concurrent, conditional removal of its function (127, 128). In another approach, an intrinsically unstable FRB tag is added to the target, which causes degradation of the entire chimera (Fig. 6c) (129–131). More than 5 kcal/mol of folding energy is rescued by CID-mediated recruitment of FKBP (130), and, thus, this method can be used in cultured cells, mice, and Xenopus laevis to regulate protein function conditionally (129, 130, 132). This method has proven useful in studies of developmental processes; in one recent example, unstable FRB fusions were used to learn that the kinase GSK3β is required for midline fusion during two brief times in mouse gestation (133). An improved version of this method has also been reported recently by the Wandless group; in these systems, stability of the tag can be controlled without the need for a recruited protein partner (134, 135). The Wandless group found that unstable FKBP proteins can be attached to a variety of targets and that a modified rapamycin analog, termed Shield-1, is sufficient to recover the stability of the chimeras. This simplified system has been applied to study falcipain function in Plasmodium falciparum (126) and the phenotypes of essential genes in Toxoplasma gondii (137). The goal of these technologies is to regulate protein function at the posttranscriptional level. By targeting this stage of a protein’s lifetime, one avoids the lengthy transcriptional and translational process to generate more rapid conditional systems. In practice, these systems typically display changes in protein levels over 1 or more hours.
Figure 6  Chemical dimerization. (a) The general components of homo- and hetero-dimerization systems are shown schematically. The binding proteins (X and Y) expressed as chimeras with the targets and these fusions are brought into proximity by the bifunctional small molecule. A variety of molecules have been used as the chemical inducer of dimerization (CID; see text for details). (b) CID-based approach to inhibiting difficult protein–protein contacts; a readily available cellular protein, FKBP, is used to preclude amyloid formation stably. (c) Inducible stabilization. Three variations on the CID system involve the conditional stabilization or destabilization of one protein target. A representative system is shown; an unstable domain leads to degradation of the target, but recruitment of the binding partner recovers protein levels and function. (d) Conditional protein mislocalization is shown schematically, with the example of nuclear import used to illustrate the general components. Reconstitution of a target with an NLS favors temporary nuclear uptake, whereas removal of the small molecule reverses this process.

Conditional control over subcellular localization

Although some biological processes occur over hours, other questions may require more rapid regulation. Conditional protein mislocalization offers fast kinetics and, thus, may be suitable for these specific situations. In these systems, a CID is used to append a subcellular address signal reversibly to the protein of interest (Fig. 6d) (138, 139). Because a protein’s function is often linked to its position within the cell (e.g., a transcription factor cannot be active in the cytoplasm), this system offers rapid and selective control.

Conclusions and future directions

Chemical dimerization and three-hybrid systems are being applied to numerous different biological systems, in a variety of organisms, and in a growing number of technological variations. Likewise, the classic two-hybrid is being reimagined in new ways, such as fragment complementation. At the same time, improvements in parallel technologies, such as high-content microscopy screening, mass spectrometry, and microarray profiling, have offered new opportunities. Together, this suite of methods is being used to study and control macromolecular contacts. As chemical biologists continue to develop, refine, and reimage two- and three-hybrid techniques, more insights will be made.

Acknowledgments

The authors thank Dennis Coles and members of our groups for comments on the manuscript.

References

20. Shih HM, Goldman PS, DeMaggio AJ, Hollenberg SM, Good-
19. Vidal M, Brachmann RK, Fattaey A, Harlow E, Boeke JD.
18. Wiwatwattana N, Kumar A. Organelle DB: a cross-species
17. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castag-
16. Serebriiskii IG, Golemis EA. Two-hybrid system and false pos-
15. Fields S. High-throughput two-hybrid analysis. The promise and
10. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A,
8. Sos is sufficient for activating the Ras signaling pathway. Cell
7. Isolation of an AP-1 repressor by a novel method for detecting
6. Approaches to detection and elimination. Methods Mol.
5. Cervantes S, Gonzalez-Duarte R, Marfany G. Homodimerization
4. Scheper W, Thaminy S, Kais S, Stagljar I, Romisch K. Coor-
3. Rechsteiner M. Natural substrates of the ubiquitin proteolytic
1. Stalperski IG, Golemis EA. Comparison of PH domain-containing targets of phosphatidylinos-
Two- and Three-Hybrid Systems

58. Strausberg RL, Schreiber SL. From knowing to controlling: forward chemical genetics: progress and obstacles.

56. Lokey RS. Forward chemical genetics: progress and obstacles.


47. Hynes TR, Mervine SM, Yost EA, Sabo JL, Berlot CH. Live cell imaging of Gs and the beta2-adrenergic receptor: demonstrates that both alphas and betagamma7 internalize upon stimulation and exhibit similar trafficking patterns that differ from that of the beta-adrenergic receptor. J. Biol. Chem. 2004;279:44101–44112.


Further Reading

The Saccharomyces Genome Database (www.yeastgenome.org) details an extensive collection of two-hybrid screens and other large-scale protein interaction studies. Ariad Pharmaceuticals maintains a useful and comprehensive database of CID-related references (http://www.ariad.com/ar/page/kits references). They also have made ARGENT homo- and hetero-dimerizer systems available for research use.

See Also

Protein Interaction Networks, Chemical Tools to Elucidate Protein-Protein Interactions. Array-Based Techniques for Proteins

Mass Spectrometry (MS) for Proteins and Protein Complexes Forward Chemical Genetics
Computational Approaches in Drug Discovery and Development

Honglin Li, Mingyue Zheng, Xiaomin Luo, Weiliang Zhu and Hualiang Jiang, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China

doi: 10.1002/9780470048672.wecb098

Recently, computational approaches have been considerably appreciated in drug discovery and development. Their applications span almost all stages in the discovery and development pipeline, from target identification to lead discovery, from lead optimization to preclinical or clinical trials. In conjunction with medicinal chemistry, molecular and cell biology, and biophysical methods as well, computational approaches will continuously play important roles in drug discovery. Several new technologies and strategies of computational drug discovery associated with target identification, new chemical entity discovery, and lead optimization will be the focus of this review.

Drug research and development (R & D) is a comprehensive, expensive, and time-consuming enterprise, and it is full of risk throughout the process (1). Numerous new technologies have been developed and applied in drug R & D to shorten the research cycle and to reduce the expenses. Among them, computational approaches have revolutionized the pipeline of discovery and development (2). In the last 40 years, computational technologies for drug R & D have evolved very quickly, especially in recent decades with the unprecedented development of biology, biomedicine, and computer capabilities. In the post-genomic era, because of the dramatic increase of small-molecule and biomacromolecule information, computational tools have been applied in almost every stage of drug R & D which has greatly changed the strategy and pipeline for drug discovery (2). Computational approaches span almost all stages in discovery and development pipeline, from target identification to lead discovery, from lead optimization to preclinical or clinical trials (see Fig. 1). In this review article, we highlight some recent advances of computational technologies for drug discovery and development; emphases are put on computational tools for target identification, lead discovery, and ADME/T (absorption, distribution, metabolism, excretion, and toxicity) prediction.

Target Identification

Target identification and validation is the first key stage in the drug-discovery pipeline (see Fig. 1). By 2000, only about 300 drug targets had been reported (3, 4). The completion of human genome project and numerous pathogen genomes unveils that there are 30,000 to 40,000 genes and at least the same number of proteins; many of these proteins are potential targets for drug discovery. However, identification and validation of druggable targets from thousands of candidate macromolecules is still a challenging task (5).

Numerous technologies for identifying targets have been developed recently. Experimental approaches such as genomic and proteomic techniques are the major tools for target identification (6, 7). However, these methods have been proved inefficient in target discovery because they are laborious and time consuming. In addition, it is extremely difficult to get relatively clear information related to drug targets from enormous information produced by genomics, expression profiling, and proteomics (5). As complementarities to the experimental methods, a series of computational (in silico) tools have also been developed for target identification in recent two decades (6, 8-12). In general, they can be categorized into sequence-based approach and structure-based approach. Sequence-based approach contributes to the processes of the target identification by providing functional information about target candidates and positioning information to biological networks. Such methods include sequence alignment for gene selection, prioritization of protein families, gene and protein annotation, and expression data analysis for microarray or gene chip. Here, we only introduce a special structure-based method for target identification—reverse docking.

Molecular docking has been a promising approach in lead discovery (2, 13) (also see discussion below). Reverse docking,
Deformylase (PDF) is a target for anti-H. pylori drugs. PDF is a potential target for screening new anti-H. pylori agents. In a similar approach, Paul et al. (15) successfully recovered the corresponding targets of four unrelated drugs with reverse docking method.

The advantage of reverse docking is obvious, in addition to identifying target candidates for active compounds, it is also possible to identify potential targets responsible for toxicity and/or side effects of a drug under the hypothesis that the target database contains all the possible targets (20). However, reverse docking still has certain limitations. The major one is that the protein entries in the proteins structure databases like PDB are not enough for covering all the protein information of disease related genomes. The second one is that this approach has not considered the flexibility of proteins during docking simulation. These two aspects will produce false negatives. Another limitation is that the scoring function for reverse docking is not accurate enough, which will produce false positives (16). One tendency to overcome these shortages is to develop new docking program including protein flexibility and accurate scoring function. Another tendency is to integrate sequence-based and structure-based approaches together.
Lead Discovery

Computational approaches for lead discovery are more mature than those for target identification. Some methods such as virtual screening and library design have become promising tools for lead discovery and optimization (2, 13). Here, we introduce two computational techniques, virtual screening, and focused combinatorial library design.

Virtual screening

The drugs developed in the past 100 years are found to interact with approximately 500 targets; in the same period, about 20,000,000 organic compounds including natural products have been synthesized or isolated. Moreover, the genomic and functional genomic projects have produced additional 1500 druggable targets for drug intervention to control human diseases (21). Therefore, it is believable that a large number of new drugs, at least many leads or hits, hide in the existing chemical databases, which contain the pharmacophore elements and may conform to the pharmacophore feature is constructed by structure-activity relationship analysis on a series of active compounds (26) or is deduced from the X-ray crystal structure of a ligand-receptor complex (27). Taking this 3-D–pharmacophore feature as a query structure, 3-D database search can be performed to select the molecules among a large number of candidate molecules likely to be active against a chosen biological receptor (23). Therefore, how to dig out this source is a hard task. Collecting all the existing compounds and screening them randomly against all the potential targets one by one are extremely impractical, because it is intolerably expensive and time consuming. However, virtual screening shows a potential to satisfy this requirement. Indeed, virtual screening has been involved into the pipeline of drug discovery as a practical tool (22).

In essence, virtual screening is designed for searching large-scale hypothetical databases of chemical structures or virtual libraries by using computational analysis for selecting a limited number of candidate molecules likely to be active against a known biological receptor (23). Therefore, virtual screening is a logical extension of 3-D pharmacophore-based database searching (PBDS) (24) or molecular docking (25), which is capable of automatically evaluating very large databases of compounds.

Two strategies have been used in virtual screening (see Fig. 3); 1) using PBDS to identify potential hits from the databases, mostly in the cases that 3-D structures of the targets are unknown and 2) using molecular docking approach to rank the databases if the 3-D structures of the targets are available.

Normally, these two approaches are used concurrently or sequentially, because the former can filter out the compounds quickly and the latter can evaluate the ligand-receptor binding more accurately.

For the pharmacophore-based screening, a 3-D-pharmacophore feature is constructed by structure-activity relationship analysis on a series of active compounds (26) or is deduced from the X-ray crystal structure of a ligand-receptor complex (27). Taking this 3-D–pharmacophore feature as a query structure, 3-D database search can be performed to select the molecules from the available chemical databases, which contain the pharmacophore elements and may conform to the pharmacophore geometric constraints. Then the selected compounds are obtained either from commercial sources or from organic synthesis for the real pharmacologic assays (see Fig. 3).

Docking-based virtual screening (DBVS) requires the structural information of both receptors and compounds. The general procedure of DBVS includes five steps: receptor modeling (virtual screening mode construction), compound database generation, computer screening, hit molecules postprocessing, and experimental bioassay. The core step of virtual screening is docking and scoring. Docking is a process to place each molecule from a 3-D small-molecule database into the binding site of a receptor protein, optimize the relative orientation and conformation for a ligand interacting with a protein, and select molecules from the database that may bind to the protein tightly. Billions of possible conformations can be created for a flexible ligand even with a few freedom degrees in the rotation space alone; current optimization algorithms cannot sample exhaustively for all conformations and orientations, needless to say to account for the flexibility of protein with thousands of degrees of freedom. Therefore, instead of searching exhaustively in the search space, an optimization algorithm should sample the solutions effectively and rapidly close to the global optimum (28). Basically, the optimization algorithms employed widely by molecular docking program can be divided into three categories: numerical optimization methods, random or stochastic methods, and hybrid optimization methods. Essentially, there are three types of scoring functions (29), which

\[
\text{IC}_{50} \text{ values)
\]

\[
\text{IC}_{50} \text{ values)
\]
include force-field based scoring functions, empirical scoring functions, and knowledge-based scoring functions. However, no single scoring function can perform satisfactorily for every system because many physical phenomena determining molecular recognition were not fully accounted for, such as entropic or solvent effects (13). Since Kuntz et al. (30) published the first docking algorithm DOCK in 1982, more than 20 docking programs have been developed in recent two decades. Of the existing docking programs, DOCK, FlexX, AutoDock, and GOLD are most frequently used, other programs such as Glide, ICM, and Surfex have been applied successfully in virtual screening. However, many limitations and challenges still exist for molecular docking, such as accurate prediction of the binding conformation and affinity, protein flexibility, entropy, and solvent effects. Without doubt, the molecular docking process is still a complex and challenging project of computational chemistry and biology (25, 31).

Virtual screening has discovered numerous active compounds and leads, more than 50 compounds have entered into clinical trials, and some have been approved as drugs (2). Virtual screening enriched the hit rate (defined as the quotient in percentage of the number of active compounds at a particular concentration divided by the number of all compounds experimentally tested) by about 100-fold to 1000-fold over random screening (22, 32). Additionally, virtual screening provides an alternative way in firstly finding new leads of some targets, whereas the techniques for high-throughput screening remain in development (e.g. potassium ion (K⁺) channels). By using docking-based virtual screening in conjunction with electrophysiological assay, we discovered 10 new blockers of the eukaryotic Shaker K⁺ channels, 4 natural product blockers (33), and 6 synthetic compounds (34).

**Focused combinatorial library design**

What big pharma and medicinal chemists are seeking is new chemical entities (NCEs), which can be strictly protected by compound patents. Nevertheless, all hits discovered by screening the existing compounds databases can only find the new medical usages for the existing compounds or old drugs. At least two kinds of computational approaches are available for NCE discovery: de novo drug design (35, 36) and combinatorial library design (37). The de novo drug design does not start from a database of complete molecules but aims at building a complete molecule from molecular bricks (“building blocks”) to fill the binding sites of target molecule chemically. The complete chemical entries could be constructed through linking the “building blocks” together, or by growing from an “embryo” molecule with the guidance of evaluation of binding affinity. The “building blocks” could be either atoms or fragments (functional groups or small molecules). However, using atoms as “building blocks” is thought to be inefficient, which is seldom used currently. In the fragment-linking approach, the binding site is mapped to identify the possible anchor points for functional groups. Then, these groups are linked together to form a complete molecule. In the sequential-growing approach, the molecule grows in the binding site controlled by an appropriate search algorithm, which evaluates each growing possibility with a scoring function. Different from docking-based virtual screening, fragment-based de novo design can perform sampling in whole compound space, which can obtain novel structures that are not limited in available databases. Nevertheless, the quality of a growing step strongly depends on previous steps. Any step chemically growing wrongly would lead to an unacceptable result. For the fragment linking approach, it remains a
demonstrated the efficiency of the strategy for focused library design and screening. In addition, the novel chemical entities reported in this study could be leads for discovering new therapies against the CypA pathway.

ADME/T Optimization

It has widely been appreciated that ADME/T should be involved in the early stage of drug discovery (lead discovery) in parallel with efficacy screening (40, 41). Many high-throughput technologies for ADME/T evaluation have been developed, and the accumulated data make it possible to construct models for predicting the ADME/T properties of compounds before structural modifications. These computational models may reduce the redesign-synthesize-test cycles. Hereinafter, we will show how computational approaches predict and model the most relevant pharmacokinetic, metabolic and toxicity endpoints, which thereby accelerate the pace of new drug development.

Prediction of intestinal absorption

Absorption of drugs from the gastrointestinal tract is a complex process that can be influenced by not only physicochemical and physiological factors but also by formulation factors. Some published models and commercially available computer programs have been developed for predicting the intestinal absorption. Among them, physiologically based approaches are of particular interest because the human physiological environment and the processes involved in drug absorption are simulated.

Mixing tank models describe the intestine as one or more well-mixed tanks, in which dissolved and solid forms of drug have a uniform concentration and are transferred between the mixing tanks by first-order transit kinetics. This model has been implemented in commercial program Intellipharm PK (http://www.intellipharm.com) and OraSpotter (http://www.akosgmbh.de). Tube models assume the intestine to be a cylindrical tube with or without changing radius, which includes the mass balance models and the models used in PK-Map and PK-Sim (http://www.bayertechnology.com). In the mass balance model, the difference between the mass flows in and out of the tube represents the mass absorbed per time unit, and it is proportional to the permeability and the concentration of the drug in the intestines. In models of PK-Map and PK-Sim, a transit function was introduced and the intestinal permeability coefficient can be calculated solely by the lipophilicity and molecular weight of the investigated compound.

Yu et al. (42) developed an absorption and transit (CAT) model, which formed a basis of many models and commercially available computer programs, such as GITA (GI-Transit Absorption), iDEA (http://www.akosgmbh.de) and GastroPlus (http://www.simulations-plus.com). iDEA uses an absorption model proposed by Grass et al. (43), and GastroPlus uses the ACAT (advanced compartmental absorption and transit) model developed by Agoram et al. (44). Parrott et al. (45) compared these two commercial programs both in their ability to predict fraction absorbed for a set of 28 drugs and in terms of the functionality offered. The results suggested that iDEA may perform better with measured input data, whereas GastroPlus presents a...
more sophisticated user interface, which shows strengths in its ability to integrate additional data.

**Prediction of blood–brain barrier (BBB) permeability**

The blood–brain barrier (BBB) is a physiological barrier in the circulatory system that restrains substances from entering into the central nervous system (CNS). The orally administered drugs that target receptors and enzymes in the CNS must cross the BBB to produce desired therapeutic effects. At the same time, the peripherally acting agents should not cross it to avoid the side effects (46). To date, various methods that have been applied for model generation include the traditional statistical approaches and state-of-art machine-learning techniques.

Generally, current BBB models have shown acceptable capabilities depending on the approaches used and particularly the type and the size of the dataset under investigation. Some insights into the molecular properties that determine the brain permeation have also been gained. However, most current in silico modeling of BBB permeability start with the assumption that most drugs are transported across the BBB by passive diffusion. Absence of data regarding active transport or P-glycoprotein (P-gp) efflux limits the development of accurate and reliable BBB models that can more closely mimic the in vivo situation. Recently, Adeno et al. (47) reported their studies based on a large heterogeneous dataset (~1700 compounds) including 91 P-gp substrates. Likewise, Garg and Verma (48) took the active transport of the molecules into consideration in the form of their probabilities of becoming a substrate to P-gp, and their results indicated that the P-gp indeed plays a role in BBB permeability for some molecules.

**Prediction of CYP-mediated metabolism**

The human cytochromes P450 (CYP450) constitute a large family of heme enzymes. In drug development, the CYP-mediated metabolism is of particular interest because it may profoundly affect the initial bioavailability, desired activity, and safety profile of compounds. A great variety of in silico modeling approaches have been applied to CYP enzymes. Connections and integrations between the “ligand-based,” “protein-based,” and “ligand-protein interaction–based” groups of methods have been extensively described in the review of Graaf et al. (49). Apart from this methodological categorization, current in silico models mainly address the following three aspects of metabolism: CYP inhibition, isoform specificity, and site of CYP-mediated metabolism.

Inhibition of CYP enzymes is unwanted because of the risk of severe side effects caused by drug–drug interactions. A novel Line-Walking Recursive Partitioning method was presented that uses only nine chemical properties of the shape, polarizability, and charge of the molecule to classify a compound as a 3A4, 2C9, or 2D6 inhibitor (50). Recently, Jensen et al. (51) constructed Gaussian kernel weighted k-NN models to predict CYP2D6 and CYP3A4 inhibition, resulting in classification models with over 80% of overall accuracy.

CYP isoform specificity predicts the subtype responsible for the main route of metabolism. This type of study started from the work by Manga et al. (52), in which the isoform specificity for CYP3A4, CYP2D6, and CYP2C9 substrates was investigated. Terfloth et al. (53) reported a similar study on the same dataset using ligand-based methods, and the accuracy of prediction was improved substantially.

The site of metabolism (SOM) refers to the place in a molecule where the metabolic reaction occurs, which is usually a starting point in metabolic pathway investigation. Recently, Sheridan et al. (54) reported a study addressing the SOM mediated by CYP3A4, CYP2C9, and CYP2D6. In this study, pure QSAR-model was constructed with only substructure and physical property descriptors. The cross-validated accuracies range from 72% to 77% for the investigated datasets, which are significantly higher than that of earlier models and seem to be comparable with the results from MetaSite (http://www.moldiscovery.com), which is a more mechanism-based method of predicting SOM.
Prediction of toxicity

At present, a variety of toxicological tests needs to be conducted by the drug regulatory authorities for safety assessment. Computational techniques are fast and cheap alternatives to biochemical as they require neither experimental materials nor physically available compounds. Some examples of available computer programs that predict toxicity are Deductive Estimation on Risk from Existing Knowledge (http://www.thesalolimited.org), Multi-Purpose Computer Automated Structure Evaluation (http://www. molsoft.com), and Toxicity Prediction by Computer Assisted Technology (current DIS TOPKAT, http://accelrys.com/products/discovery-studio/toxicology). Recently, developments in artificial intelligence research and the improvement of computational resources have led to efficient data-mining methods that can automatically extract structure toxicity relations (STRs) from toxicity databases with structurally diverse compounds. Helma et al. (55) described some general procedures used for generating QSTR models from experimental data. Here, we briefly present a study made to model the mutagenic probability (56).

Mutagenity is the ability of a compound to cause mutations in DNA, which is one of the toxicological liabilities closely evaluated in drug discovery and their toxic mechanisms have been relatively well understood. In the development of QSTR models, it is essential to select the structural or chemical properties most relevant to the endpoint of interest. For mutagenicity, we designed a molecular electrophilicity vector (MEV) to depict the electrophilicity of chemicals, which is a major cause of chemical/DNA interaction. A model for the classification of a chemical compound into either mutagen or nonmutagen was then obtained by a support vector machine together with a F-score based feature selection. For model construction, the SVM-c-MEV method shows its superior efficiency in the data fitting, with a concordance rate of 91.86%. For validation, a prediction accuracy of 84.80% for the external test set was yielded, which is close to the experimental reproducibility of the Salmonella assay. The F-score based feature weighting analyses highlight the important role of atomic partial charges in describing the noncovalent intermolecular interactions with DNA, which is usually a hard part in the mutagenicity prediction.

The Future

The technological progress of computational chemistry and biology brought a paradigm change to both pharma and research institutes. Computational tools have been considerably appreciated in drug discovery and development, which play increasingly important roles in target identification, lead discovery, and ADME/T prediction. In the future, in addition to improving individual existing computational techniques, such as perfection of the accuracy and effectiveness of virtual screening, one major tendency is to integrate computational chemistry and biology together with chemoinformatics and bioinformatics. This research is leading to a new topic known as pharmacoinformatics, which will impact the pharmaceutical development process and increase the success rate of development candidates (57).

Another tendency is to use more accurate computational approaches, such as molecular dynamics simulation, to design specific inhibitors or activators against the pathway of target protein folding (58). By using this strategy, Broglio et al. (59) obtained inhibitors that may block HIV-1 protease monomer to fold into its native structure, which causes the inhibitors not to create resistance. Similarly, by targeting the C-terminal β-sheet region of an Aβ intermediate structure extracted from molecular dynamics simulations of Aβ conformational transition (60), we obtained a new inhibitor that abolishes Aβ fibrillation using virtual screening in conjunction with thioflavin T fluorescence assay and atomic force microscopy determination (61).

After the completion of the human genome and numerous pathogen genomes, efforts are underway to understand the role of gene products in biological pathways and human diseases and to exploit their functions for the sake of discovering new drug targets (62). Small and cell-permeable chemical ligands are used increasingly in genomics approaches to understand the global functions of genome and proteome. This approach is referred to as chemical biology (or chemogenomics) (63). Another emerging field is computational systems biology to model or simulate intracellular and intercellular events using data gathered from genomic, proteomic, or metabolomic experiments (64). This field highly impacts drug discovery and development.

References


Further Reading


LOTUS: http://www.dddt.ac.cn/lotus


TarFisDock: http://www.dddc.ac.cn/tarfisdock


See Also

ADME Properties of Drugs

Computational Chemistry in Biology

Lead Optimization in Drug Discovery
Pharmacokinetics of Drug Candidates, Determination of

Ronald E. White, Schering-Plough Research Institute, Kenilworth, New Jersey

doi: 10.1002/9780470048672.wecb644

Pharmacokinetics is a mathematically based discipline that describes the time course of uptake, distribution, and elimination of a drug in an organism. This article explains the physiologic basis and clinical interpretation of the various pharmacokinetic parameters, such as half-life, clearance, and oral bioavailability. Because of the central importance of pharmacokinetics to the clinical use of drugs, the determination of pharmacokinetics is required at several points in the research and development cycle of new drug candidates. The rationale and methods of determination are discussed.

The therapeutic effectiveness of drugs generally can be related to the concentrations achieved at the physiologic site of action and to the length of time that effective concentrations are maintained (1, p. 1). For instance, molecules of an antibiotic must reach the site of infection (e.g., inner ear, lung, or bladder) to interact with the molecular target (typically an enzyme) in the infecting bacteria. Similarly, cough suppression or relief of clinical depression may require a medicine to penetrate the blood-brain barrier to access molecular targets (typically pharmacologic receptors) in the central nervous system. Therefore, an assessment of the effectiveness of the therapy partly depends on our ability to monitor concentrations of the drug in these organs. However, because often it is not feasible to measure intraorgan concentrations of drugs directly, we usually measure blood concentrations and then infer the corresponding tissue concentrations by using predetermined relationships for the particular drug and organ in question. In practice, we establish the blood concentration that corresponds to the actual tissue concentration necessary for therapeutic effectiveness for each clinical drug, which allows us to discuss therapy in terms of blood concentrations only without reference to tissue levels. A useful clinical drug must exhibit several practical characteristics (2):

- **Efficacy**: the intrinsic ability of the compound to produce a desired pharmacologic effect.
- **Availability**: the ability of the compound to reach the target organ.
- **Safety**: the sufficient selectivity at the therapeutic dose so that undesirable pharmacologic actions are acceptably mild.
- **Persistence**: the sufficient residence time in the body to allow a clinically useful duration of action, usually expressed as the plasma elimination half-life.

Pharmacokinetics (PK) is a mathematically based scientific discipline that describes the time course of uptake, distribution, and elimination of a drug in an organism. A central goal of PK is to determine the length of time that target pharmaceutical receptors are exposed to pharmacologically effective concentrations of the drug molecules. Therefore, PK is of the utmost importance in understanding drug action, especially the duration of action, optimal dose size, individual variation in response, interactions between drugs, and developed resistance or tolerance to drugs.

**Clinical Pharmacokinetics**

The following discussion addresses only small-molecule drugs, chemical compounds with molecular weights less than 1,000 Daltons. Large-molecule drugs (biologics) such as vaccines, cytokines, antibodies, and genes also can be studied by the methods of pharmacokinetics, but they are distributed and cleared in fundamentally different ways than small molecules. (3, 4).

**Dose and administration**

Dose refers to the amount of the drug to be administered to the patient and may be expressed in several ways. The most common form is the amount (expressed in milligrams) taken in an oral formulation such as a tablet, capsule, or syrup. Often, the formulation is provided by the manufacturer in several dose strengths, such as 10, 25, and 50 mg. Some medicines are delivered by a parenteral route such as inhalation, topical, subcutaneous, or intravenous to deliver the medicine more directly to the target organ, to control the rate of delivery, or to substitute for when oral dosing is impractical. Depending on the rate of elimination of the drug from the body, additional doses are administered at intervals to maintain adequate therapeutic.
blood levels. Because the presence of food in the gastrointestinal tract may affect the absorption of some drugs, the patient may be instructed to take the drug either with or between meals. Thus, a dosing regimen must be designed to achieve a particular concentration over a particular interval or time of day in relation to meals, such as a 10-mg tablet taken once a day after dinner.

**Meaning and interpretation of pharmacokinetic parameters**

Plasma, the fluid component of blood left after the separation of cells, commonly is used for pharmacokinetic determinations. Figure 1 illustrates the time course of drug concentrations in plasma following either an intravenous (iv) or an oral (po) dose of a hypothetical drug. In the case of the iv bolus dose, the drug is delivered completely to the bloodstream essentially instantaneously and then concentrations immediately begin to decline as the drug distributes within and is eliminated from the body (see the dashed curve in Fig. 1). The solid curve shows the drug concentration profile for the same dose given orally. In this case, because the drug first must be absorbed from the gastrointestinal tract to reach the blood, concentrations initially rise (absorption phase); however, when the rate of absorption equals the rate of elimination, the input and output processes reach a balance and the peak concentration (Cmax) is seen. Thereafter, when absorption is complete and only the elimination process is left, the elimination phase is reached and concentrations decline as the body acts to remove the drug.

During the absorption phase, the plasma concentration attains the minimum therapeutic threshold accompanied by onset of the efficacious effect. A desirable medicine has a large window between the efficacious concentration and a concentration that produces undesirable side effects (adverse event threshold). As concentrations fall after Cmax, the duration of effect is determined by the time above the therapeutic threshold. Thus, the illustrated drug loses effect after about 12 hours and requires a second dose by that time.

**Figure 1** Plasma concentration time profiles for a single 42-mg dose of a hypothetical drug with a half-life of 8 hours, Vd 42 liters, and oral bioavailability of 80%. Dashed curve: intravenous bolus dose. Solid curve: oral dose. Dotted horizontal lines represent plasma concentrations required for efficacy (green) and for the onset of adverse events (red).
A successful dosing regimen will keep $C_{\min}$ accumulation index of these two is called the time points (Fig. 4 trapezoids formed between successive measured concentration to approach the adverse event threshold.

state (Cave plasma concentrations over a single dosing interval at steady state (Fig. 3 each successive dose (Fig. 3). The time-weighted average of plasma concentrations over a single dosing interval at steady state ($C_{\text{SS}}$, SS) is higher than that on the first dose, and the ratio of these two is called the Accumulation Index (R) (1, p. 87). A successful dosing regimen will keep $C_{\text{SS}}$, SS well above the therapeutic threshold at all times while never allowing $C_{\text{SS}}$, SS to approach the adverse event threshold.

$AUC$ is a measure of the total exposure of the body to the drug and can be estimated as the sum of the areas of a series of trapezoids formed between successive measured concentration time points (Fig. 4 and Equation 1) (1, p. 469). Because a drug may not necessarily be completely absorbed or may undergo presystemic metabolism or other elimination during the absorption process, the related parameter $F$ indicates the fraction of an orally administered dose that actually reaches the systemic circulation. As a practical matter, $F$, which usually is expressed as a percentage, is calculated as the ratio of the $AUC$ measured from an oral dose to the $AUC$ from an equivalent intravenous dose that is delivered completely to the bloodstream (Equation 2) (1, p. 42). Because two different drugs generally have different $C_{\text{max}}$, values, bioavailabilities, and half-lives, the observed exposures (i.e., $AUC$s) will be different even though equal doses of the drugs may have been given.

$AUC = \sum_{i=1}^{n} \frac{1}{2} (C_i + C_{i-1}) \times (t_i - t_{i-1})$ (1)

$F = \frac{AUC_{\text{oral}}}{AUC_{\text{iv}}}$ (2)

$V_d$ measures the volume of plasma in which the drug appears to be dissolved and is calculated by dividing the amount of drug in the body by the observed concentration in plasma at that time (Equation 3) (1, p. 20). For a drug that is confined to the plasma compartment, $V_d$ is the volume of plasma in the body (ca. 3 liters (L) in humans) (6). A drug that freely passes into and out of the cells of tissues will have a $V_d$ value of about 42 L (i.e., nominal total body water for an adult human) (6). However, many drugs show $V_d$ values that are much greater than 42 L, which cannot correspond to any physiologic compartment. In these cases, the drug must be sequestered significantly outside of the plasma compartment into organs and tissues. For this reason, $V_d$ sometimes is called the apparent volume of distribution. Because it is difficult to know at any moment the exact amount of drug in the body because of continuous elimination, we usually cannot apply Equation 3 directly to determine $V_d$ and we must use an indirect method (see below).

$V_d = \frac{\text{amount \ in \ plasma}}{C_{\text{plasma}}}$ (3)

Finally, $CL$ describes how quickly the drug is eliminated completely from plasma or blood and has the units of volume per unit time. $CL$ is calculated easily from the plasma $AUC$ observed for a given dose (Equation 4) (1, p. 37). For an intravenous dose $F = 1$, Equation 4 simplifies to $CL = \text{Dose/AUC}_{\text{iv}}$. For an oral dose, if the bioavailability is not known, then the $Dose/AUC$ calculation yields $CL/F$ rather than the true clearance.

$CL = \frac{\text{Dose}_{\text{oral}}}{\text{AUC}_{\text{oral}}}$ $F = \frac{\text{Dose}_{\text{iv}}}{\text{AUC}_{\text{iv}}}$ (4)

$CL$ is a useful working concept because it is related to the passage of drug-containing blood through a clearance organ such as liver or kidney. A drug that is cleared only by the liver (by liver metabolism, for instance) can be cleared only as fast as blood flows to the liver. This mass transport limitation is expressed in Equation 5 (1, p. 166), where $Q$ is the total blood flow to the liver (about 1.4 L per minute (min)) (1, p. 138), $CL_{\text{int}}$ is the intrinsic ability of the liver to clear the drug if blood flow were not a limitation, and $f_l$ is the fraction of the drug not bound to proteins in plasma (7). We can see from Equation 5 that in the two extreme cases, $CL$ is equal to...
elimination, instance, for a drug that has both hepatic and renal routes of clearance is the sum of the individual organ clearances. For clearance is that it is additive. In other words, the total systemic from urine back into blood. A final useful characteristic of concentration time profile is treated descriptively by the method of noncompartmental analysis, persistence in the body is described by a new parameter, mean residence time (MAT, the average time that an drug molecule resides in the body) rather than terminal elimination half-life. For many drugs, MRT is a better indicator of the clinically effective duration. The difference between MRT values for oral and intravenous doses gives a descriptor of the rate of intestinal absorption called the mean absorption time (MAT). Noncompartmental analysis also defines a new volume parameter, \( V_{\text{ss}} \), the volume of distribution at steady state, which is useful for calculations of plasma levels after multiple dosing or during intravenous infusion. A second descriptor of volume, \( V_{\text{s}} \), represents the volume of distribution during the terminal elimination phase and is calculated exactly the same as \( V_{\text{ss}} \) in the one-compartment model (see Equation 6). \( V_{\text{s}} \) usually is different from \( V_{\text{ss}} \) and seldom is used in clinical dosing regimen considerations.

**Role of PK in the Drug Discovery and Development Process**

**Discovery**

Basic research elucidates the complex biochemical events that comprise a biologic process, such as the regulation of blood glucose, and identifies key control points that are mediated by enzymes or receptors. Most drugs exert their pharmacologic action by modulating the activity of one of these enzymes or receptors within cells of the abnormally functioning organ. So, the next step in the drug discovery process is lead generation, the design or discovery of a small molecule that will bind to the molecular target to modulate its activity. Once a lead has been generated, medicinal chemists typically synthesize thousands of analogs by systematically varying the structure to create a compound that has been optimized with respect to potency and selectivity toward the molecular target. PK characteristics, and

---

**Equations**

1. When \( CL_{\text{int}} \) is small compared with \( Q \), and \( CL \) is equal to \( Q \) when \( CL_{\text{int}} \) is large compared with \( Q \):

   \[
   CL = \frac{Q \times f_u \times CL_{\text{int}}}{Q + f_u \times CL_{\text{int}}} \tag{5}
   \]

2. Similarly, renal clearance is limited ultimately by kidney blood flow (ca. 1,100 mL/min) (1, p. 138). Clearance by kidneys comprises three processes: glomerular filtration, the rate at which the kidneys filter plasma (ca. 120 mL/min) (1, p. 169); direct secretion of drugs into urine; and reabsorption of drugs from urine back into blood. A final useful characteristic of clearance is that it is additive. In other words, the total systemic clearance is the sum of the individual organ clearances. For instance, for a drug that has both hepatic and renal routes of elimination, \( CL_{\text{tot}} = CL_h + CL_r \). Such a drug still can be cleared, albeit more slowly, in patients with either liver or kidney failure.

   For an ideal, so-called one-compartment drug (i.e., one that is in rapid equilibrium with all tissues), the parameters half-life, clearance, and volume of distribution are interrelated by Equation 6 (1, p. 24). Because half-life and clearance can be determined independently, it is possible to calculate \( V_{\text{ss}} \):

   \[
   CL = \frac{Q \times f_u \times CL_{\text{int}}}{Q + f_u \times CL_{\text{int}}} \times 0.693 \times V_{\text{ss}} \tag{6}
   \]

   As mentioned above, many drugs do not conform to the simple one-compartment model. These cases may require a two- or three-compartment model characterized by a bi- or tri-exponential decline (8). Alternatively, a simpler, commonly used approach is noncompartmental analysis, in which the concentration time profile is treated descriptively by the method of statistical moments (9). Whereas CL has exactly the same definition in noncompartmental analysis, persistence in the body is described by a new parameter, mean residence time (MAT, the average time that a drug molecule resides in the body) rather than terminal elimination half-life. For many drugs, MRT is a better indicator of the clinically effective duration. The difference between MRT values for oral and intravenous doses gives a descriptor of the rate of intestinal absorption called the mean absorption time (MAT). Noncompartmental analysis also defines a new volume parameter, \( V_{\text{ss}} \), the volume of distribution at steady state, which is useful for calculations of plasma levels after multiple dosing or during intravenous infusion. A second descriptor of volume, \( V_{\text{s}} \), represents the volume of distribution during the terminal elimination phase and is calculated exactly the same as \( V_{\text{ss}} \) in the one-compartment model (see Equation 6). \( V_{\text{s}} \) usually is different from \( V_{\text{ss}} \) and seldom is used in clinical dosing regimen considerations.

---

**Figure 4**

Estimation of the total AUC for an oral dose of the drug by use of the Trapezoidal Rule. Vertical lines from measured concentration time points define a family of trapezoids, the sum of the areas of these trapezoids approximates the area under the curve. The area beyond the last time point (24 hour) is estimated by multiplying the last measured concentration by the factor \( 0.693 \times T_{1/2} \). Area (24-hour) = \( C_{\text{final}} \times T_{1/2} \).
safety. This compound then is designated as a clinical candidate, and it enters the development process (10).

Given the importance of PK in the clinical use of drugs, a suitable human PK profile is an important criterion in the selection from many discovery compounds of a single candidate for advancement to clinical development. Therefore, contemporary drug discovery screening of compounds includes pharmacologic potency, selectivity, and PK properties in the acceptable range for the medical indication (11). Because drug discovery normally takes place before the clinical phase, the human PK characteristics of a drug candidate must be inferred from in vitro animal models or from various human-derived in vitro systems (12).

**PK screening**

**In vitro studies**

Because clearance at the whole-body level often is determined by metabolism at the cellular level, it is possible to use a variety of human-derived in vitro systems to determine rates of metabolism. These systems include pure human enzymes (such as cytochrome P450 enzymes) (13) and human liver subcellular fractions (microsomes) (14). However, with enzymes and subcellular fractions, some information is lost because the whole-cell integration of subcellular processes has been disrupted. The use of cultured human hepatocytes retains the whole-cell integration at the expense of greater experimental complexity (15). Each system provides a different window on the metabolic processes, is relatively easy to use, and can be obtained from commercial sources. Rates and pathways of metabolism may be compared with a series of discovery candidates to identify those with the greatest relative metabolic stability or with a benchmark compound of known human PK characteristics to provide a more absolute estimate of hepatic metabolic clearance.

**In vivo studies**

PK studies in typical laboratory animals (typically mice, rats, dogs, or monkeys) are useful because they directly determine the various PK parameters discussed above, which affords an understanding of the whole-body characteristics of absorption, distribution, metabolism, and elimination. Because these translate with some fidelity to humans, animal PK commonly is used to assess the PK acceptability of discovery compounds. Unfortunately, these studies tend to be too slow to permit rapid evaluation of dozens or hundreds of discovery compounds. Therefore, two modifications have been introduced that provide the PK acceptability of discovery compounds. First, if the compound was cleared mainly by hepatic metabolism in the animal species tested and if human hepatocytes in vitro suggest the same will be true in humans, then the measured hepatocyte clearance may be used in a process called in vitro/in vivo scaling (20, pp. 207–228) to provide an estimate of the human intrinsic clearance. The application of Equation 5 then gives an estimate of the human systemic clearance. Second, the animal PK parameters of CL and Vd can be subjected to allometric scaling (20, pp. 207–228) whereby the PK parameter is related to a measurable allometric variable such as body mass, body surface area, heart rate, and so forth (21) by fitting these parameter-variable pairs for several species to an empirical power equation of the form

\[ \text{PK parameter} = \alpha \times \text{(allometric variable)}^\beta \]  

where \( \alpha \) and \( \beta \) are adjustable constants. A common application is to scale CL according to body mass.

\[ \text{CL} = \alpha \times \text{(Body Mass)}^\beta \]  

Both CL and Vd from the animal PK studies may be scaled allometrically to provide the estimates of human CL and Vd needed to project the human half-life. In addition to calculating the half-life, one also can estimate the human therapeutic dose if the blood levels that correspond to efficacy in a relevant animal pharmacologic model have been measured. The AUC measured in the animal model can be equated with the AUC needed for efficacy in humans, and then the application of Equation 4 with the estimated human CL gives an estimated human dose.

For drugs intended for oral administration, the bioavailability \( F \) often is assumed to lie in the range between the lowest and highest values observed in the animal species tested. These projected human PK parameters need not be extremely accurate to be useful for planning.

**Preclinical development**

Before a new chemical entity can be tested in humans, its safety in various in vitro and in vivo pharmacologic and toxicity tests is
must be assessed (22). This assessment includes an administration of high doses to animals (typically mice or rats and larger animals such as dogs or monkeys) and in vitro tests of genotoxic potential (such as the Ames test). PK, normally will be determined in the animal species to be used for safety testing to assess the relationship between the oral dose and the exposure for dose selection and to select time points for the monitoring of exposure during the safety studies. In parallel with the safety testing, a large-scale chemical synthetic process must be developed to produce clinical supplies of the drug candidate to enable clinical trials. Finally, a clinical formulation must be devised to allow adequate and reproducible exposure in the clinical trials, and animal PK studies are helpful in the selection process. These studies usually examine Cmax, Tmax, and AUC for several possible formulations to allow matching to the desired clinical delivery profile. Data from the discovery and preclinical activities form the basis of a petition to government health authorities for permission to begin testing in humans. Once this dossier, called the Investigational New Drug Application (IND) in the United States, is approved, clinical trials may proceed (23). Information about the IND and other regulatory guidelines and documentation is available through the web link to the United States Food and Drug Administration at the end of this article.

**Clinical development**

Clinical trials are divided formally into four distinct but temporarily overlapping phases, defined by the objectives of each phase (24). Phase I tests the safety and tolerability of the drug candidate in normal healthy volunteer subjects. An important goal is to determine the maximum dose that may be safely given to establish the dose–exposure relationship and also the half-life. Phase II investigates whether the candidate modulates the pharmacologic target in ten to twenty patients (Phase IIb) using PK parameters, which are needed to plan the next phases. Phase IIc tests the drug in much larger groups of patients to establish the comprehensive clinical profile of the candidate, including confirming the therapeutic benefit with adequate statistics and investigating the variation of therapeutic response and the safety in various severities of the disease and in special populations such as ethnic groups or comorbidities. Phase IV usually occurs after the drug has been approved and is on the market and involves testing even larger populations for longer periods of time and testing against additional disease indications.

**Phase I**

Human PK is determined for the first time in Phase I, initially as a series of single doses escalated through several dose levels to establish the dose-exposure relationship and also the half-life. Because an oral dose may have an extended period of absorption, it is not possible to get the true clearance or volume of distribution from an oral dose; a confounding with bioavailability may occur. Furthermore, sometimes absorption is slower than elimination and a phenomenon known as “flip-flop kinetics” occurs in which the observed half-life really is a reflection of absorption and not elimination. Thus, when feasible, an intravenous dose is given to allow absolute oral bioavailability and the true values of total clearance, renal clearance, volume of distribution, and half-life to be determined. Administration of an intravenous dose requires the development of a sterile solution formulation and acute intravenous animal safety studies and may not be possible always with poorly soluble drugs.

The next step is multiple dosing at several dose levels within the range determined to be safe from the single-dose study. In addition to more investigation of clinical safety, the main PK purpose of multiple dosing is to establish how many doses are required to reach steady state and what steady state Cmax and AUC values are reached. A well-behaved drug will reach steady state in about five doses (see Fig. 3), and a general rule is that if doses are given at intervals of one half-life, then the steady state Cmax will be about twice that following a single dose (25, p. 123). Dosing more frequently than the half-life results in a more than twofold accumulation, and dosing less frequently than the half-life results in a less than twofold accumulation. The average plasma level to be expected at steady state can be calculated by Equation 9 where \( r \) is the dosing interval (1, p. 80),

\[
C_{\text{ave, ss}} = \frac{F \times \text{Dose}}{C \times r}
\]  

or can be determined from a numerical simulation using the Principle of Superposition (25, pp. 451-457). Importantly, a significant deviation of the accumulation index R-value from that predicted by the single-dose PK indicates a time-dependent alteration of clearance. For example, continuing exposure of the body to some drugs causes an adaptive response called autoinduction in which drug-metabolizing enzymes are upregulated to allow the body to eliminate the drug more rapidly (26). Such a drug would then show an R-value of less than one, which means that subsequent doses would lead to lower Cmax and AUC values. It is possible for autoinduction to be so severe that the drug becomes ineffective after a few doses. The opposite situation, although it is rare, also can occur; that is, R values can be much larger than expected, which means that when multiple dosing, the body becomes less able to clear the drug and plasma levels continue to rise. Greater–than–expected accumulation usually indicates some form of toxicity, such as organ toxicity or irreversible inhibition of clearance enzymes. The cytochrome P450 enzyme designated CYP3A4 is especially susceptible to irreversible inhibition (27), so the discovery process usually includes a screen for this type of behavior to prevent such compounds from reaching the clinic. Again, this type of nonideal PK behavior could be enough to cause the drug candidate to be clinically unusable.

An additional Phase I PK studies include the determination of the effect of food on orally administered drugs (28), formulation testing (29), detection of circulating metabolites of the drug (30), and checking for PK interactions between coadministered drugs (31). The most common form of drug-drug PK interaction is the competition of the two drugs for the same clearance mechanism, especially the CYP-P450 family of drug-metabolizing enzymes, which results in a limitation on which drugs ultimately
can be used clinically with the new agent (32). All clinical PK studies rely on some method of quantification of drug levels in blood or plasma, and the major bioanalytic technique in use today for this purpose is liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) (33, 34). Current technology permits the determination of picogram-per-mL levels of a drug in the presence of many interfering substances, with sample-to-sample cycle times less than 2 minutes. Because of individual biologic variability and inevitable experimental error, the measured drug concentrations at each time point will show variation among a group of subjects, which results in the need to express derived pharmacokinetic parameters as a mean value with an associated coefficient of variation (1, p. 203).

In addition to the conventional first-in-human approach described above, another available approach allows rapid access to the clinic to answer specific questions addressable by a single dose in humans. The U.S. Food and Drug Administration allows a sponsor to proceed to limited clinical dosing with an Exploratory IND (35), which has more moderate requirements than the standard IND. Determination of human PK is the most common use of the new approach, which can be thought of as an extension of the discovery process to check human PK before the decision is made to commit full development resources to pursue the conventional IND. One option within the Exploratory IND is microdosing, which has even lower requirements (36, 37). A microdose is defined as no more than 1/100 of the expected therapeutic dose but not exceeding 100 micrograms. The advantages of a microdose are that minimal safety testing is required because only minute quantities of the drug candidate are administered and that large-scale chemical synthesis can be postponed until acceptable clinical PK is confirmed.

**Phase II**

Although the main goal of Phase II clinical trials is the determination of efficacy toward the disease indication, PK is a fundamental part of this determination because to understand the clinical use of the new drug candidate, it is necessary to determine the temporal relationship between the plasma levels and the pharmacodynamics of the beneficial effect (PK/PD). A secondary PK goal in Phase II is to determine if the disease state affects the PK of the candidate drug compared with the PK observed in healthy subjects.

During the dose range-finding portion of Phase II, several doses are assessed for efficacy in patients with the disease to be treated. The PK sampling at these doses is critical because it allows the construction of a pharmacokinetic/pharmacodynamic (PK/PD) model whereby the observed drug concentrations are used to predict the effect of the drug (20, pp. 189-205, 38). The simplest PK/PD model is the direct or E\textsubscript{max} model in which a direct relationship exists between the concentration and the effect; changes in plasma levels are reflected immediately in the pharmacologic response. Often, however, as concentrations increase, eventually the pharmacologic responses reach a peak or nadir and greater increases in concentration result in only a small change in effect. Ideally, the drug candidate is safe enough to dose to the maximal effect. However, it is more common to strike a balance between the magnitude of response and the incidence of adverse side effects. Some drugs follow a more complex PK/PD relationship called the indirect model in which the pharmacologic response temporally lags behind changes in plasma levels (hysteresis). In this model, the response seen from a particular plasma concentration in the declining phase of the PK curve may be much greater than it was for the same plasma concentration during the rising phase of the curve. Hysteresis occurs when the initial stimulus of the drug binding to the receptor is uncoupled in time to the observed clinical response because a biologic cascade of several biochemical or physiologic events must occur for the clinical response to become manifest. After it is determined which type of relationship applies, a theoretical PK/PD model can be fit to the concentration-response data, which thereby generates a quantitative understanding of the action of the drug and the ability to predict the response to be expected in any situation.

With the availability of a PK/PD model, clinical trial simulation (39, 40) can be performed to explore alternative dosing regimens and to help achieve the optimal design of the Phase III program. The inputs for simulation include the number of patients to be tested, the duration of the testing, the potential clinical end points, and most importantly, the dose(s) to be tested. With Monte Carlo methods to simulate the variability in PK and response observed in Phases I and II, hundreds of computer simulations are conducted for each design and dose level to estimate the statistical probability of a successful outcome (i.e., one that meets predetermined end points of efficacy and safety). The goal is to find a design and dose that balances a high likelihood of a successful trial with a low incidence of adverse events. Ideally, this simulation activity is done in collaboration with government health authorities to minimize the approval time for the Phase III design and the chance of a failed Phase III study (41).

**Phase III**

In Phase III, large-scale trials are mounted with the goal of showing statistically significant benefit against either placebo or, more usually, standard-of-care therapy. The use of many patients affords the opportunity to create a Population PK model for the drug, which is a statistical means to account for individual variation in PK by means of clinically measurable covariates such as age, gender, race, ethnicity, body weight, kidney function, and so forth (42). In this way, the PK characteristics of a particular patient can be estimated before dosing, which allows a rational means to select the dose most likely to prevent good benefit while minimizing risk. After Phase III is complete, the sponsor of the drug candidate may submit an application (called the New Drug Application in the United States) for approval to government health authorities of all countries in which marketing is desired (43).

**Summary**

Pharmacokinetics studies the relationships for a given drug between blood concentrations, uptake, distribution, elimination, pharmacologic effect, and time. Because drugs that require large doses must be dosed frequently, cannot be dosed orally, must be timed in relation to food intake or have strong interactions...
with other drugs, may have issues of patient compliance, variable response, or toxicity, the pharmacokinetic aspects of the behavior of a drug are critical to its successful clinical use. Accordingly, the scientists who design drugs must use preclinical pharmacokinetic screening methods to maximize the likelihood that a drug candidate is acceptable pharmacokinetically. Correspondingly, the researchers who test the drug in clinical trials must ensure that all aspects of pharmacokinetic behavior have been characterized thoroughly before submitting the drug for approval to regulatory authorities.

References

Pharmacokinetics of Drug Candidates: Determination of


Further Reading


